

Corpus Linguistics: An International Handbook, Volume 2

Edited by
Anke Lüdeling
Merja Kytö

Walter de Gruyter

Corpus Linguistics

HSK 29.2



Handbücher zur Sprach- und Kommunikations- wissenschaft

Handbooks of Linguistics
and Communication Science

Manuels de linguistique et
des sciences de communication

Mitbegründet von Gerold Ungeheuer (†)
Mitherausgegeben 1985–2001 von Hugo Steger

Herausgegeben von / Edited by / Edités par
Herbert Ernst Wiegand

Band 29.2

Walter de Gruyter · Berlin · New York

Corpus Linguistics

An International Handbook

Edited by

Anke Lüdeling and Merja Kytö

Volume 2

Walter de Gruyter · Berlin · New York

② Printed on acid-free paper which falls within the guidelines
of the ANSI to ensure permanence and durability.

Library of Congress Cataloging-in-Publication Data

Corpus linguistics : an international handbook / edited by Anke Lüdeling and Merja Kytö
p. cm. — (Handbooks of linguistics and communication science ; 29.1 – 29.2)
Includes bibliographical references and indexes.
ISBN 978-3-11-018043-5 (hardcover : alk. paper) —
ISBN 978-3-11-020733-0 (hardcover : alk. paper) — I. Corpora (Linguistics) 2. computational linguistics.
I. Lüdeling, Anke, 1968 — II. Kytö, Merja.
p126.C68C663 2008
410—dc22

2008042529

ISBN 978-3-11-020733-0

ISSN 1861-5090

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© Copyright 2009 by Walter de Gruyter GmbH & Co. KG, 10785 Berlin, Germany.
All rights reserved, including those of translation into foreign languages. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording or any information storage and retrieval system, without permission in writing from the publisher.

Printed in Germany

Typesetting: META Systems GmbH, Wustermark

Coverdesign: Martin Zech, Bremen

Contents

Volume 2

V. Use and exploitation of corpora	
36. Marco Baroni/Stefan Evert, Statistical methods for corpus exploitation	777
37. Marco Baroni, Distributions in text	803
38. Douglas Biber, Multi-dimensional approaches	822
39. Antal van den Bosch, Machine learning	855
40. Hermann Moisl, Exploratory multivariate analysis	874
41. R. Harald Baayen, Corpus linguistics in morphology: Morphological productivity	899
42. W. Detmar Meurers/Stefan Müller, Corpora and syntax	920
43. Anatol Stefanowitsch/Stefan Th. Gries, Corpora and grammar	933
44. Sabine Schulte im Walde, The induction of verb frames and verb classes from corpora	952
45. Michael Hoey, Corpus linguistics and word meaning	972
46. Richard Xiao, Theory-driven corpus research: Using corpora to inform aspect theory	987
47. Michael McCarthy/Anne O'Keeffe, Corpora and spoken language	1008
48. Anders Lindström/Robert Eklund, Cross-lingual influence: The integration of foreign items	1024
49. Tuija Virtanen, Corpora and discourse analysis	1043
50. Michael P. Oakes, Corpus linguistics and stylometry	1070
51. Anne Curzan, Historical corpus linguistics and evidence of language change	1091
52. Christian Mair, Corpora and the study of recent change in language	1109
53. Lieselotte Anderwald/Benedikt Szmrecsanyi, Corpus linguistics and dialectology	1126
54. Josef Schmied, Contrastive corpus studies	1140
55. Silvia Hansen-Schirra/Elke Teich, Corpora in human translation	1159
56. Harold Somers, Corpora and machine translation	1175
57. Holger Diessel, Corpus linguistics and first language acquisition	1197
58. Stefan Evert, Corpora and collocations	1212
59. Paul Clough/Rob Gaizauskas, Corpora and text re-use	1249
60. Constantin Orasan/Laura Hasler/Ruslan Mitkov, Corpora for text summarisation	1271
61. Douglas Biber/James K. Jones, Quantitative methods in corpus linguistics	1287
Indexes	
Index of names	1305
Index of corpora and repositories	1333
Subject index	1343

Volume 1

Introduction	v
Acknowledgments	xiii
I. Origin and history of corpus linguistics – corpus linguistics vis-à-vis other disciplines	
1. Charles F. Meyer, Pre-electronic corpora	1
2. Fred Karlsson, Early generative linguistics and empirical methodology	14
3. Stig Johansson, Some aspects of the development of corpus linguistics in the 1970s and 1980s	33
4. Matti Rissanen, Corpus linguistics and historical linguistics	53
5. Stefanie Dipper, Theory-driven and corpus-driven computational linguistics, and the use of corpora	68
6. Suzanne Romaine, Corpus linguistics and sociolinguistics	96
7. Ute Römer, Corpora and language teaching	112
8. Ulrich Heid, Corpus linguistics and lexicography	131
II. Corpus compilation and corpus types	
9. Susan Hunston, Collection strategies and design decisions	154
10. Marianne Hundt, Text corpora	168
11. Anne Wichmann, Speech corpora and spoken corpora	187
12. Jens Allwood, Multimodal corpora	207
13. Joakim Nivre, Treebanks	225
14. Claudia Claridge, Historical corpora	242
15. Sylviane Granger, Learner corpora	259
16. Karin Aijmer, Parallel and comparable corpora	275
17. Michael Beißwenger/Angelika Storrer, Corpora of computer-mediated communication	292
18. Gunnar Bergh/Eros Zanchetta, Web linguistics	309
19. Alexander Mehler, Large text networks as an object of corpus linguistic studies	328
III. Existing corpora	
20. Richard Xiao, Well-known and influential corpora	383
21. Nicholas Ostler, Corpora of less studied languages	457
IV. Preprocessing corpora	
22. Timm Lehmberg/Kai Wörner, Annotation standards	484
23. Eric Atwell, Development of tag sets for part-of-speech tagging	501
24. Helmut Schmid, Tokenizing and part-of-speech tagging	527
25. Arne Fitschen/Piklu Gupta, Lemmatising and morphological tagging .	552

26.	Paul Rayson/Mark Stevenson, Sense and semantic tagging	564
27.	Ruslan Mitkov, Corpora for anaphora and coreference resolution	579
28.	Hannah Kermes, Syntactic preprocessing	598
29.	Dawn Archer/Jonathan Culpeper/Matthew Davies, Pragmatic annotation	613
30.	Nelleke Oostdijk/Lou Boves, Preprocessing speech corpora: Transcription and phonological annotation	642
31.	Peter Wittenburg, Preprocessing multimodal corpora	664
32.	Michael P. Oakes, Preprocessing multilingual corpora	685
33.	Martin Wynne, Searching and concordancing	706
34.	Sean Wallis, Searching treebanks and other structured corpora	738
35.	Heike Zinsmeister/Erhard Hinrichs/Sandra Kübler/Andreas Witt, Linguistically annotated corpora: Quality assurance, reusability and sustainability	759

V. Use and exploitation of corpora

36. Statistical methods for corpus exploitation

1. Introduction
2. The logic behind hypothesis testing
3. Estimation and effect size
4. The normal approximation
5. Two-sample tests
6. Linguistic units and populations
7. Non-randomness and the unit of sampling
8. Other techniques
9. Directions for further study
10. Literature

1. Introduction

Linguists look for generalizations and explanations of various kinds for linguistic phenomena. While the interest is usually in an *intensional* view of these phenomena, to be explained in terms of the human language competence, such competence cannot be directly observed. Thus, evidence has to come from an external reflection of it, i. e., it has to be based on an *extensional* view of language. According to this extensional view, a language is defined as the set of all utterances produced by speakers of the language (with all the paradoxes that this view implies – see, e. g., Chomsky 1986, chapter 2). Corpora are finite samples from the infinite set that constitutes a language in this extensional sense. For example, in this perspective, the Brown corpus (see article 20) is a finite sample of all the utterances produced in written form by American English speakers. Psycholinguistic experiments, such as eye-tracking tests, priming, and even traditional grammaticality judgments (Schütze 1996) constitute other sources of evidence. It is important to observe that the empirical analysis of these other sources also requires an extensional view of language.

It is rarely the case that linguists are interested in the samples *per se*, rather than in generalizations from the samples to the infinite amount of text corresponding to the extensional definition of a (sub)language. For example, a linguist studying a pattern in the 500 text samples of the Brown corpus will typically be interested in drawing conclusions about (written) American English as a whole, and not just about the specific texts that compose the Brown. Statistical inference allows the linguist to generalize from properties observed in a specific sample (corpus) to the same properties in the language as a whole (statistical inference, on the other hand, will not be of help in solving thorny issues such as what the appropriate extensional definition of a “language as a whole” is and how we can sample from that).

Statistical inference requires that the problem at hand is *operationalized* in quantitative terms, typically in the form of units that can be *counted* in the available sample(s). This is the case we will concentrate on here (but see section 8 for other kinds of measure-

ments). For example, a linguist might be interested in the issue of whether a certain variety of English is more “formal” than another (as in some of Douglas Biber’s work, see article 38). In order to operationalize this research question, the linguist might decide to take passivization as a cue of formality, and count the number of sentences that display passivization in samples from the two varieties. Statistical inference can then be used to generalize from the difference in number of passives between the two samples to the difference between the two varieties that the samples represent (we will discuss this example and the appropriate techniques further in section 5). Similarly, a linguist might be interested in whether (certain classes of) idiomatic constructions have a tendency to repel passive formation (as observed by Culicover/Jackendoff 2005 and many others). In order to operationalize this question, the linguist may count the number of passives in idiomatic and non-idiomatic phrases in a corpus. Statistical inference will then help to determine how reliably the attested difference in passive frequency would generalize to idiomatic and non-idiomatic phrases at large. Of course, it is up to the linguist to interpret the generalizations about frequencies produced by statistical analysis in terms of the linguistic phenomena of interest.

Statistical inference is necessary because any sample from a language is subject to random variation. Suppose that someone doubted the claim that non-idiomatic constructions are more prone to passivization than idiomatic constructions, and we wanted to dispel these doubts. A sample of language that reveals a higher proportion of passives among the non-idiomatic constructions, especially if the difference in proportions is small, would not allow us to reject the doubters’ hypothesis: even if they were right, we could not expect the proportions to be exactly identical in *all* samples of a language. Statistical inference can help us to determine to what extent the difference between a sample-based observation and a theoretical prediction can be taken as serious evidence that the prediction made by the theory is wrong, and to what extent it can reasonably be attributed to random variation. In the case at hand, statistical inference would tell us whether the difference in passive rates in the two samples can be explained by random variation, or whether it is the symptom of a true underlying difference. It is perhaps worth clarifying from the outset that randomness due to sampling has to be distinguished from measurement errors, such as those introduced by the automatic annotation and analysis of corpus data (something that statistical methods will not help us correct). Suppose that a very skilled linguist sampled 100 English sentences and recorded very carefully how many of them have passives, without making any errors. It should be intuitive that, given another random sample of 100 sentences and the same error-free linguist, the exact number of passives would probably be different from the one found in the previous sample. This is the random variation we are referring to here.

Notice that the necessity of statistical inference pertains to the need to generalize from a finite (random) sample of language data to the theoretically infinite amount of text corresponding to the extensional definition of an entire (sub)language, and it has nothing to do with whether our theory about the phenomenon at hand, or about language competence in general, includes a probabilistic component. The prediction that idiomatic sentences repel the passive construction might stem from a completely categorical theory of how passives and idiomticity interact – still, randomly sampled English sentences will display a certain amount of variation in the exact proportion of passives they contain.

The rest of this article introduces the basics of statistical inference. We use the artificially simple example of testing a hypothesis about the proportion of passives in English

sentences (and later proportions of passives in sentences from different English genres), in order to focus on the general philosophy and methodology of statistical inference as applied to corpus linguistics, rather than on the technical details of carrying out the relevant computations, which can be found in many general books on the subject and are implemented in all standard statistical packages (see references in section 9). Section 6 gives examples of how statistical inference can be applied to more realistic linguistic analysis settings.

2. The logic behind hypothesis testing

Imagine that an American English style guide claims that 15% of the sentences in the English language are in the passive voice (as of June 2006, <http://www.ego4u.com/en/business-english/grammar/passive> makes the even bolder statement that no more than 10% of English sentences are in the passive voice and writers should be careful to use passives sparingly). This is a fairly easy claim to operationalize, since it is already phrased in terms of a proportion. However, we still need to define what we understand by “the English language”, and what it means for a sentence to be in the passive voice. Given the source of the claim and our need for an extensional definition, it makes sense to take “English” to mean the set of all English texts published in the US and produced by professional writers. Regarding the second issue, we consider a sentence to be in the passive voice if it contains at least one verb in the passive form, which seems to be a plausible interpretation of what the style guide means (after all, it is warning against the overuse of passives), and at the same time makes it easier to count the number of sentences in passive voice using automated pattern matching techniques (which might not be relevant with the small samples we use here, but would be important when dealing with large amounts of data).

It is of course impossible to look at all sentences in all the publications satisfying the criteria above – what we can do, at best, is to select a random sample of them. In particular, we took a random sample of 100 sentences of the relevant kind, and we counted the number of them containing a passive. For convenience, we restricted ourselves to publications from 1961, because we are lucky enough to already own a random sample of sentences of the relevant kind from that year – namely, the Brown corpus! All we had to do was select 100 random sentences from this random sample (we will see in section 7 that it is not entirely correct to treat sentences from the Brown as a random sample, but we ignore this for now).

If the style guide’s claim is true, we would expect 15 sentences to be in the passive voice. Instead, we found 19 passives. This seems to indicate that the proportion is higher than 15% and rather close to 20%. However, it is obvious that, even if the claim of the style guide was correct, not *all* samples of size 100 would have exactly 15 passives, because of random variation. In light of this, how do we decide whether 19 passives are enough to reject the style guide’s claim?

In statistical terms, the claim that we want to verify is called a *null hypothesis*, $H_0: \pi = 15\%$, where π is the putative proportion of passives in the set of sentences that constitute our extensional definition of American English. This set of sentences is usually called a *population* in statistical parlance, and the goal of statistical inference is to draw

conclusions about certain properties of this population from an available sample (the population itself is practically infinite for all intents and purposes, and we can only access a small finite subset of it). We will often refer to π as a population proportion or parameter in what follows. The number of sentences we have randomly sampled from the population is called the *sample size*, $n = 100$. Intuitively, we expect $e = n \cdot \pi = 15$ passives in the sample if the null hypothesis is true. This is called the *expected frequency*. The number of passives we actually observed in the sample, $f = 19$, is called the *observed frequency*.

Having introduced the terminology, we can rephrase the problem above as follows. If we are prepared to reject the null hypothesis that $\pi = 15\%$ for an observation of $f = 19$, there is a certain risk that in doing so we are making the wrong decision. The question is how we can quantify this risk and decide whether it is an acceptable risk to take. Imagine that the null hypothesis in fact holds, and that a large number of linguists perform the same experiment, sampling 100 sentences and counting the passives. We can then formally define risk by the percentage of linguists who wrongly reject the null hypothesis, and thus publish incorrect results. In particular, if our observation of $f = 19$ is deemed sufficient for rejection, all the other linguists who observed 19 or even more passives in their samples would also reject the hypothesis. The risk is thus given by the percentage of samples containing 19 or more passives that would be drawn from a language in which the true proportion of passives is indeed 15%, as stipulated by H_0 . Rejecting the null hypothesis when it is in fact true is known as a *type-1 error* in the technical literature (failure to reject H_0 when it does not hold constitutes a *type-2 error*, which we do not discuss here, but see, e.g., DeGroot/Schervish 2002, chapter 8).

Fortunately, we do not need to hire hundreds of linguists to compute the risk of wrong rejection, since the thought experiment above is fully equivalent to drawing balls from an urn. Each ball represents a sentence of the language, with red balls for passive sentences and white balls for sentences in other voices. The null hypothesis stipulates that the proportion of red balls in the urn is 15%. The observed number of red balls (passives) changes from sample to sample. In statistical terminology, it is called a *random variable*, typically denoted by a capital letter such as X . We simulate a large number of samples from the urn with a computer and tabulate how often each possible value k of the random variable X is observed. The result of this simulation is shown in Figure 36.1, which reports percentages of samples that yield $X = k$ for k ranging from 0 to 30 (the percentage is indistinguishable from 0 for all values outside this range). For instance, the value $X = 19$ can be observed in 5.6% of the samples. The information presented in this graph is called the *sampling distribution* of X under H_0 . The percentage of samples with $X = k$ is called the *probability* $\Pr(X = k)$. For example, $\Pr(X = 19) = 5.6\%$ (our reasoning in this section has led us to what is known as the *frequentist* definition of probabilities; we do not discuss the alternative *Bayesian* interpretation of probability theory here, but see for example section 1.2. of DeGroot/Schervish 2002).

Following the discussion above, the risk of wrongly rejecting the null hypothesis for an observation of $f = 19$ is given by the percentage of samples with $X \geq 19$ in the sampling distribution, i.e., the probability $\Pr(X \geq 19)$. This probability can be computed by summing over the shaded bars in Figure 36.1:

$$\Pr(X \geq 19) = \Pr(X = 19) + \Pr(X = 20) + \dots + \Pr(X = 100) = 16.3\% \quad (1)$$

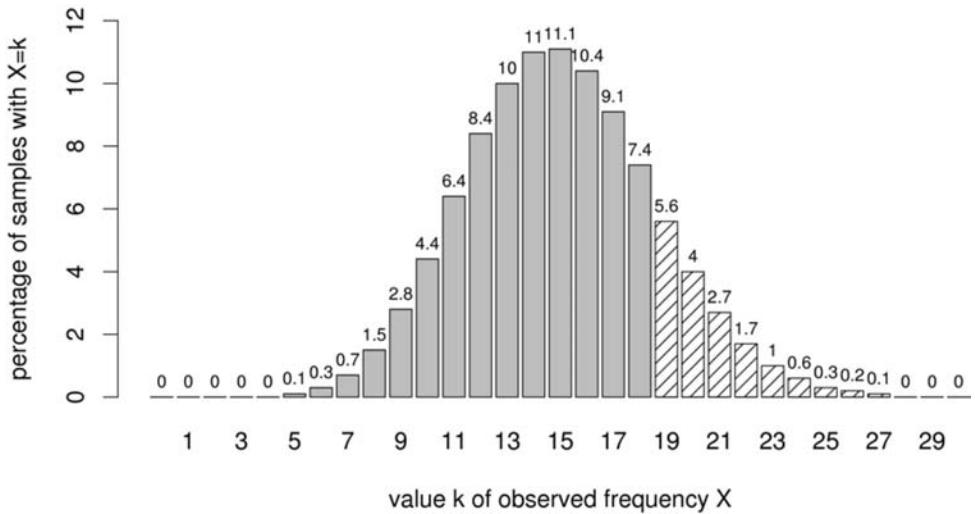


Fig. 36.1: Sampling distribution of X with $n = 100$ and $\pi = 15\%$

This is called a *tail probability* because it sums over the right-hand “tail” of the distribution. In the same way, we can compute the risk associated with any other value f , namely the probability:

$$\Pr(X \geq f) := \sum_{k=f}^n \Pr(X = k) \quad (2)$$

We refer to this risk as the *p-value* of an observation f . Notice that *smaller* p-values are *more* significant, since they indicate that it is less risky to reject the null hypothesis, and hence they allow greater confidence in the conclusion that the null hypothesis should be rejected. In our example, the p-value associated with $f = 19$ indicates that the risk of false rejection is unacceptably high at $p = 16.3\%$. If we used $f = 19$ as the threshold for rejection and the null hypothesis happened to be true, about one in six experiments would lead to wrong conclusions. In order to decide whether to reject H_0 or not, the computed p-value is often compared with a conventional scale of *significance levels*. A p-value below 5% is always required to consider a result significant. Other common significance levels are 1% and 0.1% (usually written as mathematical fractions rather than percentages and denoted by the symbol α , viz. $\alpha = .05$, $\alpha = .01$ and $\alpha = .001$).

So far, we have only been considering cases in which the observed frequency is greater than the one predicted under H_0 – reflecting our intuition that the proportion of passives proposed by the style guide errs on the side of being too low, rather than too high. However, in principle it would also be possible that the proportion of passives is *lower* than predicted by H_0 . Coming back to our passive-counting linguists, if they are prepared to reject H_0 for $f = 19$, they should also reject it for $X = 7$ or $X = 8$, since these values are even more “extreme” than 19 with respect to H_0 . Thus, when computing the p-value of $f = 19$, it is typically appropriate to sum over the probabilities of all values that are at least as “extreme” as the observed value, to either side of the expected frequency e , since they add to the risk of false rejection.

It is difficult to determine exactly which of the values below e should count as equally extreme as, or more extreme than f , but one reasonable approach is to include all the values of X with an absolute difference $|X - e| \geq |f - e|$. Using $|f - e|$ (or, more precisely, $(f - e)^2$, which has certain mathematical advantages) as a measure of “extremeness” leads to a class of statistical tests known as *chi-squared tests*. An alternative approach would rather compare the probabilities $\Pr(X = k)$ and $\Pr(X = f)$ as a measure of extremeness, resulting in a class known as likelihood (ratio) tests. In many common cases, both classes of tests give very similar results. We will focus on chi-squared tests in this article, but see article 58 for an application where likelihood tests are known to be superior.

In the case at hand, using the chi-squared criterion, the p-value would be computed by adding up the probabilities of $X \geq 19$ and $X \leq 11$ (since $|19 - 15| = 4 = |11 - 15|$). In the illustration shown in Figure 36.1, we would add the bars for $k = 1\dots11$ to the shaded area. This way of computing the p-value, taking both “extreme” tails of the distribution into account, is called a *two-tailed test* (the approach above, where we considered only one side, is known as a *one-tailed test*). Of course, a two-tailed p-value is always greater than (or equal to) the corresponding one-tailed p-value. In our running example, the two-tailed p-value obtained by summing over the bars for $X \leq 11$ and $X \geq 19$ turns out to be 32.6%, indicating a very high risk if we chose to reject the null hypothesis for $f = 19$ (you might obtain a different p-value if the binomial test implemented in your software package uses the likelihood criterion, although the value will still indicate a very high risk in case of rejection). Had our experiment yielded 22 passives instead, the one-tailed test would have produced a p-value of 3.9%, while the two-tailed test would have given a p-value of 6.7%. Thus, by adopting the common 5% significance threshold, we would have had enough evidence to reject the null hypothesis according to the one-tailed test, but not enough according to the two-tailed test.

As a general rule, one should always use the more conservative two-tailed test, unless there are very strong reasons to believe that the null hypothesis could only be violated in one direction – but it is hard to think of linguistic problems where this is the case (in many situations we can predict the probable direction of the violation, but there are very few cases where we would be ready to claim that a violation in the other direction is absolutely impossible). If we use a two-tailed test, the interpretation of a significant result will of course have to take into account whether f is greater or smaller than e . Observed frequencies of 25 and 5 passives, respectively, would both lead to a clear rejection of the null hypothesis that 15% of all sentences are in the passive voice, but they would require rather different explanations.

Not only have we been spared the expense of hiring passive-counting linguists to repeat the experiment; it is not even necessary to perform expensive computer simulation experiments in order to carry out the sort of tests we just illustrated, because $\Pr(X = k)$ – the percentage of samples of size n from a population with proportion π of passive sentences that would result in a certain value k of the random variable X – can be computed with the following formula, known as the *binomial distribution* (the hypothesis test we have described above, unsurprisingly, is called a *binomial test*):

$$\Pr(X = k) = \binom{n}{k} (\pi)^k (1 - \pi)^{n-k} \quad (3)$$

The binomial coefficient $\binom{n}{k}$, “ n choose k ”, represents the number of ways in which an unordered set of k elements can be selected from n elements. Any elementary textbook on probability theory or statistics will show how to compute it; see, e.g., DeGroot/Schervish (2002, section 1.8.). Of course, all statistical software packages implement binomial coefficients and the binomial distribution. For a different null hypothesis about the population proportion π or a different sample size n , we obtain sampling distributions with different peaks and shapes – in statistical terminology, π and n are the *parameters* of the binomial distribution. In particular, the value of π affects the location of the peak in the histogram. For example, if we hypothesized that $\pi = 30\%$, we would see a peak around the expected value $e = n \cdot \pi = 30$ in the histogram corresponding to Figure 36.1. Intuitively, experiments in which we draw 1,000 balls will tend to produce outcomes that are closer to the expected value than experiments in which we draw 100 balls. Thus, by decreasing or increasing n , we obtain distributions that have narrower or wider shapes, respectively. A sample of size 100 is small by the standards of statistical inference. As Karl Pearson, one of the founding fathers of modern statistics, once put it: “Only naughty brewers deal in small samples!” (cf. Pearson 1990, 73; this quip was a reference to W. S. Gosset, an employee of the Guinness brewery who developed and published the now famous t-test under the pseudonym of “Student”). It will typically be difficult to reject H_0 based on such a sample, unless the true proportion is very far away from the null hypothesis, exactly because a small sample size leads to a wide sampling distribution. Had we taken a sample of 1,000 sentences and counted 190 passives, the null hypothesis would have been clearly rejected (a two-sided binomial test with $f = 190$, $n = 1000$ and $H_0: \pi = 15\%$ gives a p-value of $p = 0.048\%$, sufficient for rejection even at the very conservative significance level $\alpha = .001$).

The procedure of hypothesis testing that we introduced in this section is fundamental to understanding statistical inference. At the same time, it is not entirely intuitive. Thus, before we move on, we want to summarize its basic steps. For the whole process to be meaningful, we must have a *null hypothesis* H_0 that operationalizes a research question in terms of a quantity that can be computed from observable data. In our case, the null hypothesis stipulates that the proportion of passives in the *population* of (professionally written American) English sentences is 15%, i.e.: $H_0: \pi = 15\%$. We draw a random *sample* of size n of the relevant units (100 sentences in our case) from the population, and count the number of units that have the property of interest (in our case, being passive sentences). Given the population proportion stipulated by the null hypothesis and the sample size, we can determine a *sampling distribution* (by simulation or using a mathematical formula). The sampling distribution specifies, for each possible outcome of the experiment (expressed by the *random variable* X , which in our case keeps track of the frequency of passives in the sample), how likely it is under the null hypothesis. This *probability* is given by the percentage of a large number of experiments that would produce the outcome X in a world in which the null hypothesis is in fact true. The sampling distribution allows us, for every possible value k of X , to compute the *risk* of making a mistake when we are prepared to reject the null hypothesis for $X = k$. This risk, known as the *p-value* corresponding to k , is given by the overall percentage of experiments that give an outcome at least as extreme as $X = k$ in a world in which the null hypothesis is true (see above for the *one-* and *two-tailed* ways to interpret what counts as “extreme”). At this point, we look at the actual outcome of the experiment in our sample, i.e., the

observed quantity f (in our case, f is the number of passives in a sample of 100 sentences), and we compute the p-value (risk) associated with f . In our example, the (two-tailed) p-value is 32.6%, indicating a rather high risk in rejecting the null hypothesis. We can compare the p-value we obtained with conventional thresholds, or *significance levels*, that correspond to “socially acceptable” levels of risk, such as the 5% threshold $\alpha = .05$. If the p-value is higher than the threshold, we say that the results of the experiment are not *statistically significant*, i. e., there is a non-negligible possibility that the results would be obtained by chance even if the null hypothesis is true.

Notice that a non-significant result simply means that our evidence is not strong enough to reject the null hypothesis. It does *not* tell us that the null hypothesis is correct. In our example, although the observed frequency is not entirely unlikely under the null hypothesis of a passive proportion of 15%, there are many other hypotheses under which the same result would be even more likely, most obviously, the hypothesis that the population proportion is 19%. Because of this indirect nature of statistical hypothesis testing, problems undergoing statistical treatment are typically operationalized in a way in which the null hypothesis is “uninteresting”, or contradicts the theory we want to support. Our hope is that the evidence we gather is strong enough to reject H_0 . We will come back to this in section 5 below, presenting a two-sample setting where this strategy should sound more natural.

While many problems require more sophisticated statistical tools than the ones described in this section, the basic principles of hypothesis testing will be exactly the same as in the example we just discussed.

3. Estimation and effect size

Suppose that we ran the experiment with a sample of $n = 1,000$ sentences, $f = 190$ of which turned out to be in the passive voice. As we saw in the previous section, this result with the larger sample leads to a clear rejection of the null hypothesis $H_0: \pi = 15\%$. At this point, we would naturally like to know what the *true* proportion of passives is in edited American English. Intuitively, our best guess is the observed proportion of passives in the sample, i. e., $\hat{\pi} = f/n$. This intuitive choice can also be justified mathematically. It is then known as a *maximum-likelihood estimate* or *MLE* (DeGroot/Schervish 2002, section 6.5.).

Since we have estimated a single value for the population proportion, $\hat{\pi}$ is called a *point estimate*. The problem with point estimates is that they are subject to the same amount of random variation as the observed frequency on which they are based: most linguists performing the same experiment would obtain a different estimate $\hat{\pi} = X/n$ (note that, mathematically speaking, $\hat{\pi}$ is a random variable just like X , which assumes a different value for each sample).

Let us put the question in a slightly different way: besides the point estimate $\hat{\pi} = 19\%$, which other values of π are also plausible given our observation of $f = 190$ passives in a sample of $n = 1,000$ sentences? Since $H_0: \pi = 15\%$ was rejected by the binomial test, we know for instance that the value $\pi = 15\%$ is *not* plausible according to our observation. This approach allows us to answer the question in an indirect way. For any potential estimate $\pi = x$, we can perform a binomial test with the null hypothesis $H_0: \pi = x$ in

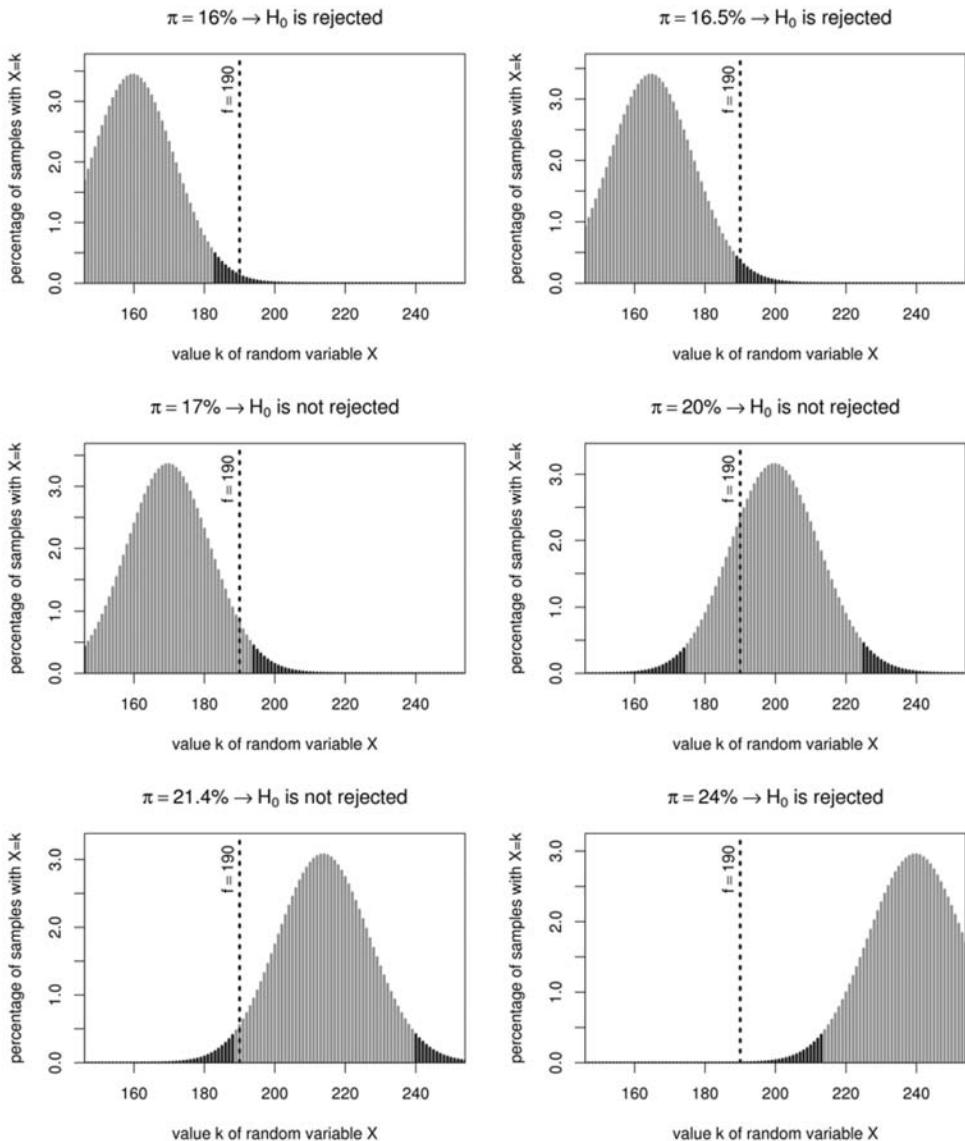


Fig. 36.2: Illustration of the procedure for estimating a confidence set

order to determine whether the value x is plausible (H_0 cannot be rejected at the chosen significance level α) or not (H_0 can be rejected). Note that failure to reject H_0 does not imply that the estimate x is very likely to be accurate, but only that we cannot rule out the possibility $\pi = x$ with sufficient confidence. Figure 36.2 illustrates this procedure for six different values of x , when $f = 190$ and $n = 1,000$. As the figure shows, $H_0 : \pi = 17\%$ would not be rejected, and thus 17% is in our set of plausible values. On the other hand, $H_0 : \pi = 16.5\%$ would be rejected, and thus 16.5% is not in our set.

Collecting all plausible values $\pi = x$, we obtain a *confidence set*. For the binomial test, this confidence set is an uninterrupted range of numbers and is called a *binomial confidence interval*. Of course, it is infeasible to perform separate hypothesis tests for the infinite number of possible null hypotheses $\pi = x$, but specialized mathematical algorithms (available in all standard statistical software packages) can be used to compute the end points of binomial confidence intervals efficiently. In our example, the observed data $f = 190$ and $n = 1000$ yield a confidence interval of $\pi \approx 16.6\% \dots 21.6\%$ (the common mathematical notation for such a range, which you may encounter in technical literature, is $[.166, .216]$).

The width of a binomial confidence interval depends on the sample size n and the significance level α used in the test. As we have seen in section 2, a larger value of n makes it easier to reject the null hypothesis. Obviously, adopting a higher (i.e., less conservative) value of α also makes it easier to reject H_0 . Hence these factors lead to a narrower confidence interval (which, to reiterate this important point, consists of all estimates x for which H_0 is *not* rejected). Table 36.1 shows confidence intervals for several different sample sizes and significance levels. A confidence interval for a significance level of $\alpha = .05$ (which keeps the risk of false rejection below 5%) is often called a 95% confidence interval, indicating that we are 95% certain that the true population value π is somewhere within the range (since we can rule out any other value with 95% certainty). Similarly, a significance level of $\alpha = .01$ leads to a 99% confidence interval.

Tab. 36.1: Binomial confidence intervals for various sample sizes n and confidence levels α . The maximum-likelihood estimate is $\hat{\pi} = 19\%$ in each case

	$n = 100$ $k = 19$	$n = 1,000$ $k = 190$	$n = 10,000$ $k = 1,900$
$\alpha = .05$	11.8% ... 28.1%	16.6% ... 21.6%	18.2% ... 19.8%
$\alpha = .01$	10.1% ... 31.0%	15.9% ... 22.4%	18.0% ... 20.0%
$\alpha = .001$	8.3% ... 34.5%	15.1% ... 23.4%	17.7% ... 20.3%

Confidence intervals can be seen as an extension of hypothesis tests. The 95% confidence interval for the observed data immediately tells us whether a given null hypothesis $H_0: \pi = x$ would be rejected by the binomial test at significance level $\alpha = .05$. Namely, H_0 is rejected if and only if the hypothesized value x does *not* fall within the confidence interval. The width of a confidence interval illustrates thus how easily a null hypothesis can be rejected, i.e., it gives an indication of how much the (unknown) true population proportion π must differ from the value stipulated by the null hypothesis (which is often denoted by the symbol π_0) so that H_0 will reliably be rejected by the hypothesis test. Intuitively speaking, the difference between π and π_0 has to be considerably larger than the width of one side of the 95% confidence interval so that it can reliably be detected by a binomial test with $\alpha = .05$ (keep in mind that, even when the difference between π and π_0 is larger than this width, because of sampling variation, $\hat{\pi}$ and π_0 might be considerably closer, leading to failure to reject H_0). The term *effect size* is sometimes used as a generic way to refer to the difference between null hypothesis and true proportion. The reliability of rejection given a certain effect and sample size is called the *power* of the hypothesis test (see DeGroot/Schervish 2002, chapter 8). In our example, the arithmetic difference $\pi - \pi_0$ is a sensible way of quantifying effect size, but many other measures

exist and may be more suitable in certain situations (we will return to this issue during the discussion of two-sample tests in section 5).

In corpus analysis, we often deal with very large samples, for which confidence intervals will be extremely narrow, so that a very small effect size may lead to highly significant rejection of H_0 . Consider the following example: Baayen (2001, 163) claims that the definite article *the* accounts for approx. 6% of all words in (British) English, including punctuation and numbers. Verifying this claim on the LOB (the British equivalent of the Brown corpus, see article 20), we find highly significant evidence against H_0 . In particular, there are $f = 68,184$ instances of *the* in a sample of $n = 1,149,864$ words. A two-sided binomial test for $H_0: \pi = 6\%$ rejects the null hypothesis with a p-value of $p \approx 0.1\%$.

However, the MLE for the true proportion π is actually very close to 6%, viz. $\hat{\pi} = 5.93\%$, and the 95% confidence interval is $\pi = 5.89\% \dots 5.97\%$. This difference is certainly not of scientific relevance, and $\hat{\pi}$ as well as the entire confidence range would be understood to fall under Baayen's claim of "approximately 6%". The highly significant rejection is merely a consequence of the large sample size and the corresponding high power of the binomial test. Gries (2005) is a recent discussion of the "significance" of statistical significance in corpus work.

At the opposite end of the scale, it is sometimes important to keep the sample size as small as possible, especially when the preparation of the sample involves time-consuming manual data annotation. Power calculations, which are provided by many statistical software packages, can be used to predict the minimum sample size necessary for a reliable rejection of H_0 , based on our conjectures about the true effect size.

4. The normal approximation

Looking back at Figure 36.1, we can see that the binomial sampling distribution has a fairly simple and symmetric shape, somewhat reminiscent of the outline of a bell. The peak of the curve appears to be located at the expected frequency $e = 15$. For other parameter values π and n , we observe the same general shape, only stretched and/or translated. This bell-shaped curve can be described by the following mathematical function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (4)$$

This is the formula of a *normal* or *Gaussian distribution* (DeGroot/Schervish 2002, section 5.6.). The parameter μ , called the mean, will determine the peak of the bell-shaped curve, and the parameter σ , called the *standard deviation*, will determine the width of the curve (the symbol π in this formula stands for Archimedes' constant $\pi = 3.14159\dots$ and not for a population proportion; to avoid another ambiguity, we write $\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for the exponential function in lieu of the more commonly encountered $e^{-(x-\mu)^2/2\sigma^2}$, since we are using e to denote the expected frequency). The roles of the two parameters are illustrated in Figure 36.3.

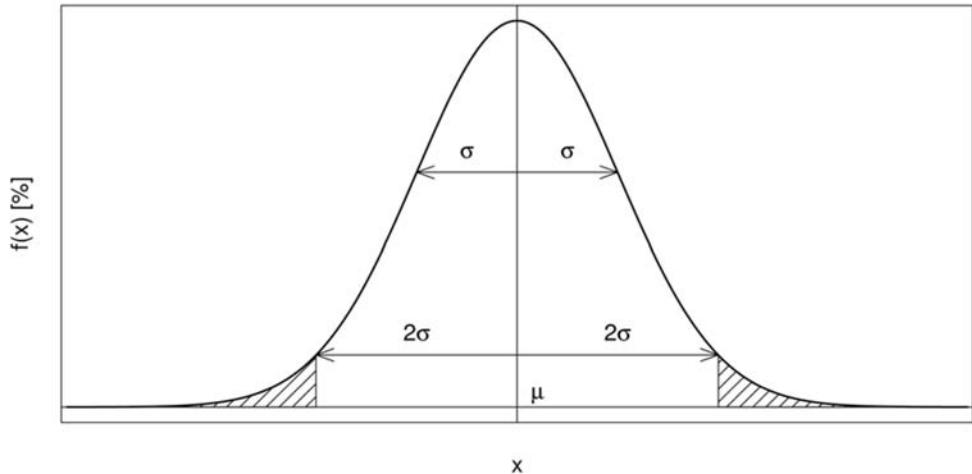


Fig. 36.3: Interpretation of the parameters μ and σ of the normal distribution

A binomial distribution with parameters n and π is approximated by a normal distribution with parameters $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1 - \pi)}$. Figure 36.4 shows the same binomial distribution illustrated in Figure 36.1 (with sample size $n = 100$ and proportion $\pi = 15\%$) and the corresponding normal approximation with parameters $\mu = 15$ and $\sigma \approx 3.57$. The quality of the approximation will increase with sample size and it will depend on π not being too skewed (i.e., not too close to 0 or 1). A rule of thumb might be to trust the approximation only if $\sigma > 3$, which is the case in our example (if you refer back to the formula for σ , you will notice that it depends, indeed, on n and the skewness of π).

The parameters of the normal approximation can be interpreted in an intuitive manner: μ coincides with the expected frequency e under H_0 (remember from section 2 that

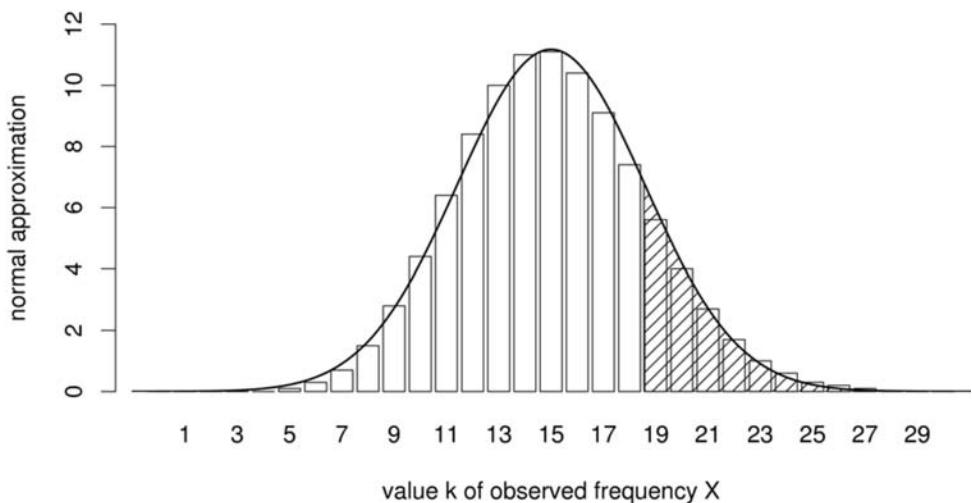


Fig. 36.4: Approximation of binomial sampling distribution by normal distribution curve

e is also given by $n\pi$; we will use μ when referring to the normal distribution formula, e otherwise, but keep in mind that the two symbols denote the same quantity). σ tells us how much random variation we have to expect between different samples. Most of the samples will lead to observed frequencies between $\mu - 2\sigma$ and $\mu + 2\sigma$ and virtually all observed values will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$ (refer to Figure 36.3 again), provided that H_0 is true.

To compute binomial tail probabilities based on a normal approximation, one calculates the corresponding area under the bell curve, as illustrated in Figure 36.4 for the tail probability $\Pr(X \geq 19)$. In this illustration, we have also applied *Yates' continuity correction* (DeGroot/Schervish 2002, section 5.8.), which many statistical software packages use to make adjustments for the discrepancies between the smooth normal curve and the discrete distribution that is approximated. In our example, Yates' correction calculates the area under the normal curve for $x \geq 18.5$ rather than $x \geq 19$.

We find that the normal approximation gives a one-tailed p-value of 16.3% for observed frequency $f = 19$, sample size $n = 100$ and null hypothesis $H_0: \pi = 15\%$. This is the same p-value we obtained from the (one-tailed) binomial test, indicating that the approximation is very good. Given that the normal distribution (unlike the binomial!) is always symmetrical, the two-tailed p-value can be obtained by simply multiplying the one-tailed value by two (which corresponds to adding up the tail areas under the curve for values that are at least as extreme as the observed value, with respect to e). In our case this gives 32.6%, again equivalent to the binomial test result.

There are two main reasons why the normal approximation is often used in place of the binomial test. First, the exact (non-approximated) binomial test and binomial confidence intervals require computationally expensive procedures that, for large sample sizes such as those often encountered in corpus-based work, can be problematic even for modern computing resources (a particularly difficult case is the extension of confidence intervals to the two-sample setting that we introduce in section 5 and beyond). Second, the normal approximation leads to a more intuitive interpretation of the difference $f - e$ between observed and expected frequency, and the amount of evidence against H_0 that it provides (the importance of a given raw difference value depends crucially on sample size and on the null hypothesis proportion π_0 , which makes it hard to compare across samples and experiments).

An interpretation of $f - e$ (or, equivalently $f - \mu$) that is comparable, e.g., between samples of different sizes, is achieved by a normalized value, the *z-score*, which divides $f - \mu$ by the standard deviation σ (you can think of this as expressing $f - \mu$ in σ 's, i.e., using σ as the “unit of measurement”):

$$z := \frac{f - \mu}{\sigma} \tag{5}$$

If two observations f_1 and f_2 (possibly coming from samples of different sizes and compared against different null hypotheses) lead to the same z-score $z_1 = z_2$, they are equally “extreme” in the sense that they provide the same amount of evidence against their respective null hypothesis (as given by the approximate p-values). To get a feel for this, refer back to Figure 36.3, which illustrates the approximate two-tailed p-value corresponding to $z = 2$ as a shaded area under the normal curve. This area has exactly the same size regardless of the specific shape of the curve implied by H_0 (in the form of the

parameters μ and σ). In other words, whenever we observe a value that translates into a z-score of $z = 2$ (according to the respective null hypothesis), we will obtain the same p-value from (the normal approximation to) the binomial test. Since we apply a two-tailed test, an observation that is two standard deviations to the left of the expected value ($z = -2$) will also lead to the same p-value.

Once an observation f has been converted into a z-score z , it is thus easy to decide whether H_0 can be rejected or not, by comparing $|z|$ with previously established thresholds for common significance levels α . For $\alpha = .05$, the (two-tailed) z-score threshold is 1.96, so the rejection criterion is $|z| \geq 1.96$; for $\alpha = .01$ the threshold is $|z| \geq 2.58$ and for $\alpha = .001$ it is $|z| \geq 3.29$. Thus, no matter what the original values of f , π and n are, if in an experiment we obtain a z-score of, say, $z = 2$ (meaning that f is two standard deviations away from e), we immediately know that the result is significant at the .05 significance level, but not at the .01 level. Statistics textbooks traditionally provide lists of z-score thresholds corresponding to various significance levels, although nowadays p-values for arbitrary z-scores can quickly be obtained from statistical software packages.

5. Two-sample tests

So far we have analyzed what is known as a *one-sample* statistical setting, where our null hypothesis concerns a certain quantity (often, the proportion of a certain phenomenon) in the set of all relevant units (e.g., all the sentences of English) and we use a sample of such units to see if the null hypothesis should be rejected. However, *two-sample* settings, where we have two samples (e.g., two corpora with different characteristics, or two sets of sentences of different types) and want to know whether they are significantly different with respect to a certain property, are much more common. Coming back to the example of passivization in idiomatic vs. non-idiomatic constructions from the introduction, our two samples would be sets of idiomatic and non-idiomatic constructions; we would count the number of passives in both sets; and we would verify the null hypothesis that there is no difference between the proportion of passives in the two samples.

It is easier to motivate one aspect of hypothesis testing that is often counter-intuitive, i.e., the fact that we pick as null hypothesis the “uninteresting” hypothesis that we hope to reject, when looking at the two-sample case. First, most linguistic theories, especially categorical ones, are more likely to predict that there is *some* difference between two sets, rather than making quantitative predictions about this difference being of a certain size. Second, in this way, if we can reject the null hypothesis, we can claim that the hypothesis that there is no difference between the groups is not tenable, i.e., that there *is* a difference between the groups, which is what our theory predicts. If, instead, we tested the null hypothesis that there is a certain difference between the groups, and we found that this hypothesis cannot be rejected, we could only claim that, for now, we have not found evidence that would lead us to reject our hypothesis: clearly, a weaker conclusion.

Probably the majority of questions that are of interest to linguists can be framed in terms of a two-sample statistical test: for several examples of applications in syntax, see article 43; for an application to the study of collocations, see article 58. Here, we discuss the example of the distribution of passives in two broad classes of written English, “in-

formative” prose (such as daily press) and “imaginative” prose (such as fiction). One plausible *a priori* hypothesis is that these two macro-genres will differ in passive ratios, with a stronger tendency to use passives in informative prose, due to the impersonal, more “objective” tone conferred to events by passive voice (for a more serious corpus-based account of the distribution of the English passive, including register-based variation, see Biber/Conrad/Reppen 1999, sections 6.4. and 11.3.). Our null hypothesis will be that there is no difference between the proportion of passives in informative prose π_1 and imaginative prose π_2 , i. e., $H_0: \pi_1 = \pi_2$. Conveniently, the documents in the Brown corpus are categorized into informative and imaginative writing – thus, we can draw random samples of $n_1 = 100$ sentences from the informative section of the corpus, and $n_2 = 100$ sentences from the imaginative section. Counting the passives, we find that the informative sample contains $f_1 = 23$ passives, whereas the imaginative sample contains $f_2 = 9$ passives.

Since there is a considerable difference between f_1 and f_2 , we are tempted to reject H_0 . However, before we can do so, we must find out to what extent the difference can be explained by random variation, i. e., we have to calculate how likely it is that the two samples come from populations with the same proportion of passives, as stated by the null hypothesis (statistics textbooks will often phrase the null hypothesis directly as: the samples *are* from the same population). In order to calculate expected frequencies, we have to estimate this common value from the available data, using maximum-likelihood estimation: $\hat{\pi} = (f_1 + f_2) / (n_1 + n_2) = 32 / 200 = 16\%$ (we sum the f 's and n 's because, if H_0 is right, then we can treat all the data we have as a larger sample from what, for our purposes, counts as the same population). Replacing H_0 by the more specific null hypothesis $H'_0: \pi_1 = \pi_2 = 16\%$, we can compute the expected frequencies under the null hypothesis, i. e., $e_1 = e_2 = 100 \cdot \hat{\pi} = 16$ (which are identical in our case since $n_1 = n_2 = 100$), as well as the binomial sampling distributions.

In the one-sample case, we looked at the overall probability of f and all other possible values that are more extreme than $|f - e|$. The natural extension to the two-sample case would be to look at the overall probability of the pair (f_1, f_2) and all the other possible pairs of values that, taken together, are more extreme than the sum of $|f_1 - e_1|$ and $|f_2 - e_2|$. The lower this probability, the more confident we can be that the null hypothesis is false. In our case, $|f_1 - e_1|$ and $|f_2 - e_2|$ are directly comparable and might be added up in this way, since the expected frequencies $e_1 = e_2 = 16$ and the sample sizes $n_1 = n_2 = 100$ are the same. However, in many real life situations, we will have to deal with samples of (sometimes vastly) different sizes (e. g., if one of the conditions is relatively rare so that only few examples can be found).

Fortunately, we know a solution to this problem from section 4: z-scores provide a measure of extremeness that is comparable between samples of different sizes. We thus compute the z-scores $z_1 = (f_1 - e_1) / \sigma_1$ and $z_2 = (f_2 - e_2) / \sigma_2$ (with σ_1 and σ_2 obtained from the estimate $\hat{\pi}$ according to H'_0). For mathematical reasons, the total extremeness is computed by adding up the squared z-scores $x^2 := (z_1)^2 + (z_2)^2$ instead of the absolute values $|z_1|$ and $|z_2|$. It should be clear that the larger this value is, the less likely the null hypothesis of no difference in population proportions is, and thus we should feel more confident in rejecting it. More precisely, the p-value associated with x^2 is the sum over the probabilities of all outcomes for which the corresponding random variable $X^2 := (Z_1)^2 + (Z_2)^2$ is at least as large as the observed x^2 , i. e. $\Pr(X^2 \geq x^2)$.

Instead of enumerating all possible pairs of outcomes with this property, we can again make use of the normal approximation, which leads to what is known as a *chi-squared*

distribution with one *degree of freedom* ($df = 1$). Using the chi-squared distribution, we can easily calculate the p-value corresponding to the observed x^2 , or compare x^2 with known rejection thresholds for different significance levels (e.g. $x^2 \geq 3.84$ for $\alpha = .05$ or $x^2 \geq 6.63$ for $\alpha = .01$). This procedure is known as (*Pearson's chi-squared test*) (Agresti 1996, section 2.4.; DeGroot/Schervish 2002, sections 9.1.–9.4.).

An alternative representation of the observed frequency data that is widely used in statistics takes the form of a *contingency table*:

	sample 1	sample 2	(6)
passives	f_1	f_2	
other	$n_1 - f_1$	$n_2 - f_2$	

The cells in the first row give the frequencies of passives in the two samples, while the cells in the second row give the frequencies of all other sentence types. Notice that each column of the contingency table adds up to the respective sample size, and that $\hat{\pi}$ (the estimated population proportion under H_0 needed to compute expected frequencies) can be obtained by summing over the first row and dividing by the overall total. Thus, the chi-squared statistic x^2 can easily be calculated from such a table (Agresti 1996, chapter 2; DeGroot/Schervish 2002, section 9.3.) and most statistical software packages expect frequency data for the chi-squared test in this form. Like in the one-sample case, the normal approximation is only valid if the sample sizes are sufficiently large. The standard rule of thumb for contingency tables is that all *expected* cell frequencies (under H_0) must be ≥ 5 (Agresti 1996, section 2.4.1.). In the usual situation in which $\hat{\pi} < 50\%$, this amounts to $n_1\hat{\pi} \geq 5$ and $n_2\hat{\pi} \geq 5$. Statistical software will usually produce a warning when the normal approximation is likely to be inaccurate.

There is also an *exact* test for contingency tables, similar to the binomial test in the one-sample case. This test is known as *Fisher's exact test* (Agresti 1996, section 2.6.). It is implemented in most statistical software packages, but it is computationally expensive and may be inaccurate for large samples (depending on the specific implementation). Therefore, use of Fisher's test is usually reserved for situations where the samples are too small to allow the normal approximations underlying the chi-squared test (as indicated by the rule of thumb above).

In the current example ($f_1 = 23$ and $f_2 = 9$), the contingency table corresponding to the observed data is

	informative	imaginative	(7)
passives	23	9	
other	77	91	

Using a statistical software package, we obtain $x^2 = 6.29$ for this contingency table, leading to rejection of H_0 at the $.05$ significance level (but not at the $.01$ level). The approximate p-value computed from x^2 is $p = 1.22\%$, while Fisher's exact test yields $p = 1.13\%$ (with expected frequencies $n_1\hat{\pi} = n_2\hat{\pi} = 16 \gg 5$, we anticipated a good agreement between the exact and the approximate test). We can thus conclude that there is,

indeed, a difference between the proportion of passives in informative vs. imaginative prose. Moreover, the direction of the difference confirms our conjecture that the proportion of passives is higher in informative prose.

A particular advantage of the contingency table notation is that it allows straightforward generalizations of the two-sample frequency comparison. One extension is the comparison of more than two samples representing different conditions (leading to a contingency table with $k > 2$ columns). For instance, we might want to compare the frequency of passives in samples from the six subtypes of imaginative prose in the Brown corpus (general fiction, mystery, science fiction, etc.). The null hypothesis for such a test is that the proportion of passives is the same for all six subtypes, i. e. $H_0: \pi_1 = \pi_2 = \dots = \pi_6$. Another extension leads to contingency tables with $m > 2$ rows. In our example, we have distinguished between passive sentences on one hand and all other types of sentences on the other. However, this second group is less homogeneous so that further distinctions may be justified, e. g., at least between sentences with intransitive and transitive constructions. From such a three-way classification, we would obtain three frequencies $f^{(p)}$, $f^{(i)}$ and $f^{(t)}$ for each sample, which add up to the sample size n . These frequencies can naturally be collected in a contingency table with three rows. The null hypothesis would now stipulate that the proportions of passives, intransitive and transitives are the same under both conditions (assuming $k = 2$), viz. $\pi_1^{(p)} = \pi_2^{(p)}$, $\pi_1^{(i)} = \pi_2^{(i)}$ and $\pi_1^{(t)} = \pi_2^{(t)}$. In general, an x^2 value can be calculated for any $m \times k$ contingency table by analogy to the 2×2 case. The p-value corresponding to x^2 can be obtained from a chi-squared distribution with $df = (m - 1)(k - 1)$ degrees of freedom. If the expected frequency in at least one of the cells is less than 5, a version of Fisher's exact test can be used (this version is considerably *more* expensive than Fisher's test for 2×2 tables, though).

Having established that the proportion of passives is different in informative vs. imaginative prose, we would again like to know how large the effect size is, i. e. by how much the proportions π_1 and π_2 differ. This is particularly important for large samples, where small (and hence linguistically irrelevant) effect sizes can easily lead to rejection of H_0 (cf. the discussion in section 3). A straightforward and intuitive measure of effect size is the difference $\delta := \pi_1 - \pi_2$. When the sample sizes are sufficiently large, normal approximations can be used to compute a confidence interval for δ . This procedure is often referred to as a *proportions test* and it is illustrated, for example, by Agresti (1996, section 2.2.). In our example, the 95% confidence interval is $\delta = 3.0\% \dots 25.0\%$, showing that the proportion of passives is at least three percentage points higher in informative prose than in imaginative prose (with 95% certainty).

In other situations, especially when π_1 and π_2 are on different orders of magnitude, other measures of effect size, such as the ratio π_1 / π_2 (known as *relative risk*) may be more appropriate. A related measure, the *odds ratio* θ , figures prominently because an exact confidence interval for θ can be obtained from Fisher's test. Most software packages that implement Fisher's test will also offer calculation of this confidence interval. In many linguistic applications (where π_1 and π_2 are relatively small), θ can simply be interpreted as an approximation to the ratio of proportions (relative risk), i. e., $\theta \approx \pi_1 / \pi_2$. On these measures see, again, section 2.2. of Agresti (1996). Effect size in general $m \times k$ contingency tables is much more difficult to define, and it is most often discussed in the setting of so-called *generalized linear models* (Agresti 1996, chapter 4).

Examples of fully worked out two-sample analyses based on contingency tables can be found in articles 43 and 58. As illustrated by article 43 in particular, contingency

tables and related two-sample tests can be tuned to a number of linguistic questions by looking at different kinds of linguistic populations. For example, if we wanted to study the distribution of *by*-phrases in passive sentences containing two classes of verbs (say, verbs with an agent vs. experiencer external argument), we could define our two populations as all passive sentences with verbs of class 1 and all passive sentences with verbs of class 2. We would then sample passive sentences of these two types, and count the number of *by*-phrases in them. As a further example, we might be interested in comparing alternative morphological and syntactic means to express the same meaning. For example, we might be interested, together with Rainer (2003), in whether various classes of Italian adjectives are more likely to be intensified by the suffix *-issimo* or by the adverb *molto*. This leads naturally to a contingency table for intensified adjectives with *-issimo* and *molto* columns, and as many rows as the adjective classes we are considering (or vice versa). The key to the successful application of statistical techniques to linguistic problems lies in being able to frame interesting linguistic questions in operational terms that lead to meaningful significance testing. The following section will discuss different ways to perform this operationalization.

6. Linguistic units and populations

As we just said, from the point of view of linguists interested in analyzing their data statistically, the most important issue is how to frame the problem at hand so that it can be operationalized in terms suitable for a statistical test. In this section, we introduce some concepts that might be useful when thinking of linguistic questions in a statistical way.

In the example used throughout the preceding sections, we have defined the population as the set of all (written American) English sentences and considered random samples of sentences from this population. However, statistical inference can equally well be based on any other linguistic unit, such as words, phrases, paragraphs, documents, etc. This *unit of measurement* is often called a *token* in corpus linguistics, at least when referring to words. Here, we use the term more generally to refer to any unit of interest.

The *population* then consists of all the utterances that have ever been produced (or could be produced) in the relevant (sub)language, broken down into tokens of the chosen type. We might also decide to focus on tokens that satisfy one or more other criteria and narrow down the population to include only these tokens. For instance, we might be concerned with the population of words that belong to a specific syntactic category; or with sentences that contain a particular verb or construction, etc.

What we are interested in is the *proportion* π of tokens (in the population) that have a certain additional property: e. g., word tokens that are nouns, verb tokens that belong to the inflectional paradigm of *to make*, sentences in the passive voice, etc. The properties used to categorize tokens for this purpose are referred to as *types* (in contrast to tokens, which are the categorized objects).

Since the full population is inaccessible, our conclusions have to be based on a (*random*) *sample* of tokens from the population. Such a sample of language data is usually called a *corpus* (or it can be a sub-corpus derived from a larger corpus: when we define the population as a set of verb tokens, for example, our sample might comprise all

instances of verbs found in the corpus). The *sample size n* is the total number of tokens in the sample, and the number of tokens that exhibit the property of interest (i. e., that belong to the relevant type) is the observed *frequency f*.

The same observed frequency can have different interpretations (with respect to the corresponding population proportion) depending on the units of measurement chosen as tokens, and the related target population. For instance, the number of passives in a sample could be seen relative to the number of sentences (π = proportion of sentences in the passive voice), relative to the number of verb phrases (π = proportion of passive verb phrases), relative to word tokens (π = relative frequency of passive verb phrases per 1,000 words), relative to all sentences containing transitive verbs (π = relative frequency of actual passives among sentences that could in principle be in passive voice). Note that each of these interpretations casts a different light on the observed frequency data. It is the linguist's task to decide which interpretation is the most meaningful, and to draw conclusions about the linguistic research questions that motivated the corpus study.

Other examples might include counting the number of deverbal nouns in a sample from the population of all nouns in a language; counting the number of words ending in a coronal stop in a sample from the population of all words in the language; counting the number of sentences with heavy-NP shift in a sample from the population of all sentences with a complement that could in principle undergo the process; counting the number of texts written in the first person in a sample from the population of literary texts in a certain language and from a certain period. Related problems can also be framed in terms of looking at *two* samples from distinct populations (cf. section 5), e. g., counting and comparing the number of deverbal nouns in samples from the populations of abstract and concrete nouns; counting the number of words ending in a coronal stop in samples from the population of all native words in the language and the population of loanwords; counting the number of texts written in the first person in samples from populations of texts belonging to two different literary genres.

In many cases, frequencies are computed not only for a single property, but for a set of mutually exclusive properties, i. e., a *classification* of the tokens into different types. In the two-sample setting this leads naturally to an $m \times 2$ contingency table (with the types in the classification as rows, and the two populations we are comparing as columns). Note that the classification has to be *complete*, so that the columns of the table add up to the respective sample sizes, which is often achieved by introducing a category labeled "other" (the single-property/two-samples cases above correspond to 2×2 contingency tables with an "other" class: e. g., deverbal vs. "other" nouns compared across the populations of abstract vs. concrete nouns).

As an example of a classification into multiple categories, word tokens might be classified into syntactic categories such as noun, verb, adjective, adverb, etc., with an "other" class for minor syntactic categories and problematic tokens. A chi-squared test might then be performed to compare the frequencies of these categories in samples from two genres. As another example, one might classify sentences according to the semantic class of their subject, and then compare the frequency of these semantic classes in samples of the populations of sentences headed by true intransitive vs. unaccusative verbs. It is not always obvious which characteristics should be operationalized as a classification of the tokens into types, and which should rather be operationalized in terms of different populations the tokens belong to. In some cases, it might make more sense to frame the task we just discussed in terms of the distribution of verb types across popula-

tions of sentences with different kinds of subjects, rather than vice versa. This decision, again, will depend on the linguistic question we want to answer.

In corpus linguistics, *lexical classifications* also play an important role. In this case, types are the distinct word forms or lemmas found in a corpus (or sequences of word forms or lemmas). Lexical classifications may lead to extremely small proportions π (sometimes measured in occurrences per million words) and huge differences between populations in the two-sample setting. Article 58 discusses some of the relevant methodologies in the context of collocation extraction.

The examples we just discussed give an idea of the range of linguistic problems that can be studied using the simple methods based on count data described in this article. Other problems (or the same problems viewed from a different angle) might require other techniques, such as those mentioned in the next two sections. For example, our study of passives could proceed with a *logistic regression* (see section 8), where we look at which factors have a significant effect on whether a sentence is in the passive voice or not. In any case, it will be fundamental for linguists interested in statistical methods to frame their questions in terms of populations, samples, types and tokens.

7. Non-randomness and the unit of sampling

So far, we have always made the (often tacit) assumption that the observed data (i.e., the corpus) are a random sample of tokens of the relevant kind (e.g., in our running example of passives, a sentence) from the population. Most plainly, we have compared a corpus study to drawing balls from an urn in section 2, which allowed us to predict the sampling distribution of observed frequencies. However, a realistic corpus will rarely be built by sampling individual tokens, but rather as a collection of contiguous stretches of text or even entire documents (such as books, newspaper editions, etc.). For example, the Brown corpus consists of 2,000-word excerpts from 500 different books (we will refer to these excerpts as “texts” in the following). The discrepancy between the *unit of measurement* (a token) and the *unit of sampling* (which will often contain hundreds or thousands of tokens) is particularly obvious for lexical phenomena, where tokens correspond to single words. Imagine the cost of building the Brown corpus by sampling a single word each from a million different books rather than 2,000 words each from only 500 different books!

Even in our example, where each token corresponds to an entire sentence, the unit of sampling is much larger than the unit of measurement: each text in the Brown contains roughly between 50 and 200 sentences. This need not be a problem for the statistical analysis, as long as each text is itself a random sample of tokens from the population, or at least sufficiently similar to one. However, various factors, such as the personal style of an author or minor differences in register or conventions within a particular subdomain, may have a *systematic* influence on how often passives are used in different texts. This means that the variability of the frequency of passives between texts may be much larger than between random samples of the same sizes (where all variation is purely due to chance effects).

Again, the problem is most apparent for lexical frequencies. Many content words (except for the most general and frequent ones) will almost only be found in texts that

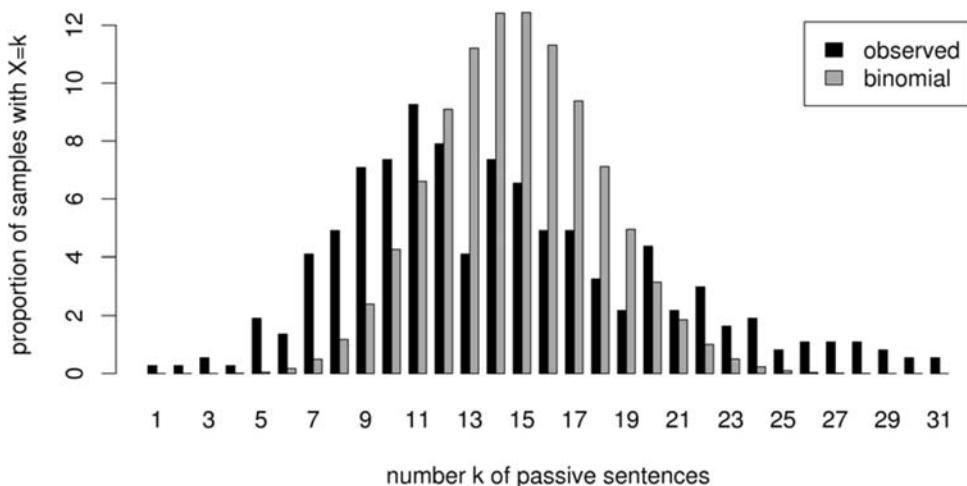


Fig. 36.5: Comparison of the frequencies of passives in the texts of the Brown corpus (informative prose only) with the binomial distribution predicted for random samples. In order to ensure comparability of the frequencies, 50 sentences were sampled from each Brown text

deal with suitable topics (think of nouns like *football* or *sushi*, or adjectives like *coronal*). On the other hand, such topical words tend to have multiple occurrences in the same text, even if these would be extremely unlikely in a random sample (indeed, the “burstiness” of words in specific texts is used as a strategy to find interesting keywords; see, e.g., Church 2000).

The increased variability of frequency between the individual texts is attenuated to some extent when corpus frequencies are obtained by summing over all the (randomly selected) texts in a corpus. However, in most cases the corpus frequencies will still show more variation than predicted by the binomial distribution.

In order to verify empirically whether a linguistic phenomenon such as the frequency of passives is subject to such non-randomness, we can compare the distribution of observed frequencies across the texts in a corpus with the distribution predicted for random samples by the binomial distribution. An example of such a comparison is shown in Figure 36.5. From each Brown text, we have taken a sample of 50 sentences (this subsampling step was necessary because the number of sentences per text varies from around 50 to more than 200). By tabulating the observed frequencies, we obtain the distribution shown as black bars in Figure 36.5. The gray bars show the binomial distribution that we would have obtained for random samples from the full population (the population proportion of passives was estimated at $\pi = 27.5\%$, based on the Brown data). Note that we have used only informative prose texts, since we already know from section 5 that the proportion of passives differs considerably between the two major sections of the corpus.

As the figure shows, the observed amount of variation is larger than the one predicted from the binomial distribution: look for example at the proportion of observed and binomial samples with $X \leq 7$. The standard deviation (which, as discussed in section 4, is a measure of the width of a distribution) is $\sigma = 6.63$ for the empirical distribution, but only $\sigma = 3.16$ for the binomial distribution. The corresponding z-scores, having σ

in the denominator (see Equation (5) in section 4), will be smaller for the empirical distribution, and thus the results are less significant than they would seem according to the binomial distribution. This means that the binomial test will lead to rejection of a true null hypothesis more easily than should be the case, given the spread of the actual distribution.

Suppose that we want to test the null hypothesis $H_0: \pi = 27.5\%$ (which is in fact true) based on a sample of $n = 50$ sentences from the informative prose in the Brown corpus. If the observed frequency of passives in this sample is $f = 7$, we feel confident to reject H_0 ($e = 13.75$ leads to a z-score of $z = -2.14$, above the $\alpha = .05$ threshold of $|z| \geq 1.96$). However, if all sentences in this sample came from the same text (rather than being sampled randomly from the entire informative prose section), Figure 36.5 shows that the risk of obtaining *precisely* $f = 7$ by chance is already around 4%! The “true” z-score (based on the standard deviation computed from the observed samples) is only $z = -1.02$, far away from any rejection threshold (in fact, this z-score indicates a risk of more than 30% that H_0 would be wrongly rejected).

Seeing how non-randomness effects can lead to a drastic misinterpretation of the observed frequency data, a question arises naturally: How can we make sure that a corpus study is not affected by non-randomness? While for many practical purposes it might be possible to ignore the issue, the only way to be absolutely sure is to ascertain that the unit of sampling coincides with the unit of measurement. When using a pre-compiled corpus (as will be the case for most studies in corpus linguistics) or when it would be prohibitively difficult and time-consuming to sample individual tokens, we have no choice but to adjust the *unit of measurement*. For example, when our data are based on the Brown corpus, the unit of sampling – and hence the unit of measurement – would be a text, i. e., a 2,000-word excerpt from a coherent document. Of course, we can no longer classify such an excerpt as “passive” or “non-passive”. Instead, what we observe for each token is a real-valued number: the proportion of passive sentences in the text.

Unlike previously, where each measurement was essentially a yes/no-decision (“passive” or “not passive”) or an m -way classification, measurements are now real-valued numbers that can in principle assume any value between 0 and 1 (3/407, 139/211, etc.). Statisticians speak of a *nominal scale* (for yes/no-decisions and classifications) vs. an *interval scale* (for numerical data). In order to analyze such data, we need an entirely different arsenal of statistical procedures, such as the well-known *t-test*. These methods are explained in any introductory textbook on statistics, and we give a brief overview of the most important terms and techniques in section 8.

This approach is only viable for phenomena, such as passive voice, that have a reasonably large number of occurrences in each text. It would not be sensible to count the proportion of occurrences of the collocation *strong tea* in the Brown texts (or even in a corpus made of larger text stretches), since the vast majority of texts would yield a proportion of 0% (in the Brown corpus, *strong tea* occurs exactly once, which means that in all texts but one the proportion will indeed be 0%).

Notice that, from a statistical perspective, the issues of representativeness and balance sometimes discussed in connection with corpus design (see article 9) involve two aspects: 1) How to define the target population precisely (is it possible to delimit a set of utterances that constitutes the population of, say, “contemporary English”?), and 2) how to take a random sample from the target population (with the complication discussed in

this section that what might constitute a random sample of, say, documents, will not be a random sample of, say, sentences). See Evert (2006) for an extended discussion of non-randomness in corpus linguistics.

8. Other techniques

Like for count data, there is a range of statistical tests that can be used to analyze data on an interval scale (such as the relative number of sentences containing passives per document discussed in the previous section, or reaction times from a psycholinguistic experiment). For the one-sample case, in which we want to test whether an observed interval-scale quantity (such as the proportion of passive sentences in a text) could plausibly come from a population where the same quantity has a certain distribution with a specific mean and standard deviation, you can use a (*one sample*) *t-test* (comparable to the binomial test for count data or, more precisely, the normal approximation based on z-scores). Unsurprisingly, when two samples are compared, the appropriate test is a *two-sample t-test* (corresponding to the chi-squared test for count data). However, in order to compare more than two samples, rather than performing a series of pairwise t-tests (a procedure that would make it much more likely that we obtain a significant result by chance), the technique to be applied is the *one-way analysis of variance (ANOVA)*. The ANOVA can only tell us whether at least one sample in the set is different from at least one other sample, and post-hoc tests must then be performed to identify the sample(s) responsible for rejection of H_0 .

In some settings, the variables in two samples have a natural pairing. For example, if we compare the proportion of passives in English and Spanish texts based on a parallel corpus, we should make use of the information that the texts are paired in order to control for irrelevant factors that may affect passive proportion (e.g., style and topic of a text), which should have a similar effect in an original and its translation. The appropriate test, in this case, is the *paired t-test*.

In many studies, it makes sense to operationalize the problem as one of assessing the association between two properties of the same unit, both measured on an interval scale. For example, we might be interested in the issue of whether there is a relation between the proportion of passives and, say, that of nominalizations (as both are plausible markers of more formal registers). Given a list of pairs specifying the (relative) frequencies of passives and nominalizations in each of the texts in our sample, we can perform a *correlation analysis*. In this case, the null hypothesis will be that there is no correlation between the two properties; and effect size will be measured in terms of how much of the variability of one variable can be explained by linear dependence on the other variable (standard correlation analysis will not capture *nonlinear* relations between variables).

A significant correlation does not imply a causal relation between two variables (even if the numbers of passives and nominalizations turn out to be correlated, it is unlikely that passives “cause” nominalization or vice versa). Often, however, we want to go beyond the mere observation of a relationship between two variables. If we hypothesize that the behavior of a certain variable depends on that of one or more other variables, we will want to use statistics to test whether our *independent* variables predict the values

of the *dependent* variable beyond chance level. In this case, we use the technique of (*multiple*) *linear regression* (which is related to correlation). In linear regression, the independent variables can be a mixture of discrete and continuous variables, but the dependent variable must be continuous.

Similar techniques can also be applied to the analysis of the kind of categorical data (resulting in a contingency table of frequency counts) that have been the focus of this article. The equivalent of linear regression in this case is *logistic regression*. For example, a logistic regression analysis could try to predict whether a sentence will be in passive voice or not (a dichotomous dependent variable) in terms of factors such as the semantic class of the verb (a categorical variable), the overall entropy of the sentence (a continuous variable), etc. A full regression analysis tests significant effects of the independent variables, but typically it also checks that the independent variables are not correlated with each other, and it might look for the optimal combination of independent variables.

The cases we listed here (detection of differences and estimation of population values in one/two/multiple paired/non-paired sample cases, assessment of association/correlation between variables, regression) constitute a nearly exhaustive survey of the analytical settings considered in inferential statistics. More advanced techniques, rather than introducing completely new scenarios, will typically deal with cases in which one or more of the assumptions of the basic models are not met or the models need to be extended. For example, more sophisticated ANOVA models can take multiple categorizations of the data and their interactions into account (akin to the analysis of $m \times k$ contingency tables with m and/or k greater than 2). Advanced regression techniques can detect non-linear relations between the dependent and independent variables. So-called “distribution-free” tests make no assumption about the distribution of the underlying population(s) nor about the sampling distribution (these are typically referred to as *non-parametric methods*). Simulation-based methods (*Monte Carlo methods*, the *Bootstrap*) provide an alternative to analytical estimation of various parameters. A wealth of exploratory and visual methods are available to evaluate the validity of assumptions and the quality of the resulting models. *Bayesian inference*, a very important branch of statistics, allows, among other things, to distinguish between “more plausible” and “less plausible” estimates within a confidence interval (the classic binomial confidence interval described in section 3 indicates a range of plausible values for the population proportion, but it does not distinguish among these values, whereas, intuitively, we would consider the MLE proportion much more plausible than, say, the values at the edges of the confidence interval).

Some important kinds of corpus data, such as distributions of word types, are characterized by the presence of a very large number of very rare types (words that occur only once or never at all in the corpus at hand) and few extremely frequent types (function words). These extremely skewed distributions make the application of standard statistical models to certain tasks problematic (mainly, estimating the number of word types in a population as well as related quantities), and demand specialized statistical tools. For a general survey of the problems involved, see article 37 and the references on statistical modeling of word frequency distributions recommended there.

Almost every elementary statistics textbook (including those listed in the next section) will introduce t-tests, ANOVA, correlation and regression. Advanced techniques are nowadays within easy reach of non-statisticians thanks to their implementation in user-friendly software packages. Here we would like to stress once more that, for all of the

large variety of available procedures and their complications, the basic logic of hypothesis testing and estimation is essentially the same as what we illustrated with very simple examples of frequency count data in the first sections of this article. It is not essential to know the mathematical details of all the techniques in order to apply them, but it is important to understand the basic principles of hypothesis testing and estimation; the assumptions of a test, its null hypothesis, and the meaning of a p-value; and to make sure that the assumptions are met by the data and that the research question can be translated into a meaningful null hypothesis. And, of course, the linguistic interpretation of the statistical results is at least as crucial as the correctness of the methods applied.

We have focused here on statistical inference for hypothesis testing and estimation, as applied to corpus data. This is only a part, albeit a fundamental one, of the role that statistical methods play in corpus-related disciplines today. For a survey of statistical procedures used for *exploratory* purposes (i.e., as an aid in uncovering interesting patterns in the data), see articles 38 and 40. Statistical methods also play a very important role as *modeling* tools for machine learning techniques applied to natural language (article 39) and more generally in so-called empirical natural language processing (see, e.g., article 56 on machine translation, and Manning/Schütze 1999 for an introduction to statistical NLP).

9. Directions for further study

A much more in-depth introduction to the statistical inference methods appropriate for count data which we discussed here is provided by Agresti (1996) or, at a more technical level, Agresti (2002). There is, of course, a vast number of introductory general statistics books. DeGroot/Schervish (2002) present a particularly thorough and clear introduction, although it requires at least a basic mathematical background. Among the less technical introductions, we recommend the one by Hays (1994), a book that provides non-mathematical but rigorous explanations of the most important notions of statistical inference (although it focuses on statistical methods for the analysis of experimental results, which are only partially relevant to corpus work). There is also a wealth of statistics “cookbooks” that illustrate when to apply a certain technique, how to apply it, and how to interpret the results. These are often and usefully linked to a specific statistical software package. For example, Dalgaard (2002) is an introduction to running various standard statistical procedures in R (see below).

There are a few older introductions to statistics explicitly geared towards linguists. The one by Woods/Fletcher/Hughes (1986) is a classic, whereas the one by Butler (1985) now has the advantage of being freely available on the Web: <http://www.uwe.ac.uk/hlss/las/statistics-in-linguistics/bkindex.shtml>

Older introductions tend to focus on techniques that are more relevant to psycholinguistics, phonetics and language testing than to corpus analysis. Oakes (1998) presents a survey of applications of statistics in corpus studies that trades depth for wider breadth of surveyed applications and methods. It is likely that, with the growing interest in corpora and statistical approaches to linguistics in general, the next few years will see the appearance of more statistics textbooks targeting corpus linguists.

There is nowadays a large number of statistical software packages to choose from. We recommend R: <http://www.r-project.org/>

R supports an impressive range of statistical procedures and, being open-source and available free of charge, it is attracting a growing community of developers who add new functionalities, including some that are of interest to corpus linguists. These extensions range from advanced data visualization techniques to modules explicitly targeting corpus work. We illustrate some of the relevant functionalities in the tutorial available at: http://purl.org/stefan.evert/SIGIL/potsdam_2007

The corpora library (also free and open-source) provides support for carrying out the statistical analyses described in this article (the Web site has a tutorial that shows how to run them), as well as several sample data sets. There is an increasing number of introductory textbooks with concrete R examples, including some that focus on statistical methods in linguistics (Baayen 2008, Gries 2008). Shravan Vasishth has written (and is constantly updating) an online book aimed at (psycho-)linguists that introduces statistics in R through a simulation approach. This book is freely available (under a Creative Commons license) from: <http://www.ling.uni-potsdam.de/~vasishth/SFLS.html>

Finally, Wulff (2005) provides a survey of online statistics facilities.

10. Literature

- Agresti, Alan (1996), *An Introduction to Categorical Data Analysis*. Chichester: Wiley.
- Agresti, Alan (2002), *Categorical Data Analysis*. Second Edition. Chichester: Wiley.
- Baayen, R. Harald (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.
- Baayen, R. Harald (2008), *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics*. Cambridge: Cambridge University Press.
- Biber, Douglas/Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (1999), *Longman Grammar of Spoken and Written English*. Harlow, UK: Pearson Education.
- Butler, Christopher (1985), *Statistics in Linguistics*. Oxford: Blackwell.
- Chomsky, Noam (1986), *Knowledge of Language: Its Nature, Origins, and Use*. New York: Praeger.
- Church, Kenneth (2000), Empirical Estimates of Adaptation: The Chance of Two Noriega's is Closer to $p/2$ than p^2 . In: *Proceedings of the 17th Conference on Computational Linguistics*. Saarbrücken, Germany, 180–186.
- Culicover, Peter/Jackendoff, Ray (2005), *Simpler Syntax*. Oxford: Oxford University Press.
- Dalgaard, Peter (2002), *Introductory Statistics with R*. New York: Springer.
- DeGroot, Morris/Schervish, Mark (2002), *Probability and Statistics*. Third Edition. Boston: Addison-Wesley.
- Evert, Stefan (2006), How Random is a Corpus? The Library Metaphor. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177–190.
- Gries, Stefan Th. (2005), Null-hypothesis Significance Testing of Word Frequencies: A Follow-up on Kilgarriff. In: *Corpus Linguistics and Linguistic Theory* 1, 277–294.
- Gries, Stefan Th. (2008), *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge.
- Hays, William (1994), *Statistics*. Fifth Edition. New York: Harcourt Brace.
- Manning, Christopher D./Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- McEnery, Tony/Wilson, Andrew (2001), *Corpus Linguistics*. Second Edition. Edinburgh: Edinburgh University Press.
- Oakes, Michael (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

- Pearson, Egon (1990), '*Student*': *A Statistical Biography of William Sealy Gosset*. Oxford: Clarendon Press.
- Rainer, Franz (2003), Studying Restrictions on Patterns of Word-formation by Means of the Internet. In: *Rivista di Linguistica* 15, 131–139.
- Schütze, Carson (1996), *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
- Woods, Anthony/Fletcher, Paul/Hughes, Arthur (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Wulff, Stefanie (2005), Online Statistics Labs. In: *Corpus Linguistics and Linguistic Theory* 1, 303–308.

Marco Baroni, Trento (Italy) and Stefan Evert, Osnabrück (Germany)

37. Distributions in text

1. Introduction
2. Distributions
3. Zipf(-Mandelbrot)'s law
4. Practical consequences
5. Conclusion
6. Acknowledgments
7. Literature

1. Introduction

The frequency of words and other linguistic units plays a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language. Thus, not surprisingly, the distribution of frequencies of words and combinations of words in corpora has played a central role in the debate between proponents and detractors of the corpus-based approach (see, e.g., Abney 1996). One would then expect that the study of word frequency distributions plays a central role in the corpus linguistics curriculum. This is not the case. The standard introductions to the field (e.g., Biber/Conrad/Reppen 1998; McEnery/Wilson 2001) do not discuss the topic at all, and even an introduction explicitly geared towards the statistical aspects of the discipline, such as Oakes (1998), mentions Zipf's law (see section 3 below) only in passing (pp. 54–55).

This state of affairs may be due to the fact that the study of word frequency distributions originated outside mainstream linguistics. George Kingsley Zipf, undoubtedly the father of *lexical statistics* (the study of word frequency distributions), was trained as a philologist and considered himself a “human ecologist”. Other important pioneers of the field were the psychologist George Miller, the mathematician Benoit Mandelbrot (of Mandelbrot set fame) and the Nobel Prize winning economist Herbert Simon. Thus, the argumentations and terminology found in the early literature often sound rather exotic

- Pearson, Egon (1990), '*Student*': *A Statistical Biography of William Sealy Gosset*. Oxford: Clarendon Press.
- Rainer, Franz (2003), Studying Restrictions on Patterns of Word-formation by Means of the Internet. In: *Rivista di Linguistica* 15, 131–139.
- Schütze, Carson (1996), *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
- Woods, Anthony/Fletcher, Paul/Hughes, Arthur (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Wulff, Stefanie (2005), Online Statistics Labs. In: *Corpus Linguistics and Linguistic Theory* 1, 303–308.

Marco Baroni, Trento (Italy) and Stefan Evert, Osnabrück (Germany)

37. Distributions in text

1. Introduction
2. Distributions
3. Zipf(-Mandelbrot)'s law
4. Practical consequences
5. Conclusion
6. Acknowledgments
7. Literature

1. Introduction

The frequency of words and other linguistic units plays a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language. Thus, not surprisingly, the distribution of frequencies of words and combinations of words in corpora has played a central role in the debate between proponents and detractors of the corpus-based approach (see, e.g., Abney 1996). One would then expect that the study of word frequency distributions plays a central role in the corpus linguistics curriculum. This is not the case. The standard introductions to the field (e.g., Biber/Conrad/Reppen 1998; McEnery/Wilson 2001) do not discuss the topic at all, and even an introduction explicitly geared towards the statistical aspects of the discipline, such as Oakes (1998), mentions Zipf's law (see section 3 below) only in passing (pp. 54–55).

This state of affairs may be due to the fact that the study of word frequency distributions originated outside mainstream linguistics. George Kingsley Zipf, undoubtedly the father of *lexical statistics* (the study of word frequency distributions), was trained as a philologist and considered himself a “human ecologist”. Other important pioneers of the field were the psychologist George Miller, the mathematician Benoit Mandelbrot (of Mandelbrot set fame) and the Nobel Prize winning economist Herbert Simon. Thus, the argumentations and terminology found in the early literature often sound rather exotic

to linguists (e.g., Mandelbrot's "temperature of discourse" approach). Still today, most articles about lexical statistics appear in relatively obscure journals and they are often rooted in traditions, in particular that of the former Soviet Union, that are not well known in the English-centered world of corpus linguistics (Sampson 2002). The heavy involvement of non-linguists in the study of lexical statistics continues to this day. Judging from the affiliations of the authors of the recent *Glottometrics* volumes in honor of Zipf, word frequency distributions are more of interest to theoretical physicists than to theoretical linguists. The publication of Baayen (2001), a thorough introduction to lexical statistics that summarizes much of the earlier work, but recasts problems and solutions in the perspective of modern corpus/computational linguistics, will probably contribute to give more prominence to the domain.

This article introduces some of the empirical phenomena pertaining to word frequency distributions and the classic models that have been proposed to capture them. In particular, section 2 introduces the basic analytical tools and discusses the patterns typically encountered in corpora/texts. Section 3 presents Zipf-Mandelbrot's law, the most famous model proposed to account for frequency distributions. Section 4 shortly reviews some of the practical consequences and applications of frequency distribution modeling. Section 5 concludes by suggesting some directions for further study.

2. Distributions

2.1. Counting tokens and types

In order to study word frequency distribution, we must first of all count all the instances (*tokens*) of all distinct words (*types*) that occur in the corpus of interest (I use the term corpus in the most general way, to refer to any text or collection of texts that is the object of a linguistic study). Neither deciding what must be counted as a token, nor mapping tokens to types are trivial tasks. Consider the following mini-corpus:

The woman went to Long Beach and to Anaheim on bus number 234. However,
the man didn't go.

First, we will have to decide whether punctuation marks are tokens or not and whether to keep or remove strings containing digits. Both choices affect the shape of frequency distributions (punctuation marks are few and very frequent, numbers are many and rare). Next, we face a number of token segmentation problems. For example, we must decide whether we should split *didn't* into two words (and if we do, where do we split it). Moreover, *Long Beach* should perhaps be counted as a single word. Again, these choices will affect our counts in a systematic way. Having decided which strings to ignore, and how to segment the remaining text, we can count the tokens in the corpus. For example, if we decide to ignore punctuation and numbers, to treat *Long Beach* as two words and *didn't* as a single word, the mini-corpus above will have 17 tokens: *The, woman, went, to, Long, Beach, and, to, Anaheim, on, bus, number, However, the, man, didn't, go.*

Now, we must map each word token to a word type. In order to do this, we have to decide whether our counts should be sensitive to the distinction between upper and lower case or not: intuitively, *The* and *the* in the mini-corpus above should be counted as instances of the same word, but it would be wrong to treat the parts of the name *Long Beach* as instances of the adjective *long* and noun *beach*, respectively. In English, ignoring the distinction between upper and lower case will have distorting effects on proper name counts, but by preserving case distinctions we will duplicate word types that occur both in sentence-initial position and elsewhere. If we distinguish between upper and lower case, the mini-corpus tokenized as above will contain 16 types, one of them (*to*) represented by two tokens.

If we have the relevant resources (most importantly, a list of word-form/lemma correspondences), we can map tokens to lemma types. In the mini-corpus above, *went* and *go* would be treated as tokens of the same lemma type. On the one hand, more sophisticated tokenization/type mapping steps are likely to lead to cleaner counts. On the other, the errors and imprecisions inherent in any form of automated pre-processing can have a serious distorting effect on the data. For example, if all the words that are not recognized by our lemmatizer are mapped to a type *unknown*, we will transform many low frequency items into a single artificial high frequency type.

In the corpora analyzed in this article, unless stated otherwise, punctuation marks, strings containing digits and strings made entirely of non-alphabetic characters are not counted as tokens; all other white-space or punctuation-delimited strings constitute separate tokens (in English, some special strings are split into multiple tokens – e.g., *wouldn't* is tokenized as *would*, *n't*); upper- and lower-case types and not merged; lemmatization is not performed. The token and type counts I report are based on this tokenization/type mapping scheme. Issues related to corpus pre-processing, tokenization and lemmatization are discussed in articles 24 and 25 of this handbook.

2.2. The basic tools of lexical statistics

Once we have tokenized a corpus and mapped each token to a type, we can count the number of tokens in the corpus, or *corpus size* (N), and the number of types, or *vocabulary size* (V). For example, in the mini-corpus above, given the tokenization and type mapping rules I adopted, N is 17 and V is 16.

The starting point for any further analysis will be a *frequency list*, i.e., a list that reports the number of instances (tokens) of each word (type) that we encountered in the corpus. Consider for example the toy frequency list in Table 37.1.

Tab. 37.1: A toy frequency list

type	f	type	f
again	2	he	1
and	3	her	1
another	1	that	2
bark	1	this	1
barks	6	will	1
dog	3	with	1
friends	1		

The data in a frequency list can be re-organized in two ways that are particularly useful to study word frequency distributions: as *rank/frequency profiles* and as *frequency spectra*. To obtain a rank/frequency profile, we simply replace the types in the frequency list with their frequency-based ranks, by assigning rank 1 to the most frequent type, rank 2 to the second most frequent word, etc. In the example of Table 37.1, *barks* would be assigned rank 1, *and* and *dog* would be assigned rank 2 and 3 (ranking of words with the same frequency is arbitrary), etc. This produces the rank/frequency profile in Table 37.2.

Tab. 37.2: A toy rank/frequency profile

r	f	r	f
1	6	8	1
2	3	9	1
3	3	10	1
4	2	11	1
5	2	12	1
6	1	13	1
7	1		

A frequency spectrum is a list reporting how many types in a frequency list have a certain frequency. The spectrum corresponding to the frequency information in Table 37.1 is presented in Table 37.3.

Tab. 37.3: A toy frequency spectrum

f	V(f)
1	8
2	2
3	2
6	1

The first row of Table 37.3 tells us that there are 8 words with frequency 1 ($V(1) = 8$; *another*, *bark*, *friends*, *he*, *her*, *this*, *will*, *with*). The second row tells us that there are 2 words with frequency 2 ($V(2) = 2$; *again*, *that*), etc.

A rank/frequency profile and the corresponding frequency spectrum contain the same information, and it is thus possible to derive one from the other. However, as we will see, rank/frequency profiles are particularly useful to study the properties of high frequency items and frequency spectra are useful to study the properties of low frequency items.

2.3. Typical frequency patterns

Table 37.4 shows the top and bottom ranks and corresponding frequencies in the Brown corpus of American English (see article 20).

The top ranks are occupied by function words such as *the*, *of* and *and*. Frequency decreases quite rapidly: the most frequent word is almost twice as frequent as the second most frequent word. The difference in frequency becomes less dramatic as we go down

Tab. 37.4: Top and bottom of the Brown frequency list

top frequencies			bottom frequencies		
rank	fq	word	rank range	fq	Randomly selected examples
1	62642	the	7967– 8522	10	recordings undergone privileges
2	35971	of	8523– 9236	9	Leonard indulge creativity
3	27831	and	9237–10042	8	unnatural Lolotte authenticity
4	25608	to	10043–11185	7	diffraction Augusta postpone
5	21883	a	11186–12510	6	uniformly throttle agglutinin
6	19474	in	12511–14369	5	Bud Councilman immoral
7	10292	that	14370–16938	4	verification gleamed groin
8	10026	is	16939–21076	3	Princes nonspecifically Arger
9	9887	was	21077–28701	2	blitz pertinence arson
10	8811	for	28702–53076	1	Salaries Evensen parentheses

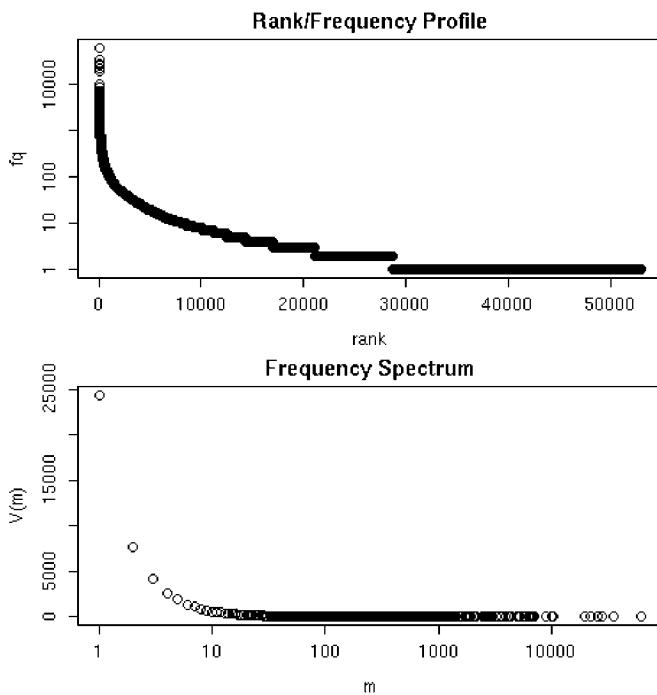


Fig. 37.1: Rank/frequency profile and frequency spectrum of the Brown corpus

the list, but the ranks are still spread across a wide frequency range. Because of their very high frequencies, the 10 top-ranked word types alone account for about 23 % of the total token count in the Brown (232,425 occurrences over 996,883 tokens in total). This

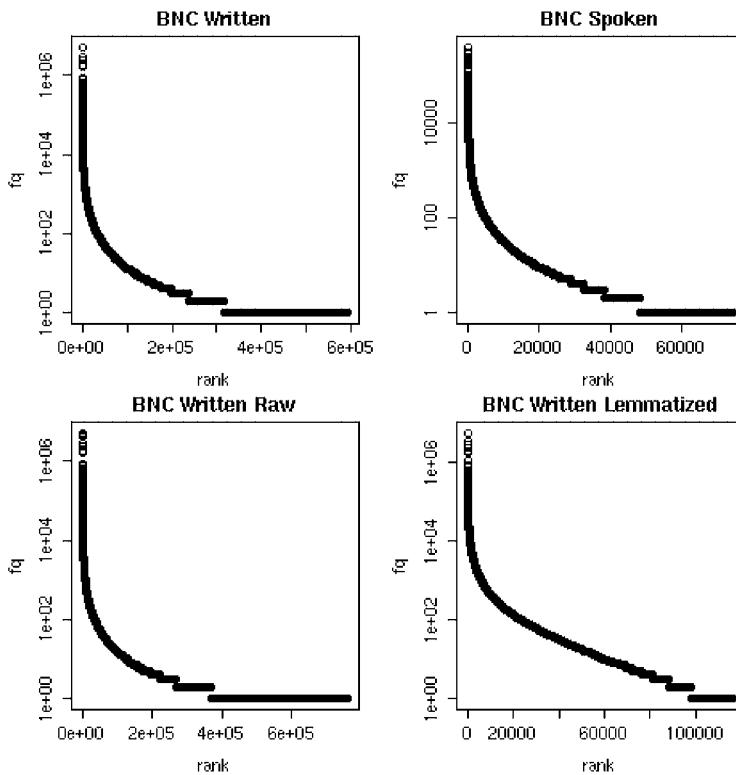


Fig. 37.2: Rank/frequency profiles of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right)

is to say that in the Brown more than one word in five comes from the set *the, of, and, to, a, in, that, is, was, for*.

The picture is very different at the bottom of the list, where there are massive frequency ties, and more ties as the frequency decreases: for example, there are 4,137 words with frequency 3 (ranks from 16939 to 21076), 7,624 words with frequency 2 (ranks from 21077 to 28701), 24,374 words with frequency 1 (ranks from 28702 to 53076). Since the Brown corpus contains 53,076 distinct types in total, the words occurring once constitute almost half of its vocabulary. The words occurring 3 times or less constitute almost 70% of the vocabulary. At the same time, this 70% of types account for only about 5% of the overall Brown token count (52,033 tokens over 996,883 total tokens). The lowest frequency elements are of course content words. As the random examples reported in the table show, not all the lowest frequency words are neologisms, new derivations or exotic forms. For example, words such as *pertinence* and *parentheses* are probably not going to strike the average English speaker as new or unusual.

The dichotomy between the extremely high token frequency of the most frequent types and the large number of low frequency types affects the classic summary statistics in peculiar ways. The average frequency of word types in the Brown is of 19 tokens.

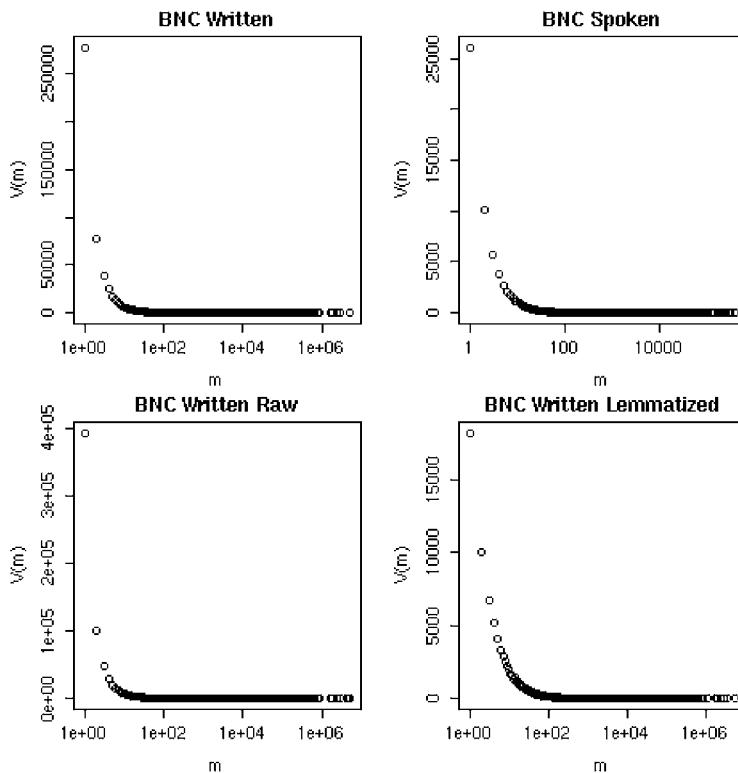


Fig. 37.3: Frequency spectra of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right)

However, this value is inflated by the very high frequencies of the most common words: more than 90 % of the types in the Brown corpus have frequency lower than the average. The median value is 2 (i.e., 50% types have frequency greater than or equal to 2, and 50 % types have frequency less than or equal to 2). The mode (the most common value), of course, is 1.

The upper panel of Figure 37.1 illustrates the rank/frequency profile of the Brown corpus. Frequency (on the y axis) is plotted on a logarithmic scale, because the frequency of the most frequent words is so much higher than the frequency of the long tail of rare words that a figure of this size without a logarithmic transformation would look like the letter L. The plot illustrates very clearly what we already observed: the frequency curve decreases very steeply from the extremely high values corresponding to the most frequent words, and it becomes progressively flatter, until it reaches a very wide plateau in correspondence to the ranks assigned to the tail of words occurring once (increasingly narrower plateaus corresponding to words occurring 2, 3, 4 times etc. are also visible). The lower panel of Figure 37.1 plots the frequency spectrum of the Brown (again, token frequency – this time on the x axis – is on a logarithmic scale). The lowest frequency classes are represented by a very large (and rapidly decreasing) number of types (the

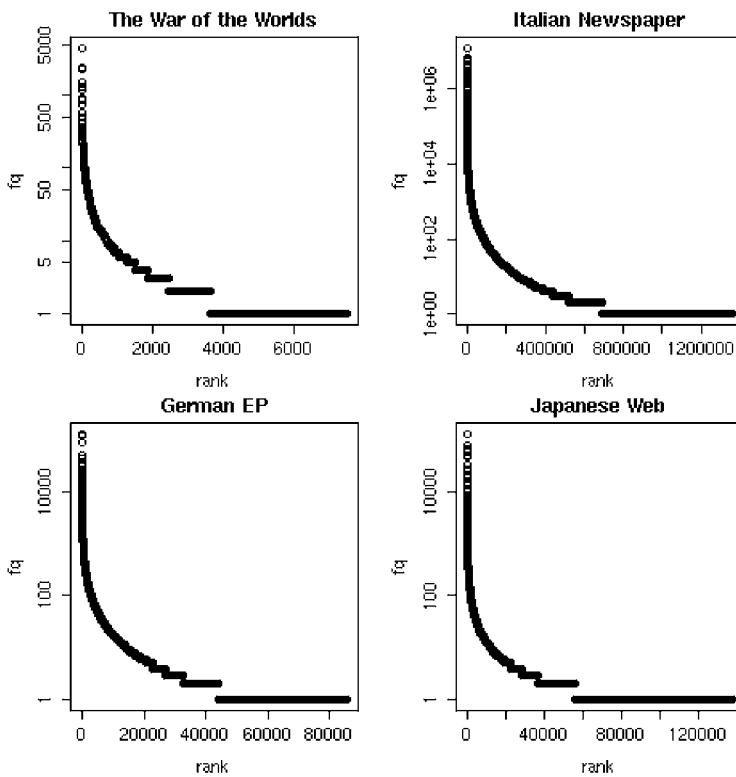


Fig. 37.4: Rank/frequency profiles of *The War of the Worlds* (top left), the Italian *la Repubblica* corpus (top right), a section of the German *EuroParl* corpus (bottom left) and a corpus of Japanese web-pages (bottom right)

types that occur once, the types that occur twice, etc.), and there is a long tail of high frequency classes represented by only 1 or 0 types.

The frequency distribution of the Brown is not specific to this corpus, but typical of natural language texts, independently of tokenization/type mapping method, size, language, textual typology, etc. To illustrate this, let us consider the British National Corpus (BNC, see article 20), which differs from the Brown in that it represents British rather than American English, it is based on more recent texts, it includes a spoken language section and, perhaps most importantly, it is much larger. The Brown contains about one million tokens, whereas the written section of the BNC contains 86,480,906 tokens, and the spoken section contains 10,423,654 tokens.

Figures 37.2 and 37.3 present rank/frequency profiles and frequency spectra for the BNC. The top two panels of Figure 37.2 show the rank/frequency profiles of the BNC written and spoken sections, respectively. The top two panels of Figure 37.3 show the corresponding spectra. The overall pattern is very similar to the one we observed in the Brown: few very frequent words, many low frequency words. This second fact is perhaps surprising: one could reasonably expect that in a very large sample of a language the words that are encountered only once become a minority. This is obviously not the case: in the written section of the BNC, the words occurring only once account for 46% of

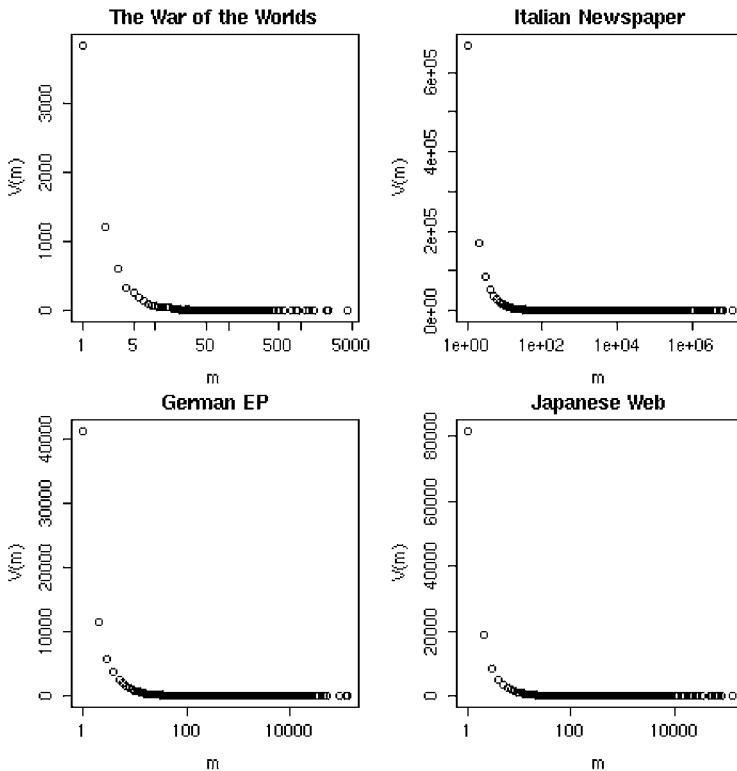


Fig. 37.5: Frequency spectra of *The War of the Worlds* (top left), the Italian *la Repubblica* corpus (top right), a section of the German *EuroParl* corpus (bottom left) and a corpus of Japanese web-pages (bottom right)

all the types, and the proportion of words occurring 3 times or less is of 66%. In the spoken section, these proportions are smaller (perhaps suggesting less lexical variety in speech?) but still very significant: 35% of the types occur only once and 56% occur 3 times or less. The mean token frequency of types in the written BNC is of about 146 tokens but more than 95% of the types have a frequency below this value. Like in the Brown, the median is 2 and the mode is 1. Corpus after corpus, we find that the mean is a value much higher than the median (and, as is intuitive, it increases in function of corpus size), the median is 2 or 1 and the mode is 1. Thus, the mean is not a meaningful indicator of central tendency, whereas the median and the mode are not very interesting since they tend to have the same values in all corpora. The third panels of Figures 37.2 and 37.3 show the rank/frequency profile and frequency spectrum in a version of the written BNC in which strings containing digits and other non-alphabetic symbols were counted as regular words. Again, we encounter a very similar pattern. Not surprisingly, the portion of the distribution taken by words occurring only once is even more prominent. The bottom right panels of Figures 37.2 and 37.3 report the rank/frequency profile and frequency spectrum of the lemmas in the written BNC. Although the number of very low frequency forms is lower than in the non-lemmatized counterpart (top left panels), the overall pattern is essentially the same, which shows that such pattern cannot

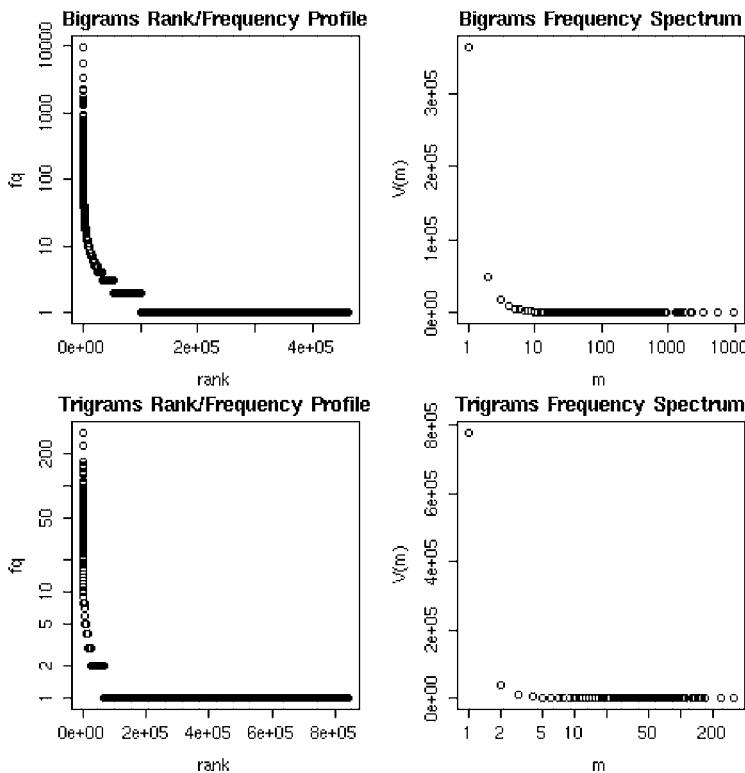


Fig. 37.6: Rank/frequency profiles and frequency spectra of the bigrams (top) and trigrams (bottom) in the Brown corpus

be simply explained in terms of the presence of inflected forms in non-lemmatized corpora.

Figures 37.4 and 37.5 display rank/frequency profiles and frequency spectra for four more texts/corpora of very different kinds. The top left panels present data from *The War of the Worlds*, the famous H. G. Wells novel from 1898, which, unlike the Brown or the BNC, is a “corpus” made of a single, coherent text. Moreover, compared to the other corpora analyzed here, it is very small, being comprised of 60,160 tokens. The top right panels present data from the *la Repubblica* corpus, containing (at the time in which the data used here were extracted) 325,290,035 tokens of Italian newspaper text (see Baroni et al. 2004). The bottom left panels present data from the year-2002 section of the German EuroParl corpus (see article 20), collecting transcriptions of European Parliament proceedings. This corpus contains 3,090,142 tokens. Finally, the bottom right panels present data from a corpus of Japanese web pages collected in 2004 with the method described in Ueyama/Baroni (2006) and tokenized with the ChaSen system (Matsumoto et al. 2000). It contains 2,175,736 tokens. Despite the obvious differences among these corpora, the rank/frequency profiles and the frequency spectra reveal strikingly similar overall patterns, in turn resembling those that we encountered in the Brown and BNC: few very high frequency types, and long tails of very rare words.

The same skewed shape also emerges if instead of looking at words we look at sequences of words, or *ngrams*, such as *bigrams* or *trigrams* (sequences of two and three words, respectively). Such distributions are even more skewed than those of words (given that the potential vocabulary of possible ngrams is much higher). This is illustrated for the Brown corpus in Figure 37.6. Among the trigrams, the types with frequency 1 constitute 92% of the vocabulary!

The distribution of word and ngram frequencies is rather different from the typical count distributions that are studied in introductory statistics classes. For example, if we divide the male students of a certain high-school into classes based on their height, we expect that most students will fall into the medium class, fewer students will be classified as tall or short, and very few students will turn out to be extremely tall or extremely short. The distribution of words is akin to finding a population made of few giants, rather few people in the medium height range and an army of dwarves.

3. Zipf(-Mandelbrot)'s law

The typical skewed structure of word frequency distributions was first systematically studied by Zipf (1949, 1965), who observed in various data-sets that frequency is a non-linearly decreasing function of rank (decreasing more sharply among high ranks than among low ranks), and proposed the following model, which became known as *Zipf's law*, to predict the frequency of a word given its rank:

$$f(w) = \frac{C}{r(w)^a} \quad (1)$$

In this formula, $f(w)$ and $r(w)$ stand for frequency and rank of word w , respectively. C and a are constants to be determined on the basis of the available data. To understand why this is a plausible model, assume for now that $a = 1$ (but the same point could be illustrated with other values of this parameter), so that equation (1) can be simplified to

$f(w) = \frac{C}{r(w)}$. Then, the most frequent word in the corpus, having rank 1, must have

frequency C . Suppose that in a certain corpus we find that the most frequent word has frequency 60,000 and thus we set $C = 60,000$. The second most frequent word is predicted to have frequency $C/2 = 30,000$, half the frequency of the first word. The third most frequent word will have frequency $C/3 = 20,000$, one third of the first word. On the other hand, the 100th most frequent word (the word with rank 100) will have frequency $C/100 = 600$. The 101st most frequent word will have frequency $C/101 = 594.06$, i. e. about 99% of the frequency of the 100th word. The 102nd most frequent word will have frequency $C/102 = 588.23$, about 98% of the frequency of the 100th word. Thus, the model predicts a very rapid decrease in frequency among the most frequent words, which becomes slower as the rank grows, leaving very long tails of words with similar low frequencies. Zipf's law does not predict frequency ties, since there are no ties among ranks, but it approximates the empirically attested plateaus by predicting a very large number of words with very similar non-integer frequencies. For example, the model above with a set to 1 and C set to 60,000 predicts that about 80,000 words will have frequencies between 1.5 and 0.5!

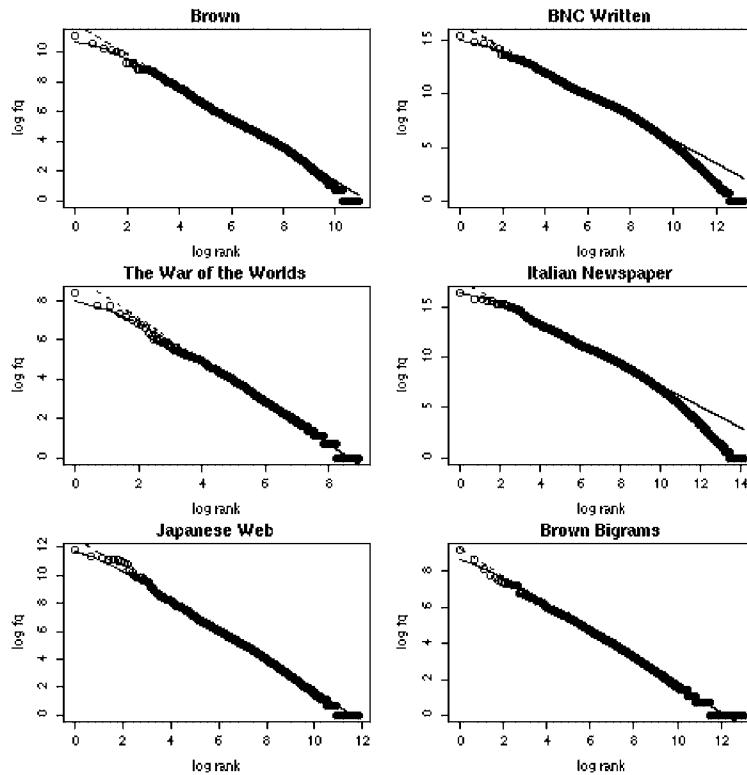


Fig. 37.7: Log rank/log frequency plots with Zipf and Zipf-Mandelbrot fits for the Brown (top left), written BNC (top right), *The War of the Worlds* (middle left), *la Repubblica* (middle right), the Japanese web-page corpus (bottom left), the Brown bigrams (bottom right)

Zipf's law is an inverse power function, i.e., frequency is proportional to a negative power ($-a$) of rank. That frequency decreases when rank increases is obvious, given that ranks are based on frequency. However, compared to other distributions commonly used to model decay in natural and artificial phenomena, such as the exponential distribution, a power law distribution decreases more slowly, leaving a long tail of low frequency items. Zipfian distributions are not limited to word frequencies, but are also encountered in completely unrelated phenomena such as city populations, incomes (in economics, a variant of Zipf's law is known as *Pareto's law*), frequency of citations of scientific papers and visits to web-sites. It should be clear that these are all distributions of the few-giants/many-dwarves type. For a short survey of attested Zipfian distributions, see Li (2002).

Mathematically, Zipf's law has the convenient property that, if we take the logarithm of both sides, we obtain a linear function (recall that the logarithm of a fraction equals the difference of the logarithms of its numerator and denominator, and that $\log x^k$ equals $k \log x$):

$$\log f(w) = \log C - a \log r(w) \quad (2)$$

This is the equation of a straight line with intercept $\log C$ and slope $-a$. Thus, Zipf's law predicts that rank/frequency profiles will appear as straight lines in double logarithmic space (i.e., plotting log frequency as a function of log rank). The values of the intercept and the slope (and thus of Zipf's law's parameters C and a) can be easily estimated using the standard method of least squares, implemented in most statistical packages (see, e.g., Dalgaard 2002 for the R implementation). Figure 37.7 presents some of the rank/frequency profiles we already saw plotted in double logarithmic space. As it can be seen, fitting a straight line to the log-log curves is not unreasonable (indeed, Zipf probably came up with his formula by looking at plots of this sort), although the fit is far from perfect, especially at the edges.

At the right edge of the curves, among the highest ranks (lowest frequencies), we notice a “bell-bottom” pattern due to the increasingly wider horizontal lines corresponding to the rare words that are assigned different ranks but have the same frequency. This is expected, since we are fitting a model predicting no ties (but many words with very near continuous frequencies) to an empirical curve that for high ranks is essentially a discrete step function. More worryingly, for the two largest corpora (BNC and *la Repubblica*) we observe a curvature suggesting that frequency is dropping more rapidly than what is predicted by Zipf's law. This tendency is already noticeable, to a lesser extent, in the Brown and Japanese web corpus curves. Zipf and other early scholars had no access to large corpora where the phenomenon is clear (we do not observe this curvature in *The War of the Worlds*). The BNC and *la Repubblica* plots suggest that we should perhaps be fitting two straight lines to the data: one for the top ranks and one, with a steeper slope, for the bottom ranks. Indeed, Ha et al. (2006) obtain a good fit to a large English corpus with two lines, one for the top 5000 ranks and another (with a slope twice as steep) for the remaining ranks, and present similar results for other languages.

At the other end of the plot (low ranks, high frequencies) we observe, again, a downward curvature of the empirical profile, i.e., the attested high frequencies tend to be lower than what would be predicted by their rank according to Zipf's law. The pattern was observed early on, and Mandelbrot (1953) added a parameter to Zipf's law to take care of this downward curvature:

$$f(w) = \frac{C}{(r(w) + b)^a} \quad (3)$$

Zipf's original law is a special case of Zipf-Mandelbrot's law with $b = 0$. A reasonably small value of b will lower the frequency of the first few ranks in a significant manner but it will hardly affect higher ranks. For example, if we assume like above that $C = 60,000$ and $a = 1$, and furthermore that $b = 1$, then for the most frequent word Zipf's law (equation 1) predicts a frequency of $60,000/1 = 60,000$ whereas the Zipf-Mandelbrot's formula (equation 3) predicts half this frequency: $60,000/(1+1) = 30,000$. On the other hand, for the word with rank 1000 the difference in predicted frequency between the two formulas is minimal ($60,000/1,000 = 60$ with Zipf's formula, and $60,000/1,001 = 59.94$ with Mandelbrot's variant). Mandelbrot's formula no longer corresponds to a straight line in double logarithmic space:

$$\log f(w) = \log C - a \log(r(w) - b) \quad (4)$$

This makes sense empirically since we just saw that the log rank/log frequency profiles are not quite straight lines, but it complicates the math since we can no longer use a simple least squares linear fit model as with Zipf's original equation. In my experience, reasonable fits can be obtained by first setting b to 0 and calculating $\log C$ and a with the least squares method, and then increasing b in small steps until the goodness of fit of equation (4) applied to the first few ranks (those that will be considerably below the predicted straight line) stops improving. Figure 37.7 presents Zipf and Zipf-Mandelbrot fits to the empirical frequency rank profiles (as dashed and continuous lines, respectively). The Zipf parameters were found with the least squares method applied to the first 10,000 ranks. The extra Zipf-Mandelbrot parameter b was calculated with the method I just described, applied to the top 20 ranks (top 2 ranks in the Japanese corpus). As expected, in all plots the difference between the Zipf and Zipf-Mandelbrot curves is noticeable only for the lowest ranks (top left).

The a parameter is close to 1 for all the word frequency curves, ranging from 1.04 (*la Repubblica*) to 1.09 (BNC and Japanese web corpus). The tendency of a to be close to 1 is well known, and it justifies the simplified version of Zipf's law sometimes found in the literature, in which the formula is reduced to $f = C/r$, by assuming $a = 1$.

In the Brown bigram rank/frequency profile, the estimated a value is 0.76, well below the values typical of single word curves. Also, the plot suggests that for bigrams there is no need for the extra parameter b . The bigram frequencies look most decidedly like a straight line, without clear signs of downward curvatures at the top or bottom. Zipf's law may provide a better fit to ngram distributions than to single words (Ha et al. 2002).

3.1. Zipf's law for frequency spectra

Zipf (1965, 40 ff.) also analyzed the frequency spectrum in terms of a power law of the form:

$$V(f) = \frac{C}{f^a} \quad (5)$$

Again, the parameters can be estimated with a simple linear least squares fit in double logarithmic space. Figure 37.8 shows that Zipf's power law for frequency spectra provides reasonable fits to the Brown and BNC corpora (parameters estimated with the least squares method using the top 50 frequency classes).

Observing how Zipf(-Mandelbrot)'s law for rank/frequency profiles fits high frequency words better and how the frequency spectrum law fits low frequency words better, Naranan/Balasubrahmanyam (1998) propose to use (a variation of) the former to model function words and (a variation of) the latter to model content words.

3.2. Explanations of Zipf's law

Language after language, we find that Zipf(-Mandelbrot)'s law fits the data reasonably well. This has prompted many scholars to seek an explanation for this pattern. Zipf famously proposed to interpret it in terms of a “least effort” principle: the tension be-

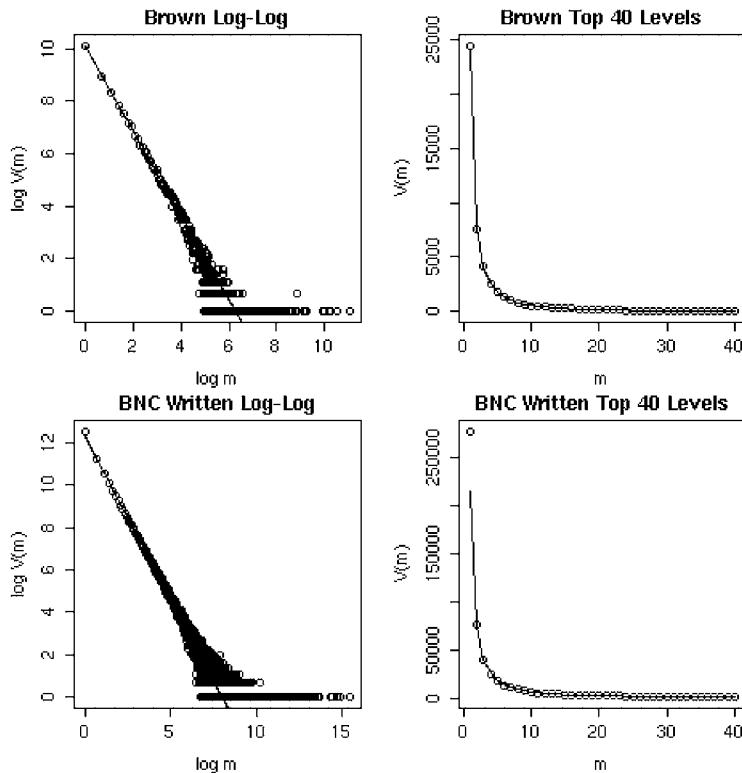


Fig. 37.8: Frequency spectra in log-log space (left panels) and first 40 classes of the frequency spectrum (right panels) in the Brown (top row) and written BNC (bottom row), with Zipfian fits based on equation (5)

tween the goal of the speaker to minimize production efforts by using only few words very frequently and the goal of the listener to minimize perceptual confusion by having a large vocabulary of distinct words would lead to the compromise distribution predicted by Zipf's law, with few high frequency types and many low frequency types. Mandelbrot derived his version of the law from information-theoretic notions, as the optimal solution to the problem of minimizing the average cost per unit of information in a text. Taking a different approach, other scholars (most famously, Simon 1955), observing how widespread Zipf's law is across phenomena that are clearly not related (such as word frequency distributions, city sizes and income distributions), studied under which general conditions such a distribution might arise.

Interestingly, texts constructed by generating characters (including white space) in random order also exhibit a Zipfian distribution (Miller 1957; Li 1992). Intuitively, when combining characters randomly, short words will be few but much more likely to occur by chance, whereas long words will be many but each of them will be extremely unlikely. Thus, in the output of the random generation process we will observe the by-now-familiar few-giants/many-dwarves pattern. Some authors (e. g., Miller 1957) take the fact that random text has a Zipfian distribution as evidence against "deep" explanations of Zipf's

law in terms of principles of language or communication. However, unlike in random text generation, the frequency with which a speaker selects a word will not depend on the length of the characters that compose it (the effect, as already observed by Zipf, is likely to go in the other direction, with a tendency for more frequently used words to be shortened). Thus, the random text experiments are not “explaining” Zipf’s law in natural language in any psychologically plausible sense.

4. Practical consequences

Although most of the literature on word frequency distributions is highly theoretical, the basic patterns of frequency in corpora have important consequences in practical work. First and most importantly, the Zipfian nature of word frequency distributions causes data sparseness problems. No matter how large a corpus is, most of the words occurring in it have very low frequency and a small set of frequent words constitutes the large majority of the tokens in the corpus. The distribution of bigrams and linguistic units larger than the word is even more skewed. Anybody working with corpora should be aware of these facts.

For example, according to the guidelines in Sinclair (2005), a trained lexicographer will need to inspect at least 20 instances of an unambiguous word to get an idea of its behavior. Even in a large corpus such as the (written) BNC, a lexicographer will find that less than 14% of the words have a frequency of 20 or higher. In a completely different area, Möbius (2003) observes that speech synthesis researchers often accept poor modeling of rare words (and other relevant units) in virtue of the fact that they are rare. However, Möbius observes that, because of the Zipfian nature of linguistic data, although each rare unit has a very low probability to occur, the overall probability that at least one rare unit will occur in a sentence approaches certainty.

Another facet of the data sparseness problem is that even large corpora do not sample the whole vocabulary of the language they represent: as sample increases, the number of types (vocabulary size) keeps increasing. This is illustrated for the Brown corpus in Figure 37.9, where I plotted the overall number of types (V) and the number of words occurring once ($V(1)$) found in the first 100K, 200K, etc. tokens, up to the full corpus size. Even at full corpus size the vocabulary is still growing.

Baayen (2001, 49–50) shows that the growth rate of the vocabulary, the rate at which the vocabulary size increases as sample size increases, can be estimated as follows:

$$G = \frac{V(1)}{N} \tag{6}$$

In equation (6), $V(1)$ is the number of words occurring once (*hapax legomena*, Ancient Greek for “said once”) in a sample of size N . The formula should make intuitive sense: the proportion of hapax legomena that we encountered up to the N th token is a reasonable estimate of how likely it is that word $N + 1$ will be a hapax legomenon, i. e., a word that we have not seen before and that will consequently increase vocabulary size. In the Brown corpus, $G = 24375 / 996883 = .024$, indicating that the vocabulary size is still growing at a relatively fast pace. The vocabulary is still growing (although at a slower pace) in much larger corpora, such as the written section of the BNC ($G = .003$) and the *la Repubblica* corpus ($G = .002$).

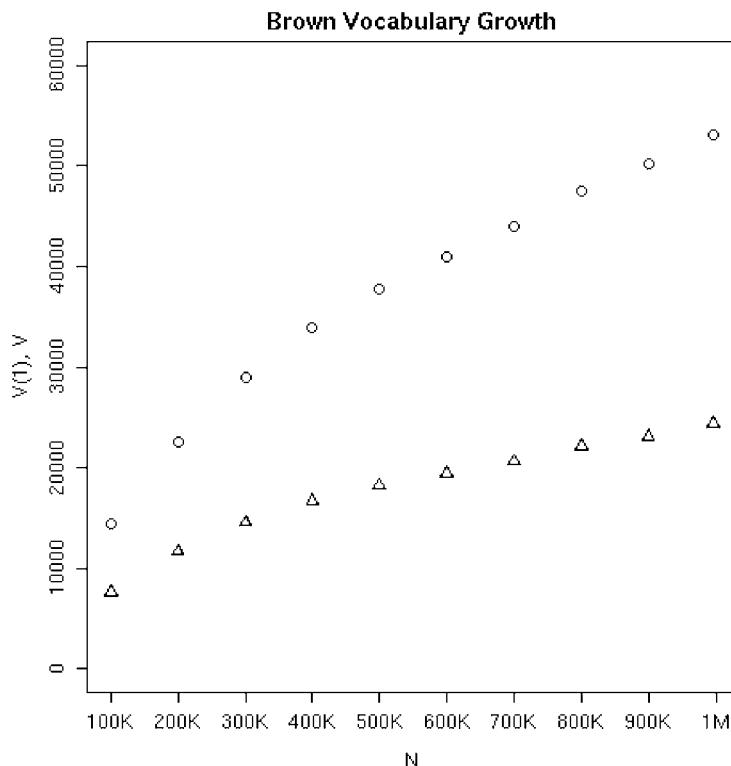


Fig. 37.9: The Brown corpus vocabulary growth curve: number of types (circles) and hapax legomena (triangles) for 10 increasingly larger token samples (N)

An important consequence of the fact that even large corpora are not sampling the full vocabulary of types they are drawn from is that the standard method of estimating the probability of occurrence of a word (or ngram) by its relative frequency in a corpus is very inaccurate. On the one hand, the word types that are not in the corpus are wrongly assigned 0 probability. On the other hand, the probability of the words that do occur in the corpus is overestimated, since they take up probability mass that should have been assigned to unseen words. Indeed, much work in corpus-based computational linguistics (see, e. g., Manning/Schütze 1999) focuses on ways to solve problems deriving from data sparseness, e. g., by assigning some probability mass to unseen words/ngrams with heuristic methods, or by clustering words into classes to obtain more robust statistics, or by using massive data collections, such as the web.

Another consequence of the fact that V keeps growing with corpus size is that we cannot use it as a measure of lexical richness when comparing corpora of different sizes: larger corpora will tend, trivially, to have more types. The fact that V increases with N (in ways that are not captured by simple functional relations) also affects nearly all the “constants” that have been proposed in the literature as measures of lexical richness (Tweedie/Baayen 1998), which turn out to vary with corpus size, and thus are not true constants. Statistical models of word frequency distributions (such as those introduced in Baayen 2001) provide formulas for the expectation (mean) and variance of quantities

such as vocabulary size at arbitrary sample sizes. Thus, they allow the comparison of corpora of different sizes (we can compute, e. g., the expected number of types we would see in a smaller corpus X if we could “stretch” it to the length of a larger corpus Y), and, under certain assumptions, to assess whether the vocabulary size difference between the corpora is statistically significant. Word frequencies require specialized statistical models, since some crucial aspects of standard methods, such as the assumption that the central limit theorem guarantees the normality of sample averages for reasonably large samples, are not appropriate for the extremely skewed word/ngram frequency data. Lexical-statistical models have been applied most extensively in stylometry and in the study of morphological productivity (see articles 50 and 41 of this handbook, respectively), but also in terminology (Kageura 1998) and collocation mining (Evert 2004). Unfortunately, the experiments reported in Evert/Baroni (2006) indicate that the prediction quality of current lexical-statistical models is not very high, probably because typical corpus data severely violate the randomness assumption that lies at the core of statistical modeling.

The Zipfian distribution of word frequencies is not only “bad news”. The fact that we can expect words in pretty much any natural language text to have this distribution (and the coefficient α to be close to 1) has found many applications, ranging from index compression (Baldi/Frasconi/Smyth 2003, section 4.1.2) to term weighting in information retrieval (Witten/Moffat/Bell 1999, section 4.4), to cryptography (Landini/Zandbergen 1998), to Bayesian modeling of morpheme frequencies (Creutz 2003).

5. Conclusion

This article introduced the typical patterns of frequency distribution encountered in corpora/texts. It also proposed Zipf-Mandelbrot’s law as a descriptive model that captures such patterns, and illustrated some of the consequences of these patterns for corpus-based work. Of course, I only scratched the surface of the large body of studies on lexical statistics.

The interested reader should proceed to Baayen (2001), a very thorough (and mathematically challenging) introduction to word frequency distributions with an emphasis on statistical modeling. I am not aware of contemporary introductions to lexical statistics at a less advanced level. Muller (1977) is an introduction in French to the basic concepts of word frequency analysis.

The *Journal of Quantitative Linguistics* and *Glottometrics* often feature articles on relevant topics. In 2002, the latter published three special issues in honor of George Kingsley Zipf. The recent HSK handbook on quantitative linguistics (Köhler/Altmann/Piotrowski 2005) features several articles on various aspects and applications of lexical statistics.

For those interested in the hands-on approach, to free *zipfR* package provides functions for the analysis of word frequency distributions integrated in the popular open source statistical package R. For information, visit the site: <http://purl.org/stefan.evert/zipfr>.

Finally, Wentian Li maintains a very up-to-date Internet bibliography on Zipf’s law and related principles at: <http://www.nslij-genetics.org/wli/zipf>.

6. Acknowledgments

I would like to thank: Stefan Evert, who patiently taught me everything I know about frequency distributions; Anke Lüdeling, for many stimulating conversations about lexical statistics and for detailed feedback on an early draft of this article; Silvia Bernardini, for useful advice about which aspects of lexical statistics might be of interest to corpus linguists; and Harald Baayen, for his very fast and helpful replies to my sudden emails about vocabulary growth and other matters. The usual disclaimers apply.

7. Literature

- Abney, Steven (1996), Statistical Methods and Linguistics. In: Klavans, Judith L./Resnik, Philip (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge MA: MIT Press, 1–26.
- Baayen, R. Harald (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.
- Baldi, Pierre/Frasconi, Paolo/Smyth, Padhraic (2003), *Modeling the Internet and the Web*. Chichester: Wiley.
- Baroni, Marco/Bernardini, Silvia/Comastri, Federica/Piccioni, Lorenzo/Volpi, Alessandra/Aston, Guy/Mazzoleni, Marco (2004), Introducing the la Repubblica corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. In: *Proceedings of LREC 2004*. Lisbon: ELDA, 1771–1774.
- Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Creutz, Mathias (2003), Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. In: *Proceedings of ACL 03*. Sapporo, Japan, 280–287.
- Dalgaard, Peter (2002), *Introductory Statistics with R*. New York: Springer.
- Evert, Stefan (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Doctoral thesis, University of Stuttgart/IMS.
- Evert, Stefan/Baroni, Marco (2006), Testing the Extrapolation Quality of Word Frequency Models. In: *Proceedings of Corpus Linguistics 2005*, available from <http://www.corpus.bham.ac.uk/PCCLC>.
- Ha, Le Quan/Sicilia-Garcia, Elvira/Ming, Ji/Smith, Francis (2002), Extension of Zipf's Law to Words and Phrases. In: *Proceedings of COLING 2002*. Taipei, Taiwan, 315–320.
- Ha, Le Quan/Stewart, Darryl/Hanna, Philip/Smith, Francis (2006), Zipf and Type-Token Rules for the English, Spanish, Irish and Latin Languages. In: *Web Journal of Formal, Computational & Cognitive Linguistics* 8, available from <http://fccl.ksu.ru/>.
- Kageura, Kyo (1998), A Statistical Analysis of Morphemes in Japanese Terminology. In: *Proceedings of COLING-ACL 98*. Montreal, Quebec, Canada, 638–645.
- Köhler, Reinhard/Altmann, Gabriel/Piotrowski, Rajmund (eds.) (2005), *Quantitative Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.
- Landini, Gabriel/Zandbergen, René (1998), A Well-kept Secret of Mediaeval Science: The Voynich Manuscript. In: *Aesculapius* 1998, available from <http://www.voynich.nu/extra/aes.html>.
- Li, Wentian (1992), Random Texts Exhibit Zipf's-Law-like Word Frequency Distribution. In: *IEEE Transactions on Information Theory* 38, 1842–1845.
- Li, Wentian (2002), Zipf's Law Everywhere. In: *Glottometrics* 5, 14–21.
- Mandelbrot, Benoit (1953), An Informational Theory of the Statistical Structure of Languages. In: Jackson, W. (ed.), *Communication Theory*. London: Butterworth, 486–502.
- Manning, Christopher D./Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

- Matsumoto, Yuji/Kitauchi, Akira/Yamashita, Tatsuo/Hirano, Yoshitaka/Matsuda, Hiroshi/Takao-ka, Kazuma/Asahara, Masayuki (2000), *Morphological Analysis System ChaSen Version 2.2.1 Manual*. NIST Technical Report.
- McEnery, Tony/Wilson, Andrew (2001), *Corpus Linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.
- Miller, George (1957), Some Effects of Intermittent Silence. In: *American Journal of Psychology* 52, 311–314.
- Möbius, Bernd (2003), Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. In: *International Journal of Speech Technology* 6, 57–71.
- Muller, Charles (1977), *Principes et méthodes de statistique lexicale*. Paris: Hachette.
- Naranan, Sundaresan/Balasubrahmanyam, Vriddhachalam K. (1998), Models for Power Law Relations in Linguistics and Information Science. In: *Journal of Quantitative Linguistics* 5, 35–61.
- Oakes, Michael (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Sampson, Geoffrey (2002), Review of Harald Baayen: Word Frequency Distributions. In: *Computational Linguistics* 28, 565–569.
- Simon, Herbert A. (1955), On a Class of Skew Distribution Functions. In: *Biometrika* 42, 425–440.
- Sinclair, John (2005), Corpus and Text: Basic Principles. In: Wynne, Martin (ed.), *Guide to Good Practice in Developing Linguistic Corpora*. Oxford: Oxford Books, 1–6. Available from <http://ahds.ac.uk/litlangling/linguistics/index.html>.
- Ueyama, Motoko/Baroni, Marco (2006), Automated Construction and Evaluation of a Japanese Web-based Reference Corpus. In: *Proceedings of Corpus Linguistics 2005*, available from <http://www.corpus.bham.ac.uk/PCLC>.
- Tweedie, Fiona/Baayen, R. Harald (1998), How Variable May a Constant Be? Measures of Lexical Richness in Perspective. In: *Computers and the Humanities* 32, 323–352.
- Witten, Ian/Moffat, Alistair/Bell, Timothy (1999), *Managing Gigabytes*. 2nd edition. San Francisco: Morgan Kaufmann.
- Zipf, George Kingsley (1949), *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.
- Zipf, George Kingsley (1965), *The Psycho-biology of Language*. Cambridge MA: MIT Press.

Marco Baroni, Trento (Italy)

38. Multi-dimensional approaches

1. Studying register and register variation
2. Theoretical background
3. Conceptual introduction to the multi-dimensional approach
4. Methodology in the multi-dimensional approach
5. Summary of the 1988 MD analysis of English registers
6. Types of MD study
7. Application of the 1988 dimensions to other discourse domains
8. Other MD analyses of English registers
9. MD analyses of other languages
10. Conclusion
11. Literature

- Matsumoto, Yuji/Kitauchi, Akira/Yamashita, Tatsuo/Hirano, Yoshitaka/Matsuda, Hiroshi/Takao-ka, Kazuma/Asahara, Masayuki (2000), *Morphological Analysis System ChaSen Version 2.2.1 Manual*. NIST Technical Report.
- McEnery, Tony/Wilson, Andrew (2001), *Corpus Linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.
- Miller, George (1957), Some Effects of Intermittent Silence. In: *American Journal of Psychology* 52, 311–314.
- Möbius, Bernd (2003), Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. In: *International Journal of Speech Technology* 6, 57–71.
- Muller, Charles (1977), *Principes et méthodes de statistique lexicale*. Paris: Hachette.
- Naranan, Sundaresan/Balasubrahmanyam, Vriddhachalam K. (1998), Models for Power Law Relations in Linguistics and Information Science. In: *Journal of Quantitative Linguistics* 5, 35–61.
- Oakes, Michael (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Sampson, Geoffrey (2002), Review of Harald Baayen: Word Frequency Distributions. In: *Computational Linguistics* 28, 565–569.
- Simon, Herbert A. (1955), On a Class of Skew Distribution Functions. In: *Biometrika* 42, 425–440.
- Sinclair, John (2005), Corpus and Text: Basic Principles. In: Wynne, Martin (ed.), *Guide to Good Practice in Developing Linguistic Corpora*. Oxford: Oxford Books, 1–6. Available from <http://ahds.ac.uk/litlangling/linguistics/index.html>.
- Ueyama, Motoko/Baroni, Marco (2006), Automated Construction and Evaluation of a Japanese Web-based Reference Corpus. In: *Proceedings of Corpus Linguistics 2005*, available from <http://www.corpus.bham.ac.uk/PCLC>.
- Tweedie, Fiona/Baayen, R. Harald (1998), How Variable May a Constant Be? Measures of Lexical Richness in Perspective. In: *Computers and the Humanities* 32, 323–352.
- Witten, Ian/Moffat, Alistair/Bell, Timothy (1999), *Managing Gigabytes*. 2nd edition. San Francisco: Morgan Kaufmann.
- Zipf, George Kingsley (1949), *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.
- Zipf, George Kingsley (1965), *The Psycho-biology of Language*. Cambridge MA: MIT Press.

Marco Baroni, Trento (Italy)

38. Multi-dimensional approaches

1. Studying register and register variation
2. Theoretical background
3. Conceptual introduction to the multi-dimensional approach
4. Methodology in the multi-dimensional approach
5. Summary of the 1988 MD analysis of English registers
6. Types of MD study
7. Application of the 1988 dimensions to other discourse domains
8. Other MD analyses of English registers
9. MD analyses of other languages
10. Conclusion
11. Literature

1. Studying register and register variation

For many years, researchers have studied the language used in different situations: the description of *registers*. *Register* is used here as a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances.

In many cases, registers are named varieties within a culture, such as novels, letters, memos, editorials, sermons, and lectures. However, registers can be defined at any level of generality, and more specialized registers may not have widely used names. For example, 'academic prose' is a very general register, while 'methodology sections in experimental psychology articles' is a much more highly specified register.

Although registers are defined in situational terms, they can also be compared with respect to their linguistic characteristics: the study of *register variation*. Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar reflecting a range of situational factors. The ubiquitous nature of register variation has been noted by a number of scholars, for example:

"register variation, in which language structure varies in accordance with the occasions of use, is all-pervasive in human language" (Ferguson 1983, 154)

"no human being talks the same way all the time ... At the very least, a variety of registers and styles is used and encountered." (Hymes 1984, 44)

However, despite the fundamental importance of register variation, there have been few comprehensive analyses of the register differences in a language. This gap is due mostly to methodological difficulties: until recently, it has been unfeasible to analyze the full range of texts, registers, and linguistic characteristics required for comprehensive analyses of register variation. With the availability of large on-line text corpora and computational analytical tools, such analyses have become possible. Multi-dimensional (MD) analysis – the focus of the present article – is a corpus-based research approach developed for the comprehensive analysis of register variation.

2. Theoretical background

In a few cases, registers can be distinguished by the presence of distinctive *register markers*: linguistic features restricted to a single register. For example, Ferguson (1983) describes how the grammatical routine known as 'the count', as in *the count is two and one*, is a distinctive register marker of baseball game broadcasts. In most cases, though, register differences are realized through the relative presence or absence of *register features* – core lexical and grammatical features – rather than by the presence of a few distinctive register markers. Register features are found to some extent in almost all texts and registers, but there are often large differences in their relative distributions across registers. In fact, many registers are distinguished only by a particularly frequent or infrequent occurrence of a set of register features.

Register analyses of these core linguistic features are necessarily quantitative, to determine the relative distribution of linguistic features. Further, such analyses require a com-

parative approach. That is, it is only by quantitative comparison to a range of other registers that we are able to determine whether a given frequency of occurrence is notably common or rare. A quantitative comparative approach allows us to treat register as a continuous construct: texts are situated within a continuous space of linguistic variation, enabling analysis of the ways in which registers are more or less different with respect to the full range of core linguistic features.

It turns out, though, that the relative distribution of common linguistic features, considered individually, cannot reliably distinguish among registers. There are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. However, when analyses are based on the *co-occurrence* and *alternation* patterns for groups of linguistic features, important differences across registers are revealed.

The theoretical importance of linguistic co-occurrence was recognized well before corpus-based methods were developed to analyze these patterns. For example, Ervin-Tripp (1972) identified ‘speech styles’ as varieties that are defined by a shared set of co-occurring linguistic features, while Brown/Fraser (1979, 38–39) observed that it can be “misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers”.

The Multi-Dimensional (MD) approach was developed as a corpus-based methodology to analyze the linguistic co-occurrence patterns associated with register variation. The following section provides a conceptual overview of the approach, while section 4 summarizes the methodological techniques used for MD analyses.

3. Conceptual introduction to the multi-dimensional approach

MD analysis was developed as a corpus-based methodological approach to: (1) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms; and (2) compare registers in the linguistic space defined by those co-occurrence patterns. The approach was first used in Biber (1985, 1986) and then developed more fully in Biber (1988).

The notion of linguistic co-occurrence has been given formal status in the MD approach, in that different co-occurrence patterns are analyzed as underlying *dimensions* of variation. The co-occurrence patterns comprising each dimension are identified quantitatively. That is, based on the actual distributions of linguistic features in a large corpus of texts, statistical techniques (specifically factor analysis) are used to identify the sets of linguistic features that frequently co-occur in texts. The methods used to identify these co-occurrence patterns are described in section 4.

Qualitative analysis is required to interpret the functions associated with each set of co-occurring linguistic features. The dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions.

A simple example is the way in which first and second person pronouns, direct questions, and imperatives are all related to interactivity. Contractions, false starts, and generalized content words (e.g., *thing*) are all related to the constraints imposed by real-time production. The functional bases of other co-occurrence patterns are less transparent, so that careful qualitative analyses of particular texts are required to help interpret the underlying functions.

In sum, the salient characteristics of the MD approach are:

- The research goal of the approach is to describe the general patterns of variation among registers, considering a comprehensive set of linguistic features and the range of registers in the target domain of use.
- The unit of analysis in the approach is each text, rather than individual linguistic constructions.
- The importance of variationist and comparative perspectives is assumed by the approach. That is, different kinds of text differ linguistically and functionally, so that analysis of a single text variety cannot adequately represent a discourse domain. From a quantitative point of view, it is not possible to determine which linguistic distributions are noteworthy without comparison to other text varieties.
- The approach is explicitly multi-dimensional. That is, it is assumed that multiple parameters of variation will operate in any discourse domain.
- The approach is empirical and quantitative. Analyses are based on normed frequency counts of linguistic features, describing the relative distributions of features across the texts in a corpus. The linguistic co-occurrence patterns that define each dimension are identified empirically using multivariate statistical techniques (see also article 40).
- The approach synthesizes quantitative and qualitative/functional methodological techniques. That is, the statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each distributional pattern. The approach is based on the assumption that statistical co-occurrence patterns reflect underlying shared communicative functions.

4. Methodology in the multi-dimensional approach

A complete multi-dimensional analysis follows eight methodological steps:

1. An appropriate corpus is designed based on previous research and analysis (cf. article 9). Texts are collected, transcribed (in the case of spoken texts), and input into the computer. (In many cases, pre-existing corpora can be used; cf. article 20.)
2. Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the linguistic features.
3. Computer programs are developed for automated grammatical analysis, to identify – or ‘tag’ – all relevant linguistic features in texts (cf. articles 23, 24).
4. The entire corpus of texts is tagged automatically by computer, and all texts are edited interactively to ensure that the linguistic features are accurately identified.
5. Additional computer programs are developed and run to compute normed frequency counts of each linguistic feature in each text of the corpus.
6. The co-occurrence patterns among linguistic features are identified through a factor analysis of the frequency counts (cf. article 40).

7. The ‘factors’ from the factor analysis are interpreted functionally as underlying dimensions of variation.
8. Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among registers.

In practice, there are two different types of MD study: those that carry out a full MD analysis and those that apply previously identified dimensions to new areas of research. Methodologically, the two types differ in whether or not they include steps 6 and 7. Full MD studies, such as the original MD studies (Biber 1985, 1986, 1988), identify and interpret underlying dimensions of register variation and then use those dimensions to characterize registers; they thus include all eight methodological steps. However, many MD studies use the dimensions identified in Biber (1988) to describe and compare additional registers. These studies omit steps 6 and 7; since they use the previously-identified dimensions, such studies do not require a separate factor analysis.

4.1. Linguistic features and tagging the corpus

One of the first tasks in a MD study is to identify the linguistic features to be used in the analysis. The goal here is to be as inclusive as possible, identifying all linguistic features that might have functional associations, including lexical classes, grammatical categories, and syntactic constructions. Thus, any feature associated with particular communicative functions, or used to differing extents in different text varieties, is included. Occurrences of these features are counted in each text of the corpus, providing the basis for all subsequent statistical analyses.

The identification of functionally important linguistic features for the 1988 study of register variation was relatively easy, due to the large body of previous research studies on speech and writing. That study was based on 67 linguistic features, taken from 16 major grammatical and functional categories:

- 1) tense and aspect markers
- 2) place and time adverbials
- 3) pronouns and pro-verbs
- 4) questions
- 5) nominal forms
- 6) passives
- 7) stative forms
- 8) subordination features
- 9) prepositional phrases, adjectives, and adverbs
- 10) lexical specificity
- 11) lexical classes
- 12) modals
- 13) specialized verb classes
- 14) reduced forms and discontinuous structures
- 15) coordination
- 16) negation

These features are identified in the texts of a corpus by using an automatic ‘tagger’ (see Biber 1988, appendix II).

The current version of the tagger has both probabilistic and rule-based components, and uses multiple large-scale dictionaries. The tagger has been developed with three primary considerations: achieving high accuracy levels; robustness across texts from different registers (with different processing options for ‘oral’ and ‘literate’ texts); and identification of a large set of linguistic characteristics (e.g., distinguishing simple past tense, perfect aspect, passive voice, and postnominal modifier for past participle verbs; identifying the gap position for *WH*-relative clauses; identifying several different kinds of complement clause, and the existence of *that*-complementizer deletion). In recent years, several linguistic distinctions have been added to the tagger as part of the analyses carried out for the *Longman Grammar of Spoken and Written English* (Biber et al. 1999); these include many lexico-grammatical features, such as mental verbs controlling *that*-clauses, or verbs of effort controlling *to*-clauses. To ensure accurate tagging, problematic linguistic features are corrected interactively using a grammar checker (see also Biber/Conrad/Reppen 1998, methodology boxes 4 and 5).

Once texts have been tagged and interactively tag-edited, other programs are used to calculate the ‘normed’ (or ‘normalized’) rate of occurrence of linguistic features in each text (e.g., the number of nouns per 1,000 words; see Biber/Conrad/Reppen 1998, methodology box 6). (Some linguistic features have non-linear distributions and so must be adjusted in other ways; see article 37.) These normed counts provide the basis for the factor analysis, described in the following section.

4.2. Factor analysis to identify the ‘dimensions’ of variation

As described above, co-occurrence patterns are central to MD analyses: each dimension represents a different set of co-occurring linguistic features. These co-occurrence patterns are identified quantitatively, using a statistical technique known as factor analysis; each set of co-occurring features is referred to as a *factor*. In a factor analysis, a large number of original variables – in this case the linguistic features – are reduced to a small set of derived, underlying variables: the factors. In the 1988 MD analysis, the 67 linguistic features were reduced to 7 factors.

Each linguistic feature has some relation to each factor, and the strength of that relation is represented by *factor loadings*. (The factor loading is essentially a correlation, representing the amount of variance that a feature has in common with the total pool of shared variance accounted for by a factor.)

The factor loadings for the 1988 MD analysis of spoken and written registers are given in Table 38.1. Factor loadings can range from 0.0, which shows the absence of any relationship, to 1.0 (positive or negative), which shows a perfect correlation. The factor loading indicates the extent to which a linguistic feature is representative of the dimension underlying a factor; the size of the loading reflects the strength of the co-occurrence relationship between the feature in question and the total grouping of co-occurring features represented by the factor.

As Table 38.1 shows, each linguistic feature has a loading on each factor. However, when interpreting a factor, only features with salient or important loadings are consid-

Tab. 38.1: Factor loadings in the factor analysis of register variation in English

LING FEATURE	FACT1	FACT2	FACT3	FACT4	FACT5	FACT6	FACT7
Past tense	-0.083	0.895	0.002	-0.249	-0.049	-0.052	0.021
Perfects	0.051	0.480	0.049	-0.016	-0.101	0.146	0.143
Present tense	0.864	-0.467	-0.008	0.229	-0.006	0.011	0.011
Place adverbs	-0.417	-0.060	-0.492	-0.094	-0.067	-0.018	-0.023
Time adverbs	-0.199	-0.062	-0.604	-0.020	-0.290	0.116	-0.046
1st pers. pro.	0.744	0.088	0.025	0.026	-0.089	0.008	-0.098
2nd pers. pro.	0.860	-0.043	-0.018	0.016	0.007	-0.168	-0.064
3rd pers. pro.	-0.053	0.727	-0.074	-0.018	-0.167	-0.076	0.138
Pronoun <i>it</i>	0.706	-0.021	-0.038	-0.034	-0.038	0.022	0.060
Dem. pronouns	0.756	-0.166	-0.001	-0.108	0.004	0.306	-0.077
Proform <i>any</i>	0.618	0.046	0.011	0.085	-0.094	-0.085	-0.032
Proform <i>do</i>	0.821	0.004	0.071	0.049	-0.057	-0.077	-0.056
<i>Wh</i> questions	0.523	-0.024	0.117	-0.111	-0.032	0.036	-0.094
Nominaliz.	-0.272	-0.237	0.357	0.179	0.277	0.129	-0.019
<i>-ing</i> nouns	-0.252	-0.127	0.216	0.177	0.087	-0.052	0.052
Other nouns	-0.799	-0.280	-0.091	-0.045	-0.294	-0.076	-0.213
Agentless pasv.	-0.388	-0.145	0.109	0.060	0.430	0.063	-0.057
By pasv.	-0.256	-0.189	0.065	-0.124	0.413	-0.089	-0.045
Stative <i>be</i>	0.713	0.056	0.075	0.008	0.014	0.292	0.180
Existential <i>there</i>	0.262	0.108	0.113	-0.124	-0.004	0.318	0.017
<i>That</i> verb clause	0.045	0.228	0.125	0.265	0.053	0.558	-0.122
<i>That</i> adj clause	-0.124	0.066	-0.080	0.123	0.171	0.360	0.183
<i>Wh</i> clause	0.467	0.143	0.221	0.032	-0.050	-0.044	-0.027
Infinitive	-0.071	0.059	0.085	0.760	-0.274	-0.005	-0.074
Advl clause <i>-ing</i>	-0.211	0.392	-0.142	-0.076	0.268	-0.217	0.121
Advl clause <i>-ed</i>	-0.025	-0.154	0.029	-0.050	0.415	-0.142	-0.059
Whiz <i>-ed</i>	-0.382	-0.336	-0.071	-0.137	0.395	-0.128	-0.103
Whiz <i>-ing</i>	-0.325	-0.114	0.080	-0.169	0.212	-0.070	-0.093
<i>That</i> rel. subj.	0.051	-0.036	0.021	0.019	-0.058	0.184	0.033
<i>That</i> rel. obj.	-0.047	0.053	0.201	0.223	-0.125	0.457	-0.065
<i>Wh-</i> rel. subj.	-0.087	-0.067	0.453	-0.027	-0.174	0.228	0.047
<i>Wh-</i> rel. obj.	-0.072	0.049	0.627	-0.060	-0.083	0.302	0.165
<i>Wh-</i> rel. pied pip.	-0.029	0.026	0.606	-0.144	0.046	0.280	0.192
Sentence rel.	0.550	-0.086	0.152	-0.118	-0.025	0.048	-0.041

Tab. 38.1: (continued)

LING FEATURE	FACT1	FACT2	FACT3	FACT4	FACT5	FACT6	FACT7
Advl. cl. – reason	0.661	-0.080	0.110	0.023	-0.061	0.078	-0.076
Advl. cl. – conc.	0.006	0.092	0.100	-0.071	0.010	-0.056	0.300
Advl. cl. – cond.	0.319	-0.076	-0.206	0.466	0.120	0.103	-0.007
Advl. cl. – other	-0.109	0.051	-0.018	0.008	0.388	0.102	0.109
Prepositions	-0.540	-0.251	0.185	-0.185	0.234	0.145	-0.008
Attributive adj.	-0.474	-0.412	0.176	-0.055	-0.038	-0.064	0.299
Predicative adj.	0.187	0.076	-0.089	0.248	0.311	-0.012	0.210
Adverbs	0.416	-0.001	-0.458	-0.020	-0.156	0.053	0.314
Typetoken ratio	-0.537	0.058	0.002	-0.005	-0.311	-0.228	0.219
Word length	-0.575	-0.314	0.270	-0.009	0.023	0.028	0.081
Conjuncts	-0.141	-0.160	0.064	0.108	0.481	0.180	0.217
Downtoners	-0.084	-0.008	0.021	-0.080	0.066	0.113	0.325
Hedges	0.582	-0.156	-0.051	-0.087	-0.022	-0.145	0.096
Amplifiers	0.563	-0.156	-0.028	-0.124	-0.124	0.225	-0.018
Emphatics	0.739	-0.216	0.015	-0.027	-0.188	-0.087	0.210
Disc. particles	0.663	-0.218	-0.128	-0.029	-0.096	0.165	-0.140
Demonstratives	0.040	-0.062	0.113	0.010	0.132	0.478	0.153
Pos. modals	0.501	-0.123	0.044	0.367	0.122	-0.022	0.115
Nec. modals	-0.007	-0.107	-0.015	0.458	0.102	0.135	0.042
Pred. modals	0.047	-0.056	-0.054	0.535	-0.072	0.063	-0.184
Public verbs	0.098	0.431	0.163	0.135	-0.030	0.046	-0.279
Private verbs	0.962	0.160	0.179	-0.054	0.084	-0.049	0.106
Suasive verbs	-0.240	-0.035	-0.017	0.486	0.051	0.016	-0.237
Seem/appear	0.054	0.128	0.160	-0.010	0.015	0.045	0.348
Contractions	0.902	-0.100	-0.141	-0.138	-0.002	-0.057	-0.032
<i>That</i> deletions	0.909	0.036	0.098	-0.059	-0.005	-0.178	-0.081
Stranded preps	0.426	0.007	-0.124	-0.210	0.023	0.340	-0.100
Split infinitives	----- DROPPED -----						
Split auxiliaries	-0.195	0.040	0.012	0.437	0.043	0.120	0.239
Phrasal coord.	-0.253	-0.091	0.355	-0.066	-0.046	-0.324	0.126
Clausal coord.	0.476	0.041	-0.052	-0.161	-0.139	0.218	-0.125
Synthetic neg.	-0.232	0.402	0.046	0.133	-0.057	0.176	0.110
Analytic neg.	0.778	0.149	0.017	0.125	0.019	0.001	0.037

ered. In the 1988 analysis, features with loadings smaller than .35 were considered not important in the interpretation of a factor. Positive or negative sign does not influence the importance of a loading; for example, nouns, with a loading of $-.799$, have a larger weight on Factor 1 than first person pronouns, with a loading of $.744$.

Rather than reflecting importance, positive and negative sign identify two groupings of features that occur in a complementary pattern as part of the same factor. That is, when the features with positive loadings occur together frequently in a text, the features with negative loadings are markedly less frequent in that text, and vice versa. (For more technical information about the factor analysis, see Biber 1995, chapter 5.)

4.2.1. Interpretation of factors as dimensions of variation

Factor interpretations depend on the assumption that linguistic co-occurrence patterns reflect underlying communicative functions. That is, linguistic features occur together in texts because they serve related communicative functions. The interpretation of a factor is based on (1) analysis of the communicative function(s) most widely shared by the set of co-occurring features, and (2) analysis of the similarities and differences among registers with respect to the factor.

For example, Table 38.2 lists the features with salient loadings on Factor 1 in the 1988 MD analysis (i. e., the features with loadings greater than 0.35). In the interpretation of a factor, it is important to consider the likely reasons for the complementary distribution between positive and negative feature sets as well as the reasons for the co-occurrence patterns within those sets.

On Factor 1, the interpretation of the negative features is relatively straightforward. Nouns, word length, prepositional phrases, type/token ratio, and attributive adjectives all have negative loadings larger than $-.45$, and none of these features has a larger loading on another factor. These features reflect an informational focus, a careful integration of information in a text, and precise lexical choice. Text sample 1 illustrates these co-occurring linguistic characteristics in an academic article:

Text sample 1: Technical academic prose

Apart from these very general group related aspects, there are also individual aspects that need to be considered. Empirical data show that similar processes can be guided quite differently by users with different views on the purpose of the communication.

This text sample is typical of written expository prose in its dense integration of information: frequent nouns and long words, with most nouns being modified by attributive adjectives or prepositional phrases (e. g., *general group related aspects, individual aspects, empirical data, similar processes, users with different views on the purpose of the communication*).

The set of features with positive loadings on Factor 1 is more complex, although all of these features have been associated with interpersonal interaction, a focus on personal stance, and real-time production circumstances. For example, first and second person pronouns, *WH*-questions, emphatics, amplifiers, and sentence relatives can all be interpreted as reflecting interpersonal interaction and the involved expression of personal stance (feelings and attitudes). Other positive features are associated with the constraints

Tab. 38.2: Factor 1 features and loadings in the 1988 MD analysis of register variation

Dimension 1: Involved vs. Informational Production	
<i>Positive features:</i>	
private verbs	.96
<i>that</i> deletion	.91
contractions	.90
present tense verbs	.86
2nd person pronouns	.86
<i>do</i> as pro-verb	.82
analytic negation	.78
demonstrative pronouns	.76
general emphatics	.74
first person pronouns	.74
pronoun <i>it</i>	.71
<i>be</i> as main verb	.71
causative subordination	.66
discourse particles	.66
indefinite pronouns	.62
general hedges	.58
amplifiers	.56
sentence relatives	.55
<i>wh</i> questions	.52
possibility modals	.50
non-phrasal coordination	.48
<i>wh</i> clauses	.47
final prepositions	.43
(adverbs)	.42)
<i>Negative features:</i>	
nouns	-.80
word length	-.58
prepositions	-.54
type/token ratio	-.54
attributive adjs.	-.47
(place adverbials	-.42)
(agentless passives	-.39)
(past participle postnominal clauses	-.38)

of real time production, resulting in a reduced surface form, a generalized or uncertain presentation of information, and a generally ‘fragmented’ production of text; these include *that*-deletions, contractions, pro-verb DO, the pronominal forms, and final (stranded) prepositions. Text sample 2 illustrates the use of positive Dimension 1 features in a formal conversation (an interview) from the London-Lund Corpus:

Text sample 2: Interview

- B: come in . come in - - ah good morning
- A: good morning
- B: you're Mrs Finney
- A: yes I am
- B: how are you - my names Hart and this is Mr Mortlake
- C: how are you

A: how do you do .
 B: won't you sit down
 A: thank you - -
 B: mm well you are proposing . taking on . quite something Mrs Finney aren't you
 A: yes I am
 B: mm
 A: I should like to anyhow
 B: you know what you'd be going into
 A: yes I do

Overall, Factor 1 seems to represent a dimension marking interactional, stance-focused, and generalized content (the features with positive loadings on Table 38.2) versus high informational density and precise word choice (the features with negative loadings). Two separate communicative parameters seem to be represented here: the primary purpose of the writer/speaker (involved versus informational), and the production circumstances (those restricted by real-time constraints versus those enabling careful editing possibilities). Reflecting both of these parameters, the interpretive label 'Involved versus Informational Production' was proposed for the dimension underlying this factor.

4.2.2. Computing dimension scores

The second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. To achieve this, *dimension scores* are computed for each text, and then texts and registers are compared with respect to those scores. Dimension scores (or *factor scores*) are computed by summing the individual scores of the features with salient loadings on a dimension. In the 1988 MD study, only features with loadings greater than $.35$ on a factor were considered important enough to be used in the computation of dimension scores. For example, the Dimension 1 score for each text was computed by adding together the frequencies of private verbs, *that*-deletions, contractions, present tense verbs, etc. – the features with positive loadings on Factor 1 (from Table 38.2) – and then subtracting the frequencies of nouns, word length, prepositions, etc. – the features with negative loadings.

All individual linguistic variables are standardized to a mean of 0.0 and a standard deviation of 1.0 before the dimension scores are computed. This process converts feature scores to a single scale representing standard deviation units, so that all linguistic features have the same range of variation and therefore equivalent weights in the computation of dimension scores (see Biber 1988, 93–97).

Once a dimension score is computed for each text, the mean dimension score for each register can be computed. Plots of these mean dimension scores allow linguistic characterization of any given register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension. Standard statistical procedures (such as ANOVA) can be used to further analyze the statistical significance of differences among the mean dimension scores.

For example Figure 38.1 plots the mean dimension scores of registers along Dimension 1. The registers with large positive values (such as face-to-face and telephone conversations), have high frequencies of present tense verbs, private verbs, first and second

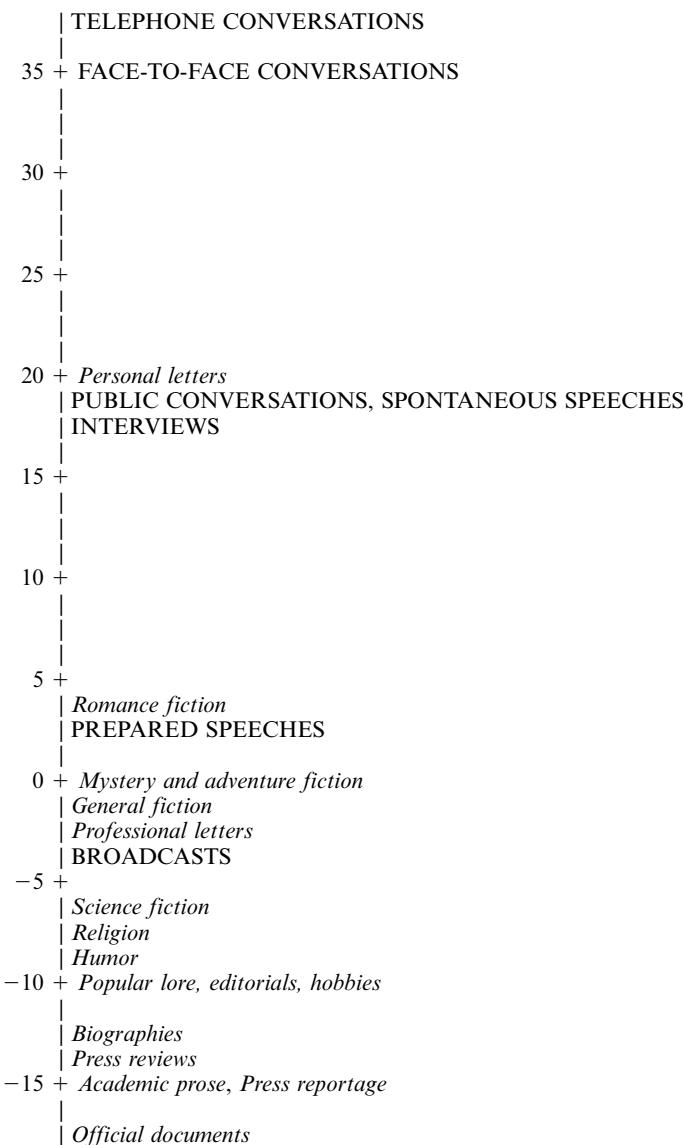


Fig. 38.1: Mean scores of registers along Dimension 1: Involved vs. Informational Production. Underlining denotes written registers; capitalization denotes spoken registers. (Adapted from Figure 7.1 in Biber 1988) ($F = 111.9$, $p < .0001$, $r^2 = 84.3\%$)

person pronouns, contractions, etc. – the features with salient positive weights on Dimension 1. At the same time, registers with large positive values have markedly low frequencies of nouns, prepositional phrases, long words, etc. – the features with salient negative weights on Dimension 1. Registers with large negative values (such as academic prose, press reportage and official documents) have the opposite linguistic characteris-

tics: very high frequencies of nouns, prepositional phrases, etc., plus low frequencies of private verbs, contractions, etc.

The relations among registers shown in Figure 38.1 confirm the interpretation of Dimension 1 as distinguishing among texts along a continuum of involved versus informational production. At the positive extreme, conversations are highly interactive and involved, with the language produced under real-time circumstances. Registers such as public conversations (interviews and panel discussions) are intermediate: they have a relatively informational purpose, but participants interact with one another and are still constrained by real time production. Finally, at the negative extreme, registers such as academic prose are non-interactive but highly informational in purpose, produced under controlled circumstances that permit extensive revision and editing.

The statistics given for F, p, and r^2 at the bottom of Figure 38.1 report the results of statistical tests. The F and p values give the results of an ANOVA, which tests whether there are statistically significant differences among the registers with respect to their mean Dimension 1 scores. The value for r^2 is a direct measure of strength or importance. The r^2 value measures the percentage of the variance among dimension scores that can be predicted by knowing the register categories (in this case, 84.3%).

The sections above have provided an overview of the methodological techniques used in MD analysis, illustrating the steps involved in the interpretation of a factor through consideration of Dimension 1 from the 1988 MD analysis. In section 5 below, I summarize the other major dimensions of variation from the 1988 study. Several publications provide fuller discussion of the relevant methodological issues, including: designing a representative corpus, the stability of linguistic feature counts, methodological details of factor analysis and computing factor scores, the stability of factor solutions, and the general goals, strengths, and weaknesses of MD analysis (see especially Biber 1988, chapters 4–5; 1995, chapter 5; 1990, 1993a, 1993b; Biber et al. 2003).

5. Summary of the 1988 MD analysis of English registers

The first major MD analysis (Biber 1988) was undertaken to investigate the relationship between spoken and written language in English. Most previous studies had been based on the assumption that speech and writing could be approached as a simple dichotomy; so, for example, a comparison of conversations and student essays was sometimes interpreted as representing general differences between speech and writing. In contrast, MD analysis is based on the assumption that all registers have distinctive linguistic characteristics (associated with their defining situational characteristics). Thus, the 1988 MD study of speech and writing set out to describe the relations among the full range of spoken registers and the full range of written registers – and to then compare speech and writing within the context of a comprehensive analysis of register variation.

For example, Figure 38.1 shows that there is a large range of variation among spoken registers with respect to the linguistic features that comprise Dimension 1 ('Involved versus Informational Production'). Conversation has extremely large positive Dimension 1 scores; spontaneous speeches and interviews have moderately large positive scores; while prepared speeches and broadcasts have scores around 0.0 (reflecting a balance of positive and negative linguistic features on this dimension). The written registers simi-

Tab. 38.3: Linguistic features on Dimensions 2–5 from the 1988 MD analysis

DIMENSION 2: Narrative vs. Non-narrative Discourse	
<i>Positive features:</i>	
past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43
synthetic negation	.40
present participial clauses	.39
<i>Negative features:</i>	
(present tense verbs	−.47)
(attributive adjs.	−.41)
DIMENSION 3: Situation-dependent vs. Elaborated Reference	
<i>Positive features:</i>	
time adverbials	.60
place adverbials	.49
adverbs	.46
<i>Negative features:</i>	
<i>Wh</i> relative clauses on object positions	−.63
pied piping constructions	−.61
<i>Wh</i> relative clauses on subject positions	−.45
phrasal coordination	−.36
nominalizations	−.36
DIMENSION 4: Overt Expression of Argumentation	
<i>Positive features:</i>	
infinitives	.76
prediction modals	.54
suasive verbs	.49
conditional subordination	.47
necessity modals	.46
split auxiliaries	.44
(possibility modals	.37)
[No negative features]	
DIMENSION 5: Abstract versus Non-abstract Style	
<i>Positive features:</i>	
conjunctions	−.48
agentless passives	−.43
past participial adverbial clauses	−.42
BY-passives	−.41
past participial postnominal clauses	−.40
other adverbial subordinators	−.39
[No negative features]	

larly show an extensive range of variation along Dimension 1. Expository informational registers, like official documents and academic prose, have very large negative scores; the fiction registers have scores around 0.0; while personal letters have a relatively large positive score.

This distribution shows that no single register can be taken as representative of the spoken or written mode. At the extremes, written informational prose is dramatically different from spoken conversation with respect to Dimension 1 scores. But written personal letters are relatively similar to spoken conversation, while spoken prepared speeches share some Dimension 1 characteristics with written fictional registers. Taken together, these Dimension 1 patterns indicate that there is extensive overlap between the spoken and written modes in these linguistic characteristics, while the extremes of each mode (i.e., conversation versus informational prose) are sharply distinguished from one another.

The overall comparison of speech and writing resulting from the 1988 MD analysis is actually even more complex, because six separate dimensions of variation were identified, and each of these defines a different set of relations among spoken and written registers. Table 38.3 displays the features and their loadings for Dimensions 2–5. (Dimension 6 has few salient linguistic features and is not considered here; see Biber 1988, 113–114, 154–160.) The name of each factor summarizes its interpretation. Dimension 1 – Involved vs. Informational Production – has been described above; the other dimensions are described and exemplified below.

5.1. Dimension 2: Narrative vs. Non-narrative Concerns

Dimension 2 is entitled Narrative vs. Non-narrative Concerns. The features with positive weights – past tense verbs, third-person pronouns, perfect aspect verbs, public verbs, synthetic negation and present participial clauses – are associated with past time narration. Past tense and perfect aspect verbs are used to describe past events, while the third-person pronouns refer to participants in the events. Public verbs (e.g., *say*, *tell*, *declare*) are used to express communication acts. Present participial clauses are typically used to add description and imagery to the narration. No features have strong negative loadings on this dimension (compared to their loadings on other dimensions); therefore, the dimension is a continuum reflecting the use of narrative features versus absence of those features.

Text sample 3 from romance fiction illustrates many of the features associated with Narrative Concerns. Particularly noticeable in this extract are the past tense verbs, third person pronouns (*he* and *his*), public verbs (particularly *said*), and the present participial clause which adds a descriptive detail to the action (*waving the manager away*).

Text sample 3: Romance fiction

But Mike Deegan was boiling mad now. When the inning was over he cursed the Anniston catcher all the way into the dugout ...

The Anniston manager came right up to the dugout in front of Mike. His face was flushed. ‘Deegan,’ the manager said, his voice pitched low, quivering.

‘That was a rotten thing to do.’

‘For God’s sake,’ Mike said, waving the manager away, ‘Stop it, will you? Tell your guys not to block the plate!’

The distribution of registers along Dimension 2, shown in Figure 38.2, further supports its interpretation as Narrative vs. Non-narrative Concerns. All types of fiction have

NARRATIVE

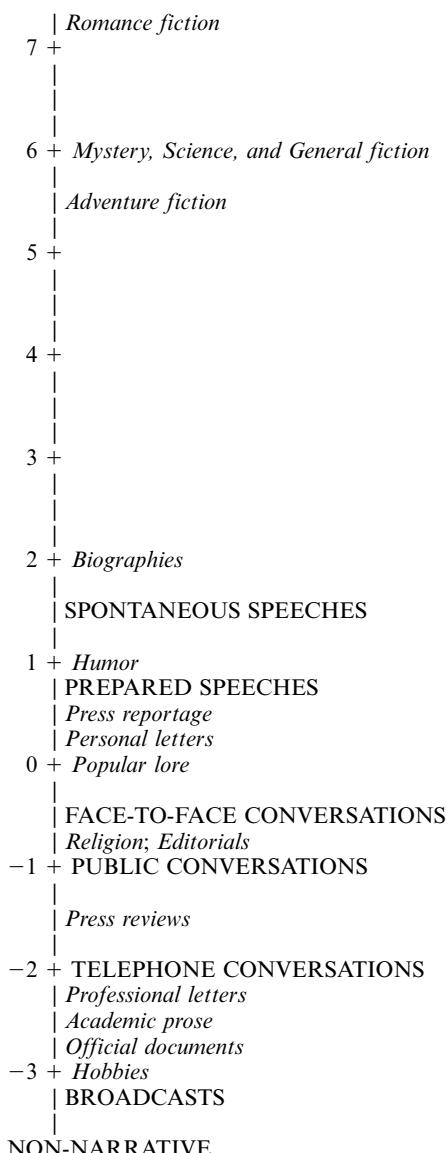


Fig. 38.2: Mean scores for registers along Dimension 2: Narrative versus Non-narrative Discourse.
 $(F = 32.3, p < .0001, r^2 = 60.8\%)$

markedly high positive scores, reflecting their emphasis on narrating events. In contrast, registers which are typically more concerned with events currently in progress (e.g., broadcasts) or with building arguments rather than narrating (e.g., academic prose) have negative scores on this dimension. Finally, some registers have scores around 0.0, reflect-

ing a mix of narrative and other features. For example, face-to-face conversation will often switch back and forth between narration of past events and discussion of current interactions.

5.2. Dimension 3: Elaborated vs. Situation-dependent Reference

Dimension 3 is labeled ‘Elaborated vs. Situation-dependent Reference’. The majority of positive features on this dimension are relative clause constructions – *WH*-relative clauses on object position, *WH*-relative clauses on subject position, and ‘pied piping’ constructions. These features explicitly identify referents or provide elaboration about referents.

In contrast, the negative features on this dimension are commonly used to refer to places and times outside of the text itself, in either the real world or an imaginary world created by the text. Place and time adverbials are used for temporal and locative reference (e.g., *earlier, soon; there, behind*). The other adverbs can have a wider range of functions, such as descriptions of manner.

Dimension 3 thus represents a continuum between texts that have elaborated, explicit reference, versus reference that is more dependent on the situational context. Figure 38.3 displays the distribution of registers along Dimension 3. Those with large positive scores – official documents, professional letters, academic prose, and press reviews – frequently use *WH*-relative clauses, along with phrasal coordinators and nominalizations (and a lack of time and place adverbials). Those with large negative scores – broadcasts and telephone conversations – rely more heavily on time and place adverbials and other adverbs in order to situate the discourse.

The two contrasting poles of Dimension 3 are exemplified by text samples 4 and 5. Text sample 4 is a short extract from an official document and illustrates the use of *WH*-relative clauses (*321 of whom were approved, 230 of whom were approved, who were awarded ...*) to elaborate noun referents. Phrasal coordination is also used.

Text sample 4: Official document

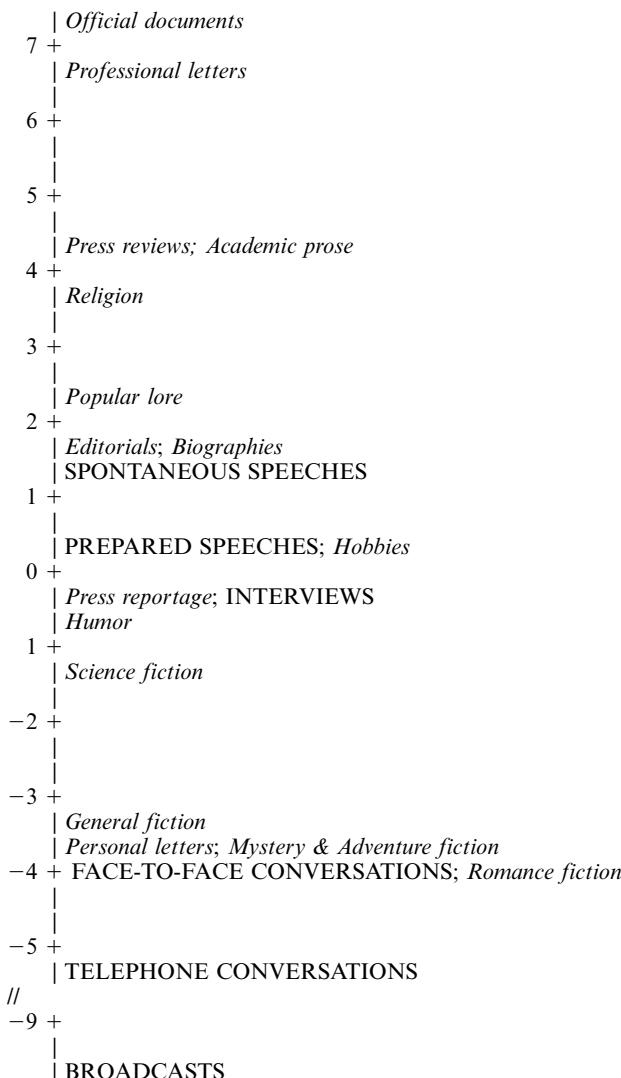
During the past year 347 candidates were examined by the Surgical Section, 321 of whom were approved, and 352 were examined by the Dental Section, 230 of whom were approved, making a total of 230 candidates who were awarded the Licence in Dental Surgery.

Text sample 5 comes from a radio broadcast of a soccer match. In contrast to the official document, time adverbials (*now*) and place adverbials (e.g., *just below us, here, forward*) are used to refer directly to the physical situation of the broadcast.

Text sample 5: Sports broadcast

and from the foot of Hemsley – the ball into touch – just below us here [...] a strike forward – but of course now turned – by manager O’Farrell [...] quickly taken by Brian Kydd – Kydd now to number seven

ELABORATED



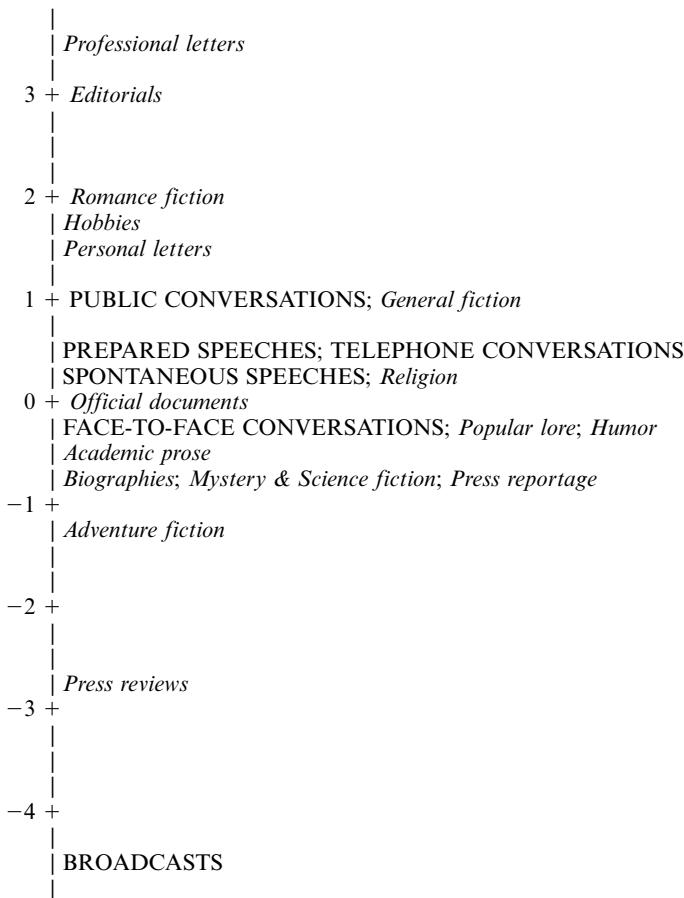
SITUATION-DEPENDENT

Fig. 38.3: Mean scores for registers along Dimension 3: Situation-Dependent versus Elaborated Reference. ($F = 31.9$, $p < .0001$, $r^2 = 60.5\%$)

5.3. Dimension 4: Overt Expression of Persuasion

Like Dimension 2, Dimension 4 has only features with positive weights: infinitives, prediction modals (e.g., *will*, *would*), suasive verbs (e.g., *agree*, *ask*, *insist*, *recommend*), conditional subordination, necessity modals (e.g., *ought*, *should*), split auxiliaries, and possibility modals (e.g., *might*, *may*).

OVERTLY ARGUMENTATIVE



NOT OVERTLY ARGUMENTATIVE

Fig. 38.4: Mean scores for registers along Dimension 4: Overt Expression of Argumentation. ($F = 4.2$, $p < .0001$, $r^2 = 16.9\%$)

This dimension has been interpreted as reflecting overt persuasion or argumentation, as exemplified in Text sample 6 from a professional letter:

Text sample 6: Professional letter

Furthermore, it really would be inappropriate for me to put words in your mouth. In short, you should really take the format of the resolution and put in your own thoughts [...] The association is already sampling opinion on a number of other matters and it may be possible to add this one. If it is not possible to add your concern this year, it would certainly be possible to add it next year.

Typical of texts with a large positive score on Dimension 4, this professional letter uses prediction modals to show what will be possible in the future (*it would be possible to add it next year*) or to discuss hypothetical situations (*it really would be inappropriate ...*).

Necessity modals express obligation for the addressee (*you should really ...*), and possibility modals convey the likelihood of certain events (*it may be possible ...*). Conditional subordination sets limits on the circumstances under which other actions or events may occur (*If it is not possible to add your concern this year ...*). The sample also illustrates the common use of infinitives as complements controlled by adjectives that encode the writer's attitude or stance (*inappropriate to put words in your mouth, possible to add this one*). Taken together, these features function to structure an argument, identify alternatives, present the author's stance about ideas, and directly encourage certain thinking or action on the part of others.

The distribution of registers along this dimension (Figure 38.4) shows that professional letters and editorials have a high frequency of these features, while press reviews and broadcasts have a relative absence of these features. Many registers are unmarked for this dimension, and thus cluster around 0 in Figure 38.4.

In MD work subsequent to 1988, Dimension 4 has been referred to both as 'Overt Expression of Persuasion' and 'Overt Expression of Argumentation'. Either 'persuasion' or 'argumentation' can characterize the use of these features.

5.4. Dimension 5: Abstract vs. Non-abstract Style

Dimension 5, like Dimensions 2 and 4, has only features with positive loadings. These features include conjuncts (e.g., *thus, however*), agentless passives, passives with *by*-phrases, past participle (passive) adverbial clauses, and past participle (passive) postnominal clauses (also called past participle WHIZ deletions). Most of these structures are passives, and are used to present information with little or no emphasis on the agent, as in this extract from an engineering report:

Text sample 7: Engineering report

Eventually however fatigue cracks were noticed in the roots of two of the blades and it was suspected that the lack of freedom in the drag hinges was the possible cause.

Later, after new blades had been fitted, it was thought better to run with drag hinges free and so reduce root stresses, experience having shown that the possibility of resonance was small [...] This question of blade fatigue is more fully discussed in the appendix.

This short extract contains many passive constructions. Agents of the actions are not mentioned; instead, inanimate referents are the focus of the discourse (e.g., *fatigue cracks were noticed, the question of blade fatigue is more fully discussed*). Two sentences use non-referential *it* as subject (*it was suspected, it was thought*), further eliminating mention of the animate agent. In other texts of this type, noun phrases are also often modified with past participle passive modifiers (e.g., *the exhaust air volume required by the 6-ft. x 4-ft. grid*).

The distribution of registers along this dimension (Figure 38.5) shows that academic prose and official documents are particularly marked in their use of these features. Thus, the register distribution reinforces the interpretation that this style of discourse is typically used with abstract or technical information. Conjuncts and adverbial subordinators co-occur with the passive forms to mark the logical relationships among clauses.

ABSTRACT

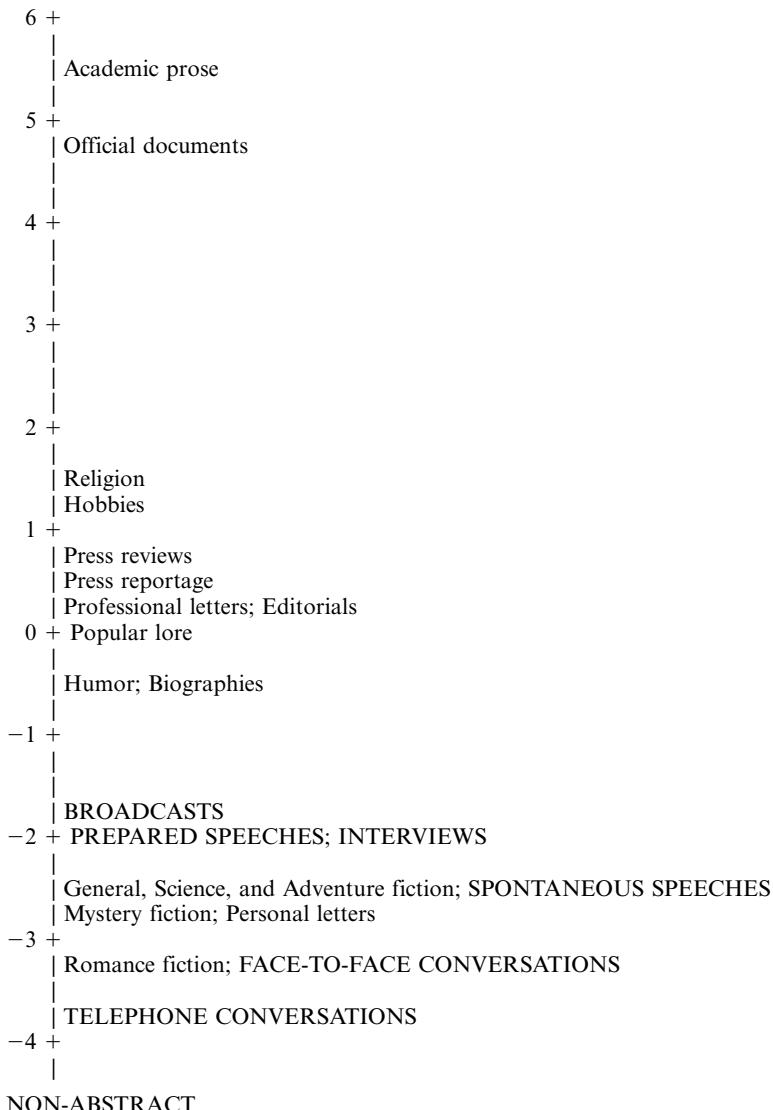


Fig. 38.5: Mean scores for registers along Dimension 5: Abstract versus Non-abstract Style. ($F = 28.8$, $p < .0001$, $r^2 = 58.0\%$)

In contrast, conversation and fiction have large negative scores, indicating an absence of these features. As text samples 2 and 3 above illustrate, the subjects of sentences in conversation and fiction are often actors, and passive constructions tend to be rare. Thus, this dimension marks a continuum of impersonal, abstract style versus a more personal, non-abstract style. (In MD studies subsequent to 1988, this dimension has also been referred to as 'Impersonal vs. Non-impersonal Style').

5.5. Overall patterns of register variation in the 1988 MD analysis

The 1988 MD analysis showed that English registers vary along several underlying dimensions associated with different functional considerations, including: interactiveness, involvement and personal stance, production circumstances, informational density, informational elaboration, narrative purposes, situated reference, persuasiveness or argumentation, and impersonal presentation of information.

Two of these dimensions have no systematic relationship to speech and writing (Dimension 2: Narrative Discourse; and Dimension 4: Argumentation). However, the other three dimensions identify sharp distinctions between ‘oral’ and ‘literate’ registers, where the term ‘oral’ is used to refer to stereotypical speech (i.e. conversation), and the term ‘literate’ is used to refer to stereotypical writing (i.e. academic prose or other kinds of formal, informational prose). On Dimension 1, conversation is at one extreme, marked as extremely involved and restricted by real time production circumstances; academic prose is at the other extreme, marked as extremely informational and carefully crafted and edited. On Dimension 3, conversation is at one extreme, marked as extremely situated in reference; academic prose is at the other extreme, marked as extremely elaborated in reference. On Dimension 5, conversation is at one extreme, marked by the absence of passive constructions; academic prose is at the other extreme, marked by impersonal styles of presentation.

Thus, the spoken and written modes can be exploited in extreme ways, resulting in register characterizations not found in the other mode. There are genuine differences in the production circumstances of speech and writing, and these differences provide the potential for styles of expression in writing that are not (normally) feasible in speech. In addition, spoken registers rarely adopt the extreme informational communicative focus of written expository registers. (Even classroom teaching is much more interactive and involved in purpose than typical written expository registers; see Biber et al. 2002). Thus, written academic prose and official documents are extremely ‘informational’ (Dimension 1), ‘elaborated in reference’ (Dimension 3), and ‘impersonal’ (Dimension 5) – extreme linguistic characterizations not found in any spoken register.

However, despite the existence of these oral/literate dimensions, no dimension identifies an absolute distinction between speech and writing. On Dimension 1, written registers can be ‘involved’ (e.g., personal letters), while spoken registers can be moderately informational (e.g., prepared speeches). And on Dimensions 3 and 5, written registers like fiction and personal letters are similar to conversation in being ‘situated’ and not ‘impersonal’.

5.6. Validation of the 1988 MD analysis

Several studies have attempted to replicate the factor analysis from the 1988 MD study, assessing the stability of that analysis and the validity of the resulting factors. Biber (1990) compares the factor structure of analyses run on various sub-corpora, finding that the factor structure is generally stable as long as the same range of variation is maintained in the target corpus. Biber (1992) uses confirmatory factor analysis to compare the goodness-of-fit for several different factorial models; de Mönnink/Brom/Oostdijk (2003) attempt to replicate the 1988 MD structure on the texts in the ICE-GB

corpus; and Lee (2000) analyzes a 4-million-word sample from the BNC, testing the influence of statistical parameters (e.g., factor extraction methods and rotation methods) and alterations to the input data (the linguistic variables and the design parameters of the corpus) on the resulting dimensions of variation.

6. Types of MD study

In practice, there are two different types of MD study: those that carry out a full MD analysis and those that apply the 1988 dimensions to new areas of research. Methodologically, the two types differ in whether or not they include a new factor analysis (steps 6 and 7 in section 4 above). Full MD studies, such as the original MD studies (Biber 1986, 1988), carry out a factor analysis to identify and interpret underlying dimensions of register variation; they thus cover all eight methodological steps. Over the years, there have been several subsequent studies that have undertaken full MD analyses of this type, identifying the underlying dimensions that operate in particular discourse domains of English and other languages (see section 8 below). However, many MD studies apply the dimensions identified in Biber (1988) to describe and compare additional registers. These studies omit steps 6 and 7; since they use the previously-identified dimensions, such studies do not require a separate factor analysis (see section 7 below).

The decision to conduct a new, complete MD analysis or to apply the 1988 dimensions depends on the research issues that are being investigated, because the two approaches will give different perspectives on register variation. Using the established dimensions allows researchers to compare new registers or specialized sub-registers to a wide range of spoken and written registers in English (i.e., the basis of the 1988 study). Other research, however, seeks to explore a particular domain and determine the dimensions of variation for that domain. Section 7 below surveys previous studies that have applied the 1988 dimensions to a range of research questions, while section 8 surveys studies that have undertaken full MD analyses. Then, section 9 briefly surveys studies that have undertaken MD analyses of languages other than English.

7. Application of the 1988 dimensions to other discourse domains

7.1. Specialized registers and discourse domains

Many studies have applied the 1988 dimensions of variation to study the linguistic characteristics of more specialized registers and discourse domains. Biber et al. (2002) describe the patterns of variation among a range of spoken and written university registers, including academic registers (e.g., lectures, office hours, study groups, textbooks, research articles) as well as non-academic articles (e.g., service encounters, course syllabi, institutional writing). However, other MD studies have usually focused on variation among written registers. For example, Biber (1987) compares the MD characteristics of matched written registers in British English and American English, showing how the AmE registers tend to be both more ‘involved’ and less ‘elaborated’ than the corresponding BrE registers. Many MD studies have focused on academic writing. For example,

Conrad (1996, 2001) compared research articles, textbooks, and student writing in two academic disciplines: biology and history. Biber/Finegan (1994b) document the systematic patterns of variation among the sections of medical research articles (i.e., introduction, methods, results, and discussion). MD studies have also been used to investigate more specialized written registers. For example, Connor/Upton (2003) describe the characteristics of direct mail letters, while Connor/Upton (2004) investigate non-profit grant proposals, comparing the different ‘move’ types used in proposals (e.g., ‘benefits’ versus ‘territory’). Two other studies have focused on author style: Connor-Linton (2001) compares the styles of authors writing about the possibility of nuclear war, while Biber/Finegan (1994a) compares the styles of 18th century authors writing essays and fiction.

Several other MD studies have focused on registers in earlier historical periods, tracking the patterns of change across time. For example, Biber/Finegan (1989, 1997) document the patterns of change for a range of written and speech-based registers, documenting a general ‘drift’ for popular written registers to become more ‘oral’, while specialist written registers have evolved to become more ‘literate’ (with respect to Dimensions 1, 3, and 5). Atkinson (1992, 1996, 1999) provides detailed descriptions of historical change for two written registers: medical research articles and scientific research articles. In these studies, Atkinson tracks the patterns of change with respect to the 1988 dimensions, and then provides detailed interpretation of those patterns relative to the socio-historical contexts of the target registers. Finally, Geisler (2002, 2003) focuses on the patterns of variation among 19th century written registers.

Fewer MD studies have focused exclusively on spoken registers. Helt (2001) compares the characteristics of BrE and AmE conversational registers, identifying parallel differences to those found for written registers (in Biber 1987; e.g., with the AmE registers being consistently more ‘involved’ than the corresponding BrE registers). Connor-Linton/Shohamy (2001) study variation with oral proficiency interviews (used for ESL assessment), showing how the language produced by students for different elicitation tasks varies systematically across dimensions. Csomay (2002) focuses specifically on academic lectures, comparing their multi-dimensional profiles across levels of instruction and interactivity.

Three MD studies have focused on dramatic dialogue. Quaglio (2004) describes how the conversations in the TV show *Friends* are very similar in their MD profiles to natural face-to-face conversations. The other two studies have added the social parameter of gender, showing how men and women talk in systematically different ways in dramatic dialogue. Rey (2001) tracks general changes in the dialogues of *Star Trek* episodes from 1966–1993, focusing especially on how women and men have been portrayed differently over time. For example, women in *Star Trek* used extremely ‘involved’ (Dimension 1) speech styles in the 1960’s but shifted to a more moderate ‘involved’ style in the most recent shows. In contrast, men used only moderately involved speech styles in the earliest shows, but they actually shifted to become slightly more involved than women in the most recent shows. Biber/Burges (2000) discuss similar kinds of patterns on a larger historical scale, considering the language of women and men as portrayed by female and male authors, in dramatic discourse over the past three centuries. This study found that the gender of the addressee was an important consideration, with both men and women using more ‘involved’ styles when speaking to a woman than when speaking to a man. Female and male authors use similarly involved styles for female speakers, but they

differ in their portrayals of male speakers: modern female authors tend to portray male speakers using ‘involved’ styles (especially when speaking to a woman), while male authors tend to portray male speakers using less involved styles.

7.2. Automatic identification of register and ‘text type’ studies

Several studies use multi-dimensional approaches to automatically determine the register or genre category of unknown texts. Biber (1993b) uses the 1988 dimensions as predictors in a discriminant analysis, while studies like Karlsgren/Cutting (1994), Beaudouin et al. (2001), and Folch et al. (2000) adopt similar approaches testing the predictive power of parameters that incorporate only linguistic features that are computationally easy to identify (see also Kessler/Nunberg/Schütze 1997; Louwerse et al. 2004). de Mönnink/Brom/Oostdijk (2003) compare the predictive power of factors based on the Biber (1988) feature set with factors derived from a larger set of word class tags. In addition, several studies have used multi-dimensional methodologies for studies of stylometry, to investigate authorship attribution (see article 50) or to automatically identify other demographic characteristics of authors (e.g., Palander-Collin 1999; Koppel/Argamon/Shimonini 2002).

A complementary perspective on textual variation is to identify and interpret the text categories that are *linguistically* well defined, referred to as *text types*. Text type distinctions have no necessary relation to register distinctions. Rather, text types are defined such that the texts within each type are maximally similar in their linguistic characteristics, regardless of their situational/register characteristics. However, because linguistic features have strong functional associations, text types can be interpreted in functional terms.

In the MD approach, text types are identified quantitatively using cluster analysis, with the dimensions of variation as predictors. Cluster analysis groups texts into ‘clusters’ on the basis of shared multi-dimensional (linguistic) characteristics: the conversations grouped in a cluster are maximally similar linguistically, while the different clusters are maximally distinguished.

Biber (1989, 1995) describes the text types in a general corpus of spoken and written English texts. Text types and registers represent complementary ways of dissecting the textual space of a language. Text types and registers are similar in that both can be described in linguistic and in situational/functional terms. However, the two constructs differ in their primary bases: registers are defined in terms of their situational characteristics, while text types are defined linguistically. Thus, a single text type can include texts from several different registers. For example, the ‘involved persuasion’ text type in the 1989 study is defined primarily by extremely large positive scores on Dimension 4 (‘Overt Expression of Persuasion’); all texts grouped into this text type share these linguistic characteristics, even though they come from 14 different registers (including interviews, spontaneous speeches, academic prose, professional letters, and personal letters). Similarly, texts from a single register can be distributed across multiple text types. For example, academic prose texts are distributed across four text types in the 1989 study: ‘scientific exposition’, ‘learned exposition’, ‘general reported exposition’, and ‘involved persuasion’. This analytical approach has also been used to identify text types in more restricted discourse domains (see section 8 below) and in other languages (Somali; see section 9 below).

8. Other MD analyses of English registers

As noted in 6 above, several studies have undertaken new MD analyses to determine the dimensions of variation operating in a particular discourse domain (i.e., including Steps 6 and 7 listed in section 3 above). An early study of this type is Grabe (1987), who investigated the dimensions of variation for written expository registers. Connor-Linton (1989) identified dimensions of variation in an extremely restricted discourse domain, comparing the conversational styles of Soviet and American participants in ‘Space-bridge’ interactions on TV talk shows. Meurman-Solin (1993) identified dimensions of variation in a corpus of early Scottish prose texts (1450–1700). White (1994) investigated the language of job interviews, comparing the language of interviewers and interviewees, while Reppen (2001) compared the MD characteristics of elementary school spoken and written registers. de Mönnink/Brom/Oostdijk (2003) analyze the dimensions in the ICE-GB corpus, comparing the multi-dimensional models derived from three different feature sets (the Biber 1988 set of 67 features; a set of 129 word class tags; and a set of 103 sentence structures).

Biber (1992) adopted a somewhat different approach, using confirmatory factor analysis to compare the ‘goodness-of-fit’ of several different multi-dimensional models for discourse complexity features in English. More recently, Biber (2001) identifies six dimensions of variation for 18th century written and speech-based registers, while Biber (2006) identifies the dimensions operating among university spoken and written registers.

Other recent studies carry out a complete MD analysis coupled with text type analysis (see section 7.2. above). For example, Biber (to appear) uses factor analysis to identify the dimensions of variation operating in a corpus of conversational texts, and then uses cluster analysis to identify the conversation text types that are well-defined in terms of those dimensions. Other studies go a step further: they first segment texts into topically coherent discourse segments. These discourse segments are then used as the ‘texts’ in a factor analysis and cluster analysis, to identify the discourse unit types that are well defined linguistically (see Csomay 2004; Biber/Jones 2005; Biber/Connor/Upton 2007).

It is interesting to compare the kinds of dimensions identified in these studies. Given that each of these studies is based on a different corpus of texts, representing a different discourse domain, it is reasonable to expect that they would each identify a unique set of dimensions. This expectation is reinforced by the fact that the more recent studies have included additional linguistic features not used in earlier MD studies (e.g., semantic classes of nouns and verbs). However, despite these differences in design and research focus, there are certain striking similarities in the set of dimensions identified by these studies.

Most importantly, in nearly all of these studies, the first dimension identified by the factor analysis is associated with an informational focus versus a personal focus (personal involvement/stance, interactivity, and/or real time production features). Table 38.4 summarizes the major features that define Dimension 1 in each of these studies.

It is perhaps not surprising that Dimension 1 in the original 1988 MD analysis was strongly associated with an informational versus (inter)personal focus, given that the corpus in that study ranged from spoken conversational texts to written expository texts. For the same reason, it is somewhat predictable that a similar dimension would have emerged from the 2001 study of 18th century written and speech-based registers, and the 2006 study of university spoken and written registers (although the corpus studied

Tab. 38.4: Comparison of Dimension 1 across multi-dimensional studies of English subsequent to Biber (1988)

Study	Corpus	Linguistic features defining the dimension
White 1994	job interviews	long words, nouns, nominalizations, prepositions, WH questions, 2nd person pronouns versus 1st person pronouns, contractions, adverbs, discourse particles, emphatics, etc.
Reppen 1994	elementary school registers	nouns, long words, nominalizations, passives, attributive adjs., prepositions versus initial <i>and</i> , time adverbials, 3rd person pronouns
Biber 2001	18th c. written and speech-based registers	prepositions, passives, nouns, long words, past tense verbs versus 1st and 2nd person pronouns, present tense, possibility and prediction modals, <i>that</i> -deletion, mental verbs, emphatics
Biber 2006	university spoken and written registers	nominalizations, long words, nouns, prepositions, abstract nouns, attributive adjectives, passives, stance noun + <i>to</i> -clause, etc. versus contractions, demonstrative pronouns, <i>it</i> , 1st person pronouns, present tense, time advs, <i>that</i> -omission, WH-questions, etc.
Biber, to appear	conversations	long words, nominalizations, prepositions, abstract nouns, relative clauses, attributive adjs. Versus contractions, 1st and 2nd person pronouns, activity verbs

for each of those two studies was more specialized than the general corpus in the 1988 study). However, it was completely unexpected that a similar oral/literate dimension – realized by essentially the same set of co-occurring linguistic features – would be fundamentally important in highly restricted discourse domains, including studies of job interviews, elementary school registers, and conversations.

A second parameter found in most MD analyses corresponds to narrative discourse, reflected by the co-occurrence of features like past tense, 3rd person pronouns, perfect aspect, and communication verbs (see, e. g., the Biber (2006) study of university registers; Biber 2001 on 18th century registers; and the Biber (to appear) study of conversation text types). In some studies, a similar narrative dimension emerged with additional special characteristics. For example, in Reppen's (2001) study of elementary school registers, 'narrative' features like past tense, perfect aspect, and communication verbs co-occurred with once-occurring words and a high type/token ratio; in this corpus, history textbooks rely on a specialized and diverse vocabulary to narrate past events. In the job interview corpus (White 1994), the narrative dimension reflected a fundamental opposition between personal/specific past events and experiences (past tense verbs co-occurring with 1st person singular pronouns) versus general practice and expectations (present tense verbs co-occurring with 1st person plural pronouns).

At the same time, most of these studies have identified some dimensions that are unique to the particular discourse domain. For example, the factor analysis in Reppen (2001) identified a dimension of 'Other-directed idea justification' in elementary student registers. The features on this dimension include 2nd person pronouns, conditional

clauses, and prediction modals; these features commonly co-occur in certain kinds of student writing (e.g., *If you wanted to watch TV a lot you would not get very much done*).

The factor analysis in Biber's (2006) study of university spoken and written registers identified four dimensions. Two of these are similar linguistically and functionally to dimensions found in other MD studies: Dimension 1: 'Oral vs. Literate Discourse'; and Dimension 3: 'Narrative Orientation'. However, the other two dimensions are specialized to the university discourse domain: Dimension 2 is interpreted as 'Procedural vs. Content-focused Discourse'. The co-occurring 'procedural' features include modals, causative verbs, 2nd person pronouns, and verbs of desire + *to*-clause; these features are especially common in classroom management talk, course syllabi, and other institutional writing. The complementary 'content-focused' features include rare nouns, rare adjectives, and simple occurrence verbs; these co-occurring features are typical of textbooks, and especially common in natural science textbooks. Dimension 4, interpreted as 'Academic stance', consists of features like stance adverbials (factual, attitudinal, likelihood) and stance nouns + *that*-clause; classroom teaching and classroom management talk is especially marked on this dimension.

A final example comes from Biber's (to appear) MD analysis of conversational text types, which identified a dimension of 'stance-focused versus context-focused discourse'. Stance focused conversational texts were marked by the co-occurrence of *that*-deletions, mental verbs, factual verb + *that*-clause, likelihood verb + *that*-clause, likelihood adverbs, etc. In contrast, context-focused texts had high frequencies of nouns and *WH*-questions, used to inquire about past events or future plans. The text type analysis identified different sets of conversations characterized by one or the other of these two extremes.

In sum, studies that have incorporated complete MD analyses of English registers (i.e., including a new factor analysis) have uncovered both surprising similarities and notable differences in the underlying dimensions of variation. Two parameters seem to be fundamentally important, regardless of the discourse domain: a dimension associated with informational focus versus (inter)personal focus, and a dimension associated with narrative discourse. At the same time, these MD studies have uncovered dimensions particular to the communicative functions and priorities of each different domain of use. The following section shows that similar patterns have emerged from MD studies of languages other than English.

9. MD analyses of other languages

In addition to the MD studies of English surveyed in sections 5–8 above, there have been several MD studies of other languages. A few of these have attempted to apply the 1988 dimensions for English to describe register variation in other languages (see, e.g., the study on Spanish carried out by Lux/Grabe 1991). However, most MD studies of other languages have recognized the need to analyze the range of linguistic devices actually found in the target language, and to carry out independent factor analyses to identify the ways in which features actually co-occur in that language. For example, Sáiz (1999) built a corpus of parallel English-Spanish expository texts and then carried out a separate MD analysis for each language, comparing the dimensions across languages. Two more recent studies have also reported on MD analyses of register variation in Spanish: Parodi (2005), and Biber et al. (2006).

Four non-western languages have been studied to date: Besnier's (1988) analysis of Nukulaelae Tuvaluan; Kim's (Kim/Biber 1994) analysis of Korean; Biber/Hared's (1992) analysis of Somali; and Jang's (1998) study of Taiwanese. Taken together, these studies provide the first comprehensive investigations of register variation in non-western languages.

Biber (1995) synthesizes these studies to investigate the extent to which the underlying dimensions of variation and the relations among registers are configured in similar ways across languages. These languages show striking similarities in their basic patterns of register variation, as reflected by:

- the co-occurring linguistic features that define the dimensions of variation in each language;
- the functional considerations represented by those dimensions; and
- the linguistic/functional relations among analogous registers.

For example, similar to the full MD analyses of English, these MD studies have all identified dimensions associated with informational versus (inter)personal purposes, and with narrative discourse.

At the same time, each of these MD analyses have identified dimensions that are unique to a language, reflecting the particular communicative priorities of that language and culture. For example, the MD analysis of Somali identified a dimension interpreted as 'Distanced, directive interaction', represented by optative clauses, 1st and 2nd person pronouns, directional pre-verbal particles, and other case particles. Only one register is especially marked for the frequent use of these co-occurring features in Somali: personal letters. This dimension reflects the particular communicative priorities of personal letters in Somali, which are typically interactive as well as explicitly directive.

The cross-linguistic comparisons further show that languages as diverse as English and Somali have undergone similar patterns of historical evolution following the introduction of written registers. For example, specialist written registers in both languages have evolved over time to styles with an increasingly dense use of noun phrase modification. Historical shifts in the use of dependent clauses is also surprising: in both languages, certain types of clausal embedding – especially complement clauses – turn out to be associated with spoken registers rather than written registers.

There are important possible confounding influences that must be considered when interpreting cross-linguistic MD comparisons. One consideration has to do with corpus design: do the corpora include the same range of spoken and written registers?

A second possible confounding influence for cross-linguistic comparisons is that each of these languages has a different inventory of structural devices and distinctions. The analytical goal in each case has been to include the full range of structural/functional distinctions found in the target language. However, the multi-dimensional patterns for each language reflect a complex interaction between the available structural resources and the register distinctions that are systematically marked by those resources. For example, the existence of subjunctive mood verbs in Spanish provides the linguistic resources for a dimension associated with irrealis discourse. Similarly, the existence of two past tenses in Spanish provides the structural resources for a specialized dimension associated with informational reports of past events. In the Korean MD analysis, personal stance features are grouped on one dimension, while features of honorification and self-humbling are grouped on a separate dimension.

The existence of structural distinctions does not necessarily entail the existence of systematic register differences, but previous MD analyses show that languages/cultures have often evolved to take advantage of these linguistic resources. However, these analyses have further shown that the ways in which a language/culture exploits such structural resources are not always what we would have anticipated. For example in Korean, the co-occurring features associated with the ‘stance’ dimension (e.g., emphatics, hedges, other epistemic and attitudinal features) are especially common in the (inter)personal registers, including all conversations and personal letters. In contrast, the features associated with the honorific/self-humbling dimension are especially common only in the *public* spoken registers, such as public interviews and public speeches. Both dimensions are generally related to the expression of stance. However, the MD analysis shows that they are exploited in different ways for specific cultural purposes.

These patterns illustrate the general finding that structural resources come to be exploited in particular (often unanticipated) ways in particular cultures. Some linguistic features are distributed widely across different languages, and they are exploited in very similar – possibly universal – ways to distinguish among registers across cultures. For example, features like 1st and 2nd person pronouns, questions, reduced/contracted forms, and simple hedging or emphatic stance features are found in many languages, and the MD analyses carried out to date indicate that these features tend to co-occur cross-linguistically associated with conversation and other (inter)personal spoken registers. Similarly, nouns, adjectives, and various kinds of nominal modifiers are found in many languages, and they tend to co-occur cross-linguistically associated with formal expository writing. In contrast, other linguistic resources are more specialized, occurring in comparatively few languages, and these resources have come to be exploited for more specialized and more distinctive dimensions of register variation.

10. Conclusion

The present article has briefly introduced the goals and methodologies of MD analysis, and surveyed the major studies undertaken with this approach. Two general patterns are especially noteworthy from these MD studies: 1) the extent to which similar dimensions of variation operate within different languages and within specific discourse domains; and 2) the specialized dimensions that are peculiar to a particular language or discourse domain. The first kind of finding relates to the possibility of universals of register variation. For example, based on prior MD studies, we could make the strong hypothesis that the texts in any language will vary systematically along at least two dimensions: one associated with an informational versus (inter)personal focus, and one associated with narrative versus non-narrative discourse. The second kind of finding relates to the distinctiveness of every language and every discourse domain, reflecting the unique communicative priorities and situational circumstances of the language/domain. These differences are systematically reflected in the dimensions of variation that exist in the language/domain, and in the relations among registers defined by each dimension. In fact, a comparison of the multi-dimensional profiles of different languages might indicate the different communicative priorities of those languages, because it reflects the way in which each one allocates linguistic resources for functional purposes (see Biber 1995, especially 264–270).

Obviously, much further research of this kind is required to confirm the existence of universal patterns of register variation, and to investigate the range of more specialized dimensions found in various languages and domains. The MD studies carried out to date, however, indicate the importance of this research approach, and the feasibility of achieving these goals through further studies of this type.

11. Literature

- Atkinson, D. (1992), The Evolution of Medical Research Writing from 1735 to 1985: The Case of the Edinburgh Medical Journal. In: *Applied Linguistics* 13, 337–374.
- Atkinson, D. (1996), The Philosophical Transactions of the Royal Society of London, 1675–1975: A Sociohistorical Discourse Analysis. In: *Language in Society* 25, 333–371.
- Atkinson, D. (1999), *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Beaudouin, V./Fleury, S./Habert, B./Illouz, G./Licoppe, C./Pasquier, M. (2001), Typweb: Decrire la toile pour mieux comprendre les parcours. In: *Colloque international sur les usages et les services des télécommunications (CIUST01)*. Paris, France. Available at: <http://www.cavi.univparis3.fr/ilpga/ilpga/sfleury/typweb.htm>.
- Besnier, N. (1988), The Linguistic Relationships of Spoken and Written Nukulaelae Registers. In: *Language* 64, 707–736.
- Biber, D. (1985), Investigating Macroscopic Textual Variation through Multi-feature / Multi-dimensional Analyses. In: *Linguistics* 23, 337–360.
- Biber, D. (1986), Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. In: *Language* 62, 384–414.
- Biber, D. (1987), A Textual Comparison of British and American Writing. In: *American Speech* 62, 99–119.
- Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. (1989), A Typology of English Texts. In: *Linguistics* 27, 3–43.
- Biber, D. (1990), Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. In: *Literary and Linguistic Computing* 5, 257–269.
- Biber, D. (1992), On the Complexity of Discourse Complexity: A Multidimensional Analysis. In: *Discourse Processes* 15, 133–163. Reprinted in: Conrad/Biber 2001, 215–240.
- Biber, D. (1993a), Representativeness in Corpus Design. In: *Literary and Linguistic Computing* 8, 1–15.
- Biber, D. (1993b), Using Register-diversified Corpora for General Language Studies. In: *Computational Linguistics* 19, 219–241.
- Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
- Biber, D. (2001), Dimensions of Variation among 18th Century Registers. In: Diller, H.-J./Görlach, M. (eds.), *Towards a History of English as a History of Genres*. Heidelberg: C. Winter, 89–110. Reprinted in: Conrad/Biber 2001, 200–214.
- Biber, D. (2006), *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, D. (to appear), Corpus-based Analyses of Discourse: Dimensions of Variation in Conversation. In: Jones, R./Bhatia, V./Flowerdew J. (eds.), *Advances in Discourse Studies*. Routledge.
- Biber, D./Burges, J. (2000), Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue. In: *Journal of English Linguistics* 28, 21–37. Reprinted in: Conrad/Biber 2001, 157–170.
- Biber, D./Connor, U./Upton, T. (2007), *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

- Biber, D./Conrad, S./Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D./Conrad, S./Reppen, R./Byrd, P./Helt, M. (2002), Speaking and Writing in the University: A Multi-dimensional Comparison. In: *TESOL Quarterly* 36, 9–48.
- Biber, D./Conrad, S./Reppen, R./Byrd, P./Helt, M. (2003), Strengths and Goals of Multi-dimensional Analysis: A Response to Ghadessy. In: *TESOL Quarterly* 37, 151–155.
- Biber, D./Davies, M./Jones, J. K./Tracy-Ventura, N. (2006), Spoken and Written Register Variation in Spanish: A Multi-dimensional Analysis. In: *Corpora* 1, 7–38.
- Biber, D./Finegan, E. (1989), Drift and the Evolution of English Style: A History of Three Genres. In: *Language* 65, 487–515.
- Biber, D./Finegan, E. (1994a), Multi-dimensional Analyses of Authors' Styles: Some Case Studies from the Eighteenth Century. In: Ross, D./Brink, D. (eds.), *Research in Humanities Computing*, Vol. III. Oxford: Oxford University Press, 3–17.
- Biber, D./Finegan, E. (1994b), Intra-textual Variation within Medical Research Articles. In: Oostdijk, N./de Haan, P. (eds.), *Corpus-based Research into Language*. Amsterdam: Rodopi, 201–222. Reprinted in: Conrad/Biber 2001, 108–123.
- Biber, D./Finegan, E. (1997), Diachronic Relations among Speech-based and Written Registers in English. In: Nevalainen, T./Kahlas-Tarkka, L. (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 253–276. Reprinted in: Conrad/Biber 2001, 66–83.
- Biber, D./Hared, M. (1992), Dimensions of Register Variation in Somali. In: *Language Variation and Change* 4, 41–75.
- Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Biber, D./Jones, J. K. (2005), Merging Corpus Linguistic and Discourse Analytic Research Goals: Discourse Units in Biology Research Articles. In: *Corpus Linguistics and Linguistic Theory* 1, 151–182.
- Brown, P./Fraser, C. (1979), Speech as a Marker of Situation. In: Scherer, K. R./Giles, H. (eds.), *Social Markers in Speech*. Cambridge: Cambridge University Press, 33–62.
- Connor, U./Upton, T. (2003), Linguistic Dimensions of Direct Mail Letters. In: Meyer, C./Leistyna, P. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 71–86.
- Connor, U./Upton, T. (2004), The Genre of Grant Proposals: A Corpus Linguistic Analysis. In: Connor, U./Upton, T. (eds.), *Discourse in the Professions*. Amsterdam: John Benjamins, 235–256.
- Connor-Linton, J. (1989), Crosstalk: A Multi-feature Analysis of Soviet-American Spacebridges. PhD dissertation, University of Southern California.
- Connor-Linton, J. (2001), Authors' Style and World-view: A Comparison of Texts about Nuclear Arms Policy. In: Conrad/Biber 2001, 84–93.
- Connor-Linton, J./Shohamy, E. (2001), Register Variation, Oral Proficiency Sampling, and the Promise of Multi-dimensional Analysis. In: Conrad/Biber 2001, 124–137.
- Conrad, S. (1996), Investigating Academic Texts with Corpus-based Techniques: An Example from Biology. In: *Linguistics and Education* 8, 299–326.
- Conrad, S. (2001), Variation among Disciplinary Texts: A Comparison of Textbooks and Journal Articles in Biology and History. In: Conrad/Biber 2001, 94–107.
- Conrad, S./Biber, D. (eds.) (2001), *Variation in English: Multi-dimensional Studies*. Harlow/London: Pearson Education.
- Csomay, E. (2002), Variation in Academic Lectures: Interactivity and Level of Instruction. In: Reppen, R./Fitzmaurice, S. M./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 203–224.
- Csomay, E. (2004), Linguistic Variation within University Classroom Talk: A Corpus-based Perspective. In: *Linguistics and Education* 15, 243–274.

- de Mönnink, I./Brom, N./Oostdijk, N. (2003), Using the MF/MD Method for Automatic Text Classification. In: Granger, S./Petch-Tyson, S. (eds.), *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam: Rodopi, 15–26.
- Ervin-Tripp, S. (1972), On Sociolinguistic Rules: Alternation and Co-occurrence. In: Gumperz, J. J./Hymes, D. (eds.), *Directions in Sociolinguistics*. New York: Holt, 213–250.
- Ferguson, C. A. (1983), Sports Announcer Talk: Syntactic Aspects of Register Variation. In: *Language in Society* 12, 153–172.
- Folch, H./Heiden, S./Habert, B./Fleury, S./Illouz, G./Lafon, P./Nioche, J./Prévost, S. (2000), Typtex: Inductive Typological Text Classification by Multivariate Statistical Analysis for NLP Systems Tuning/Evaluation. In: *Proceedings of the Second Language Resources and Evaluation Conference*. Athens, Greece, 141–148.
- Geisler, C. (2002), Investigating Register Variation in Nineteenth-century English: A Multi-dimensional Comparison. In: Reppen, R./Fitzmaurice, S. M./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 249–271.
- Geisler, C. (2003), Gender-based Variation in Nineteenth-century English Letter Writing. In: Meyer, C./Leistyna, P. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 87–106.
- Gorsuch, Richard L. (1983), *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.
- Grabe, W. (1987), Contrastive Rhetoric and Text-type Research. In: Connor, U./Kaplan, R. B. (eds.), *Writing across Languages: Analysis of L2 Text*. Reading, MA: Addison-Wesley, 115–138.
- Helt, M. (2001), A Multi-dimensional Comparison of British and American Spoken English. In: Conrad/Biber 2001, 171–184.
- Hymes, D. (1984), Sociolinguistics: Stability and Consolidation. In: *International Journal of the Sociology of Language* 45, 39–45.
- Jang, S-C. (1998), Dimensions of Spoken and Written Taiwanese: A Corpus-based Register Study. PhD dissertation, University of Hawaii.
- Karlsgren, J./Cutting, D. (1994), Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: *Proceedings of Coling 94*. Kyoto, Japan, 1071–1075. Reprinted in: Karlsgren, J. (2000), *Stylistic Experiments for Informational Retrieval*, chapter 7. PhD dissertation, Stockholm University.
- Kessler, B./Nunberg, G./Schütze, H. (1997), Automatic Detection of Text Genre. In: *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 32–38.
- Kim, Y./Biber, D. (1994), A Corpus-based Analysis of Register Variation in Korean. In: Biber, D./Finegan, E. (eds.), *Sociolinguistic Perspectives on Register*. New York: Oxford University Press, 157–181.
- Koppell, M./Argamon, S./Shimoni, A. R. (2002), Automatically Categorizing Written Texts by Author Gender. In: *Literary and Linguistic Computing* 17, 401–412.
- Lee, D. (2000), Modelling Variation in Spoken and Written English: The Multi-dimensional Approach Revisited. PhD dissertation, Lancaster University.
- Louwerse, M./McCarthy, P. M./McNamara, D. S./Graesser, A. (2004), Variation in Language and Cohesion across Written and Spoken Registers. In: Forbus, K./Gentner, D./Regier, T. (eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, 843–848.
- Lux, P./Grabe, W. (1991), Multivariate Approaches to Contrastive Rhetoric. In: *Lenguas Modernas* 18, 133–160.
- Meurman-Solin, A. (1993), *Variation and Change in Early Scottish Prose*. (Dissertationes Humanarum Litterarum 65.) Helsinki: Suomalainen Tiedeakatemia.
- Palander-Collin, M. (1999), Male and Female Styles in 17th Century Correspondence. In: *Language Variation and Change* 11, 123–141.
- Parodi, G. (2005), Lingüística de corpus y análisis multidimensional: Exploración de la variación en el corpus PUCV-2003. In: Parodi, G. (ed.), *Discurso Especializado e Instituciones Formadoras*. Valparaíso, Chile: Ediciones Universitarias de Valparaíso, 83–126.

- Quaglio, P. M. (2004), The Language of NBC's Friends: A Comparison with Face-to-face Conversation. PhD dissertation, Northern Arizona University.
- Reppen, R. (1994), Variation in Elementary Student Language: A Multi-dimensional Perspective. Unpublished PhD dissertation, Northern Arizona University, Flagstaff, AZ.
- Reppen, R. (2001), Register Variation in Student and Adult Speech and Writing. In: Conrad/Biber 2001, 187–199.
- Rey, J. M. (2001), Changing Gender Roles in Popular Culture: Dialogue in the *Star Trek* episodes from 1966 to 1993. In: Conrad/Biber 2001, 138–156.
- Sáiz, M. (1999), A Cross-linguistic Corpus-based Analysis of Linguistic Variation. Manchester: PhD dissertation, UMIST.
- White, M. (1994), Language in Job Interviews: Differences Relating to Success and Socioeconomic Variables. PhD dissertation, Northern Arizona University.

Douglas Biber, Flagstaff, AZ (USA)

39. Machine learning

1. Introduction
2. A brief machine learning primer
3. Applying machine learning to annotated corpora
4. Machine learning for corpus linguistics
5. Summary and conclusion
6. Literature

1. Introduction

In corpus linguistics the computer can play various roles as assistant to the corpus linguist, alleviating some of the linguist's tasks. This article describes methods which can automatically learn to assign linguistic annotations to digital corpora, on the basis of example annotations presented to them in a training phase. If successful, these methods can generate annotation layers automatically at superhuman speeds; alternatively, they can be integrated into the human annotation process as automatic pre-annotators that do at least part of the work, or post-annotators that search for inconsistencies in the annotations. These methods come from machine learning, a subfield of artificial intelligence.

Research in machine learning focuses on computer programs that learn from experience. Learning is a manifold concept, and is usually indirectly quantified as the measurable improvement of a learner on some task after learning has taken place (Mitchell 1997). The role of machine learning this article focuses on is that of a means to automate knowledge discovery and linguistic modeling on the basis of annotated corpus material, where the annotation is the target of learning. Annotation layers usually represent some abstraction over the text of the corpus, ranging from word-level abstractions (e. g. part-of-speech tags) to text-level abstractions (e. g. topical text categorization).

- Quaglio, P. M. (2004), The Language of NBC's Friends: A Comparison with Face-to-face Conversation. PhD dissertation, Northern Arizona University.
- Reppen, R. (1994), Variation in Elementary Student Language: A Multi-dimensional Perspective. Unpublished PhD dissertation, Northern Arizona University, Flagstaff, AZ.
- Reppen, R. (2001), Register Variation in Student and Adult Speech and Writing. In: Conrad/Biber 2001, 187–199.
- Rey, J. M. (2001), Changing Gender Roles in Popular Culture: Dialogue in the *Star Trek* episodes from 1966 to 1993. In: Conrad/Biber 2001, 138–156.
- Sáiz, M. (1999), A Cross-linguistic Corpus-based Analysis of Linguistic Variation. Manchester: PhD dissertation, UMIST.
- White, M. (1994), Language in Job Interviews: Differences Relating to Success and Socioeconomic Variables. PhD dissertation, Northern Arizona University.

Douglas Biber, Flagstaff, AZ (USA)

39. Machine learning

1. Introduction
2. A brief machine learning primer
3. Applying machine learning to annotated corpora
4. Machine learning for corpus linguistics
5. Summary and conclusion
6. Literature

1. Introduction

In corpus linguistics the computer can play various roles as assistant to the corpus linguist, alleviating some of the linguist's tasks. This article describes methods which can automatically learn to assign linguistic annotations to digital corpora, on the basis of example annotations presented to them in a training phase. If successful, these methods can generate annotation layers automatically at superhuman speeds; alternatively, they can be integrated into the human annotation process as automatic pre-annotators that do at least part of the work, or post-annotators that search for inconsistencies in the annotations. These methods come from machine learning, a subfield of artificial intelligence.

Research in machine learning focuses on computer programs that learn from experience. Learning is a manifold concept, and is usually indirectly quantified as the measurable improvement of a learner on some task after learning has taken place (Mitchell 1997). The role of machine learning this article focuses on is that of a means to automate knowledge discovery and linguistic modeling on the basis of annotated corpus material, where the annotation is the target of learning. Annotation layers usually represent some abstraction over the text of the corpus, ranging from word-level abstractions (e. g. part-of-speech tags) to text-level abstractions (e. g. topical text categorization).

In machine learning terms, this article focuses on supervised classification methods, complementing article 40 on ‘clustering techniques’, i.e., unsupervised methods, which do not assume annotation layers and rather focus on the text itself.

This article aims to provide a brief history and review of the state of the art in supervised classification methods from machine learning applied to learning natural language processing (NLP) tasks on the basis of annotated corpora. The article is structured as follows. Section 2 offers a primer on machine learning. Section 3 explores the ways in which corpora and machine learning can be combined to produce natural language processing models. Section 4 reviews the use of machine-learning-based methods for automatic and semi-automatic corpus annotation. To conclude, I review two issues regarding mismatching assumptions of machine learning algorithms on distributions in language data in section 5.

2. A brief machine learning primer

The prime goal of machine learning is to develop automatic learning algorithms by which a computer can learn to perform real-world tasks, not by being told (programmed) beforehand how the problem is solved, but by discovering the solution on the basis of examples. A typical interpretation of this goal is that a machine learning algorithm, presented with examples of correct executions of the task to be learned, should be able to construct an abstracting model that solves the problem that underlies the task. This means that the algorithm cannot simply resort to remembering the examples it is given as example material; rather, it should somehow be able to predict the outcome of new, unseen cases of the same problem. Finding appropriate models that capture the solution to a problem is, in general and in essence, a massively large search problem. Machine learning offers methods that search this massive space of possible abstractions in different ways.

Machine learning algorithms are typically developed to be applicable in wide ranges of real-world tasks. Algorithms that are trained to perform medical diagnoses, object recognition in images, or DNA sequencing, can also be trained to perform natural language processing tasks. In all cases, the prerequisites are merely (1) that there are examples to learn from, and (2) that these examples are represented in such a way that the learner has all the important information to solve the problem. To meet requirement (1) in the corpus linguistics context means that trustworthy annotated linguistic data is available in sufficient amounts; to meet requirement (2), typically some degree of linguistic knowledge is needed. To take a concrete example, consider a corpus linguist who wants to add part-of-speech tags to all the words in his large digital corpus of texts, containing a hundred million words. A team of human annotators has already assigned part-of-speech tags to one hundred thousand words, but tagging the rest manually would take too long. Here machine learning can step in. On the basis of the first hundred thousand annotated words, a machine learning algorithm can learn to tag individual words in context. The hundred thousand words with their respective manually assigned parts-of-speech are converted into one hundred thousand examples of single words in a context of, for example, two neighboring words to the left and two to the right, mapping to the appropriate tag of the focus word. The limited linguistic knowledge employed here is that in part-of-speech tagging, a local context usually offers enough information to disambiguate between the possible tags the focus word can have.

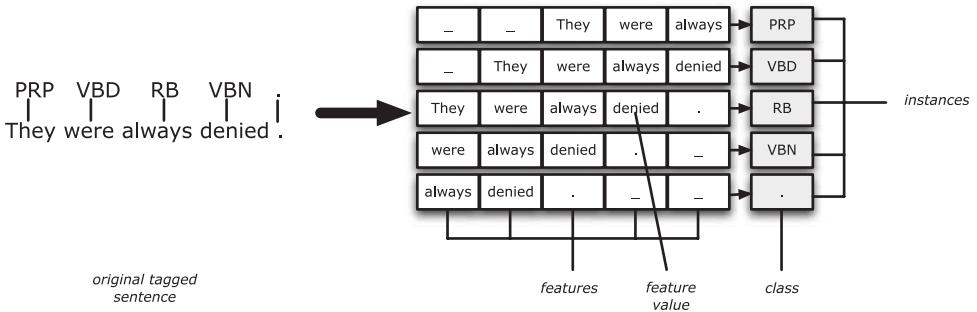


Fig. 39.1: Example conversion of a part-of-speech tagged sentence to five instances. Each instance represents a window of five adjacent words, the feature values, surrounding a focus word. The class of each instance is the part-of-speech tag of the focus word

In machine learning terms, the task is represented by one hundred thousand *instances*, each encoded by five *features*, i. e. the focus word and its four left and right neighboring words, mapping to a *class* representing the part-of-speech tag of the focus word. Figure 39.1 displays an example of a five-word sentence with part-of-speech tags, converted into five *windowed* classification instances. The example in the figure, and the part-of-speech tags, are taken from the Penn Treebank (Marcus/Santorini/Marcinkiewicz 1993). The figure illustrates that the term *feature* refers to the relative position in the *window* around the focus position, while in each instance particular *feature values*, i. e., words, fill each position. Based on a training set of classifications, a machine learning algorithm can generate a model of the classification task, by which it can then tag new text. Processing new text means that the new text is also converted into windowed instances, now with unknown classes. Machine learning algorithms differ widely in how they construct their models and how they use these models to classify new instances.

Machine learning has been known under that name since the 1970s, when it developed as a separate branch in artificial intelligence, out of earlier work on pattern recognition on the one hand, and knowledge discovery on the other hand. A reader of classic papers is edited by Shavlik/Dietterich (1990). Both pattern recognition and knowledge discovery were developed as alternatives to knowledge-based or expert-based methods (i. e., methods in which experts are able to explain how the problem is structured so that a solution can be programmed). Even though experts can sometimes formalize solutions to problems, these formalizations often turn out to be incomplete, or inefficient (e. g. too slow) when executed on a computer.

Relevant for the current context is the fact that this situation applies to many problems in language processing. It would appear from many linguistic text books that most issues in many languages, and in natural language in general, have been acknowledged, explored, and laid down in rule systems. Nevertheless, many rule systems in natural language only cover a certain “core” of cases, leaving a “periphery” unexplained. Many rule systems defer to a system of markedness, or exception lexicons, to store information not covered by the rules. Also, many phenomena in language are tied to lexical issues such as idioms, multi-word units, word senses, and lexical subcategorization preferences that defy abstraction by rules.

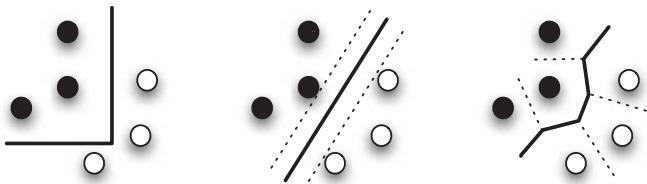


Fig. 39.2: Schematic visualization of three prototypical ways in machine learning to draw decision boundaries to separate two class spaces, given examples of two classes (the black and white circles). Left: axis-orthogonal decision boundaries. Middle: a maximum-margin hyperplane. Right: Voronoi tiles

Hence, the field of natural language processing has to some extent turned its attention to applying machine learning algorithms to natural language processing tasks. Part of this attention is spent on figuring out which type of learning strategy is most suited to particular problems. Each machine learning algorithm is equipped with a particular *bias*, a preset constrained perspective on how its abstraction is going to be shaped. Figure 39.2 schematically visualizes three prototypical biases that all make use of decision boundaries, the bold lines, separating examples from one concept or class (black) from the other (white). Earlier we presented the example of part-of-speech tagging where in fact the task is to classify instances into many different classes, i. e., all possible part-of-speech tags. A linguistic example distinction between two classes is the decision whether *with a fork* attaches to *eat* or *pizza* in the English sentence *Eleni eats pizza with a fork* – the problem of prepositional-phrase attachment (Ratnaparkhi/Reynar/Roukos 1994).

The decision boundaries attempt to approximate the real but unknown boundary that separates the black class space from the white. On the left, the decision boundary is composed of two axis-orthogonal line pieces. Every example left of the vertical line and above the horizontal line is black; all other examples are white. This type of decision boundary is typical for decision tree induction algorithms, such as the classical ID3 (Quinlan 1986), its successor C4.5 (Quinlan 1993), and CART (Breiman et al. 1984), and rule learners such as CN2 (Clark/Niblett 1989), IREP (Fürnkranz/Widmer 1994), and RIPPER (Cohen 1995).

In the middle, the black and white examples are separated by one oblique separator. Several oblique separators, or hyperplanes in spaces of higher dimensions than the two dimensions of the figure, could be drawn between the two groups of examples, but the particular line drawn in the figure is the line with the maximal amount of margin (demarcated by the dotted lines to the left and right) between the two groups. Example machine learning algorithms that employ hyperplanes are perceptrons (Rosenblatt 1958), multi-layered perceptrons (Rumelhart/Hinton/Williams 1986), Winnow networks (Littlestone 1988), and Support Vector Machines (Cortes/Vapnik 1995).

On the right, the boundary separator is composed of the edges of so-called Voronoi tiles of adjacent examples that have different classes. This is a less clean-cut boundary that would take so much space to store (i. e., all the coordinates of its joints) that in fact algorithms that make use of this type of separator do not compute it globally and beforehand, but locally and only when needed, on the basis of the individually memorized examples. Algorithms that make use of this class separation method are called k-nearest neighbor classifiers (Cover/Hart 1967; Dasarathy 1991; Aha/Kibler/Albert 1991).

The two left types of decision boundaries depicted in Figure 39.2 represent a general theme in machine learning, namely to search for minimally-sized abstractions. Two axis-orthogonal lines or one oblique line both represent major reductions of the amount of information that is needed to represent the two class spaces, as compared to what is needed to store the label and coordinates of the six individual examples. The goal of finding minimally-sized abstractions has been formally defined in the minimal description length principle (Rissanen 1983). Analogous to the medieval Occam's razor principle, the minimal description length principle states that smaller abstracted problem solutions are better.

A solution's size is defined as the sum of the size of the abstracted model plus the size of the list of exceptions the model does not cover. Both are measurable in the number of bits it takes to store them. To attain the goal of minimal description, a machine learning algorithm needs to find a balance between on the one hand compressing too much, which amounts to *overgeneralization* or *underfitting*, and on the other hand staying too close to the training material and thereby possibly staying too far off from the true class boundaries, or *overfitting*.

3. Applying machine learning to annotated corpora

Currently it is safe to say that computational linguistics is largely corpus-based. Most current literature is based on the application of some stochastic or discrete classification or clustering method to annotated corpora. Typically, these corpora have been annotated at levels representing phonetic, morphological, syntactic, semantic, or pragmatic abstractions at various granularities. Taking the Western alphabets as reference, the typical levels of granularity follow the typical units of letters, words, sentences, paragraphs, texts, and unbounded discourse:

- Typically, **morphological** and **phonological** tasks are annotated at the character level, e. g. mapping roman letters to **phonemes**, or to segmentation markers denoting the boundaries of **syllables** (which can carry syllabic **stress**) or **morphemes** (e. g. compounds in morphologically complex words).
- At the word level, mainly **word-level syntax and (lexical) semantics** are annotated. Syntactical word-level annotations are typically **part-of-speech tags** (where typically the morphological complexity of the language determines the granularity of the tag set; see article 24), or **shallow parsing** (where each word is annotated with a tag denoting the membership of that word in a multi-word chunk of words sharing a syntactic function). Semantic word-level annotations concern **word senses** (see article 26), or the membership in a certain **named entity multi-word unit**. Also, **prosody** (intonation, melody, pauses and breaks, and sentence accents; see article 30) is typically annotated at the word level, at or between words. Another frequent word-level annotation is the one where for each word the **lemma** is stored, that is, some reference base form (e. g. infinitives for verbs, and singular forms for nouns; see article 25).
- At the sentence level, complete **syntactic analyses** are annotated that typically take the form of nested labeled constituent structures, or dependency graphs (see article 28).
- At the paragraph or text level, with “text” meaning written text here, typically **topic** or **category** labels are annotated; a text may have several topics or categories, overlap-

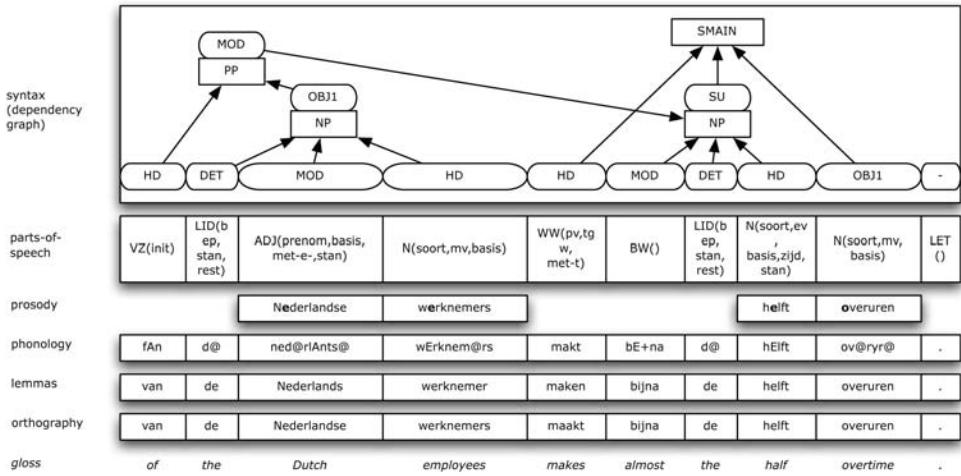


Fig. 39.3: Example of an orthographically transcribed utterance from the Spoken Dutch Corpus, annotated at the levels of phonology, lemmas, prosody, parts-of-speech, and syntax

ping or in a particular sequence with clear segmentation boundaries. Text may also have a **rhetorical structure** that can be annotated in various ways.

- Particularly in spoken language corpora, such as transcribed dialogues, where the linguistic utterances cannot be meaningfully interpreted without the context they were uttered in, annotations incorporate pragmatic phenomena such as **dialogue acts**, non-textual phenomena such as physical **gestures**, and bookkeeping features such as **speaker identification** and **timings**.

Figure 39.3 displays an excerpt from the Spoken Dutch Corpus (Oostdijk et al. 2002) featuring several parallel layers of annotation. The example is a transcribed utterance (which translates, freely, into “half of the Dutch employees work overtime”), where the words are the anchor positions to which other layers are aligned. Some layers, such as the phonemic representation and the parts-of-speech layer, have as many elements as the utterance has words, since each element maps to each word in the original utterance. Other layers, such as the prosodic layer, may only signal events at certain positions, such as words bearing sentence accents. Again other layers, such as the syntactic layer, may represent more complex structures, such as directed labeled graphs.

3.1. Feature spaces

When an annotated corpus is available, processing tasks can be formulated that map some input, for example a sentence, to one or more of the annotation layers. The supervised machine learning paradigm dictates that a task be described as a mapping of input to output, where the input is a vector of features, and the output is a label out of a set of possible labels; and where a single example is represented by a single input vector, instantiated with particular values, and mapping to a single output label. Input features can be numeric, binary, or symbolically multi-valued. Output labels, or classes, can be binary or symbolically multi-valued. (Like features, classes can be numeric as well; such

tasks can be learned by regression algorithms. We will not treat regression here.) These constraints on what a supervised machine learning algorithm can handle still leave everything open. It is up to the experimenter to define what is given to the learner as input, and it is up to the experimenter to ensure that this input contains all necessary information that could be relevant for making the mapping to the correct output class successfully. In other words, it is important that the right dimensions are selected to design a feature space in which the machine learning algorithm could draw its decision boundaries, hyperplanes, or Voronoi tiles.

What constitutes suitable feature spaces for natural language processing (or modeling) tasks is bounded by (1) the task at hand, and (2) fundamental constraints of natural language. Before turning to the tasks themselves, the three most general constraints imposed by natural language relevant to NLP are often thought to be the following (also, see articles 37 and 41 for further discussion of these issues):

- **The locality of disambiguating information.** Most tasks in NLP are complicated due to ambiguity (or choice) in the mapping from the one to the other level. Yet, much of this ambiguity can be solved quite locally. For instance, many ambiguities in part-of-speech tagging can be resolved with a context of one word to the left of the ambiguously tagged word. This general constraint has its obvious exceptions; take syntactic long-distance dependencies, for example. Yet, it has been argued that long-distance disambiguation can always be translated back to local disambiguation, provided that the right preprocessing has taken place (Daelemans 1999).
- **The Zipfian distribution of words.** (Also see article 37.) In any text a relatively small set of words tends to recur frequently, while a much larger set of words occurs less frequently or hardly at all (Zipf 1935). If examples with words as features are presented in their naturally occurring frequencies, a machine learner will quickly encounter many examples of the first kind of words (usually, function words), while on the other hand it will typically encounter about half of the words types only once. Also, when presented with new examples to be labeled, some of them are bound to contain words the learner has never seen before – which makes context all the more important for classification.
- **The burstiness of words.** Words in the low-frequency range of the Zipfian distribution are not uniformly distributed over texts, but rather come in bursts. Once a rare word is seen in a text, chances become high that it will re-occur soon. This phenomenon is best known from news texts (where names of unknown people or phenomena suddenly become household names), but can also be seen as related to hidden parameters that generate text, such as authorship, topic, and genre. The Dirichlet distribution is typically used to model word burstiness (Blei/Ng/Jordan 2003; Madsen/Kauchak/Elkan 2005).

The first constraint appears easy to meet; it limits the amount of features that need to be collected for a task to be learnable by a machine learning algorithm. Nevertheless, even with the locality constraint, many mixtures of features representing language at different granularities and abstraction levels are possible, still leaving a large feature selection search space. The other two constraints pose more direct problems to certain types of learning algorithms; we will return to them in section 5.

In the current state of the art in machine learning of natural language processing, designing feature sets for tasks to be learned is not an automatic process. Rather, it is mostly based on a combination of expert and heuristic knowledge. Many studies have

sentence accent placement

van	de	Nederlandse	werknamers	maakt	
VZ(init)	LID(b ep, stan, rest)	ADJ(prenom,basis, met-e-,stan)	N(soort,mv,basis)	WW(pv,tg w, met-t)	→ accented

dependency relation assignment

Nederlandse		werknamers	maakt	bijna	de	
ADJ(prenom,basis, met-e-,stan)		N(soort,mv,basis)	WW(pv,tg w, met-t)	BW()	LID(b ep, stan, rest)	→ right element has OB1 relation to left element
de	helft	overuren	.	maakt left of overuren	3 words between maakt and overuren	
LID(b ep, stan, rest)	N(soort,ev ,basis,zijd, stan)	N(soort,mv, basis)	LET ()			

Fig. 39.4: Example instances of two different tasks drawn from the Spoken Dutch Corpus: placing sentence accent on *Nederlandse* based on a local window of words and parts-of-speech (top), and establishing a dependency relation between *maakt* and *overuren* based on two local windows and information on the relative position of the two words (bottom). In both tasks, the features are on the left, and the class to be predicted is on the right hand side of the arrow

already confirmed the locality constraint, and many have suggested that it is useful to combine features from various levels of linguistic abstraction. Figure 39.4 illustrates two instances, created from handcrafted feature sets for two processing tasks, based on annotations in the Spoken Dutch Corpus. The one displayed on top represents one instance of a word (*Nederlandse*, Dutch), which, when the sentence is read aloud (e.g. by a speech synthesizer), should receive sentence accent (typically realized as a combination of a pitch rise and fall, a temporary lengthening of the vowel of the syllable with primary stress, and a temporary volume increase). The task of sentence accent placement is intricately bound to the meaning of the text being pronounced, and the individual importance of each word to the meaning of the text. As Marsi et al. (2003) show, the most important predictive features are the words and the parts-of-speech of the words in a local window of the focus word and its four neighboring words; two to the left and two to the right.

The bottom instance in Figure 39.4 exemplifies a somewhat more complicated classification. The example represents the classification of one particular dependency relation in a larger dependency graph (displayed at the top of Figure 39.3), namely the one between the words *maakt* (makes) and *overuren* (overtime). This labeled example shows how a non-local issue (relations can span entire sentences) can be made local, and also shows how the full task of dependency parsing can be decomposed into singular decisions that together make up the final solution. The exemplified dependency relation is pointing leftward, from the right element, the object *overuren*, to the left element, *maakt*.

The complex feature set is composed of two local windows, each surrounding one of the two words, containing the parts-of-speech and the identities of two words to the left and to the right of the focus words. Also, two non-positional features represent their distance (in the number of words, in this case 3) and their relative position (*maakt* is positioned left of *overuren*). This feature set was used in (Canisius et al. 2006) as part of a dependency parser, that further used an inference step to construct a full dependency parse from all individual pair-wise relationship assignments. This “learning plus inference” division of labor has been adopted widely in sentence-level tasks where the output is non-trivially structured (Punyakanok/Roth 2001; Lafferty/McCallum/Pereira 2001; Carreras et al. 2002).

A summary recipe for an average feature selection at the sentence level would be to select an appropriately wide window of words, and take the words themselves as features, along with one or two features that will still match when the word mismatches, such as the lemma or the part-of-speech of each word. Additionally, one might take into account more global features such as the numeric features representing word distance, and the binary feature representing relative position.

Certain machine learning algorithms have constraints, too, on the possible shape of the feature space. Most decision-tree and rule learners as well as the k-nearest neighbor classifier can naturally handle multi-valued features, such as “the word in the position immediately to the right of the focus word”, as well as multi-valued output labels. Most other machine learning algorithms such as the Naive Bayes classifier, the maximum-entropy classifier, artificial neural networks, and Support Vector Machines actually convert all multi-valued features into binary features. For instance, a multi-valued feature carrying all 26 letters of the alphabet as possible values, is converted into 26 binary features, each representing the presence of one letter. Support Vector Machines, in their basic definition, cannot handle more than two classes, so that tasks with more than two classes have to be decoded into strings of binary classifications, amounting to one classifier for each class label (“one versus all”), one classifier for each pair of class labels (“pairwise”). Or, alternatively, through *error correction output codes*, a method for coding the output class space into a string of bits the length of which lies between the minimal number to encode the different classes, and the “one versus all” encoding, such that decoding from the error-correcting output code to the actual class is to some extent noise-resistant (Dietterich/Bakiri 1991).

3.2. Annotated corpora for supervised learning

As the proceedings of conferences in the area of computational linguistics and NLP of recent years readily witness, the use of machine learning methods and related stochastic methods has become ubiquitous. Nevertheless, most of this work, and certainly the work in supervised learning which this article focuses on, is totally dependent on the availability of annotated corpora. The current amount of linguistically annotated corpora is actually quite small; some of it is in the public domain, some is available under specific licenses, and an unknown amount of annotated corpora is not available for general research purposes due to copyright issues or commercial reasons. Corpora that have had the largest impact on the development of NLP tools have been largely of the second type, released under limiting licenses. Arguably, the distribution of corpora in the litera-

ture follows the Pareto “80%–20%” rule: 80% of all work uses 20% of the available corpora. Within this small set of frequently-used corpora, the Penn Treebank corpus (Marcus/Santorini/Marcinkiewicz 1993; Marcus et al. 1994) accounts for a substantial percentage on its own. Like the above-mentioned Spoken Dutch Corpus, the Penn Treebank has grown to contain a number of annotation layers. Starting from tokenized versions of Wall Street Journal news articles, the ATIS corpus of human-machine dialogues, the classic balanced Brown corpus, and the SWITCHBOARD corpus of human-human dialogues, part-of-speech tags were added and constituent-based parsing trees were annotated on top of sentences (Marcus/Santorini/Marcinkiewicz 1993). The simple 48-tag part-of-speech tagset (36 syntactic tags and 12 punctuation tags) has effectively become a standard in English NLP. Later on, semantic functional tags (such as “subject”, “direct object”, “temporal phrase”) were added to the treebank (Marcus et al. 1994). Subsequently, the corpus has been used as a basis for additional annotations. For example, the SWITCHBOARD corpus has been annotated on the level of dysfluencies; the PropBank corpus (Kingsbury/Palmer/Marcus 2002) has added verb-argument patterns, to mark basic semantic propositions in the treebank.

The Penn Treebank is released by the Linguistic Data Consortium (<http://www.ldc.upenn.edu/>), which also holds the rights to license many other annotated (and unannotated) corpora. The European Language Resources Association (ELRA, <http://www.elra.info/>) offers a similar service. Advantages of having organized access to linguistically annotated corpora is that institutions such as LDC and ELRA have developed quality management guidelines, and are able to offer maintenance and archival services that would be hard to uphold by the university research groups that created many of the corpora.

4. Machine learning for corpus linguistics

Linguistic annotation of a corpus of written or spoken language is a notoriously costly process. Assuming that a protocol for annotation has been established, one obvious possibility is to use automatic means to pre-annotate a corpus, and have human annotators check and correct the predictions of the automatic module. At least the following three non-trivial conditions have to be met in order to make this setup effective and efficient:

1. The automatic annotation has to be correct beyond a certain level of accuracy. If many of the predictions are inaccurate, the human annotator will need to spend time on inspecting all data, and on correcting the many errors. There is a break-even point where annotating from scratch would take as much time as correcting partially incorrect pre-annotations; below the break-even point there is no point in using pre-annotations. The actual accuracy threshold depends on the task. In the context of part-of-speech tagging, even manually inspecting a seemingly high 90% correctness in tags has been reported to cost as much time as annotation from scratch, as one error in every ten words roughly amounts to at least one error per sentence. Only accuracies above 95% have clearly saved substantial time (van den Bosch/Schuurman/Vandeghinste 2006).
2. The automatic annotator has to be able to show its confidence in individual predictions, and this confidence needs to be correlated significantly to prediction errors.

Otherwise, the annotation task becomes the task of spotting the hidden errors, and this can take as much time, if not more than plain annotation from scratch, certainly if the amount of errors is large (see first point).

3. If machine learning methods are used (instead of manually-written knowledge-based pre-annotators), a *seed* corpus is needed with manual annotations, in order to train the initial pre-annotator. Given the first condition, this seed corpus needs to be large enough to allow an above-threshold accuracy of the initial pre-annotator. This means that manual annotation from scratch is unavoidable. The minimal size of the seed corpus depends on the annotation task; in practice, seed corpora typically contain in the order of some thousands of elementary annotations.

Besides these basic conditions, it is desirable that the automatic pre-annotator be able to incorporate the corrections made to its output during manual correction, and that the annotators are aided as much as possible with intuitive (e.g. graphical) interfaces with the proper visual means to show the data itself, the pre-annotations, the pre-annotator's confidence levels, and means to select correction candidates quickly. It is also important to instruct annotators to remain critical of the predictions of a machine learner even if it is correct most of the time, and its confidence levels often indicate errors; there is the danger that the annotator, in case of doubt, is biased to believing that the system is right.

As a more silent precondition to the above, the seed corpus mentioned in the third condition above will need to be sufficiently reliably annotated. What is "sufficient" rather depends on the demands further along the way, and some types of tasks are simply hard to annotate or agree on. As a rule of thumb it can be stated that predictions of a machine learner will typically not be better than the quality of the annotations of the training data. When annotation quality is low, the predictions of the machine learning algorithm will likely be of the same quality.

There is no a-priori reason to select machine-learning methods over knowledge-based methods, but the following reasons may sway the corpus developer to using the former:

- Machine-learning methods may be retrained automatically, e.g. on the seed material plus any new manually corrected samples. Some algorithms (such as the k-nearest neighbor classifier) even allow immediate incremental learning on an existing model. Many other algorithms have to retrain from scratch, and it may be computationally expensive to retrain them often, while this may be desired in the annotation process.
- Although a seed corpus may be hard to come by, there are methods to reuse pre-existing automatic NLP modules that perform not the particular annotation task at hand, but a related task, such as part-of-speech tagging with a different tagset (van Halteren/Zavrel/Daelemans 2001); this way an initial pre-annotator trained on a small corpus may be quite accurate due to educated guesses it receives from the external modules.
- Most machine learning algorithms have numeric ways of expressing their confidence about individual predictions. Often when such algorithms make errors, they do so with low confidence; correct predictions tend to be correlated with high confidence. Nevertheless, confidence estimations remain quite unreliable; so-called *calibration* methods are nowadays developed to learn the reliability of classifier confidence (Caruana/Niculescu-Mizil 2004).

In this section we offer three views on corpus annotation partly automated by machine learning. First, we exemplify the above standard scenario with a case study. Second, we present active learning, which uses human annotation as part of a loop to boost the performance of NLP, producing annotations as a by-product. Third, we pay attention to the use of machine learning in post-hoc correction of annotation layers in corpora that have already been annotated.

4.1. Automatic pre-annotation: The Spoken Dutch Corpus case

In the Spoken Dutch Corpus, machine learning was used intensively to assist in the automatic pre-annotation of part-of-speech tagging and lemmatization. We present this case as an example. At the start of the annotation project, four annotators jointly annotated a seed corpus of about ten thousand words of orthographically transcribed Dutch spoken utterances with a newly designed tagset (van Eynde 2004). All subsequent annotations in the remainder of the five-year project were done manually on the output of an ensemble of taggers trained on different tagsets and corpora, including the seed corpus. Taggers in the ensemble each employed different machine learning algorithms, with a meta-tagger on top that learned to integrate the different tagger outputs (van Halteren/Zavrel/Daelemans 2001). A subset of the combined sub-taggers was trained (and re-trained at regular intervals) on the growing Spoken Dutch Corpus itself, while the other taggers were static existing taggers for written Dutch (Zavrel/Daelemans 2000) using different tag sets. Performance (tagging accuracy) on random 10% held-out test data taken from the new corpus grew from an initial 94.2% to 97.1% at the last training. Figure 39.5 visualizes the learning curve of the meta-tagger along with those of the retrained sub-taggers participating in the combination. Two of the sub-taggers, the Brill tagger (Brill 1995) and a maximum-entropy-based tagger, MXPOST (Ratnaparkhi 1996), were not used throughout the entire period since it became too time-intensive to retrain them. The other two sub-taggers, a memory-based tagger (Daelemans et al. 1996) and a hidden-markov-based approach, Trigrams 'n' Tags (Brants 2000), were employed and retrained until the end of the project.

As can be seen in Figure 39.5, meta-learning yielded a major performance boost over a set of heterogeneous taggers in the early stages of the project, when between 10,000 and 100,000 tagged words were available for training. While some of the sub-taggers still performed at under 90% tagging accuracy on unseen data when 100,000 manually corrected tagged words were available, the meta-learning tagger had a fairly stable performance above 94% even at the early stage when only the seed corpus of ten thousand training words was available. Later in training, meta-learning continued to perform better than the best sub-taggers, but clearly the scores of the taggers almost converged.

In the correction process we made use of the fact that the tagger ensemble was able to output confidence values for each individual tag prediction. More specifically, each prediction consists of a set of predicted tags, ordered by their likelihood according to the tagger; these likelihood values were taken as confidences. A tool was developed that presents the human annotator only with those cases in which one or more possible tags had a below-threshold confidence. All other cases were not presented to the annotator, resulting in a lower workload.

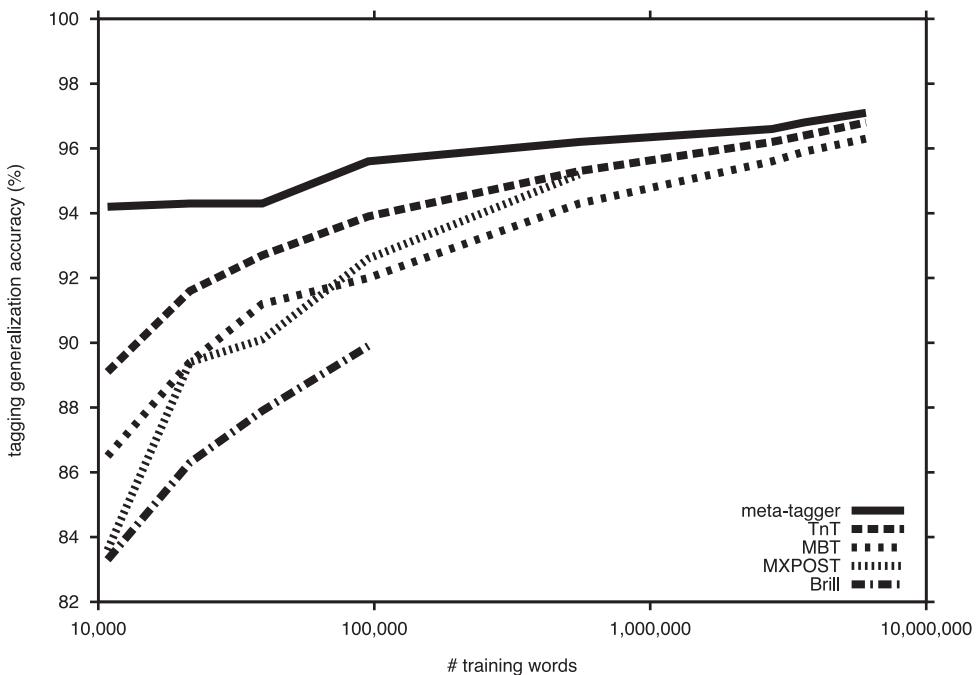


Fig. 39.5: Learning curves of the CGN meta tagger (top solid line) and its component sub-taggers (bottom dashed lines) with increasing amounts of training material. The x-axis has a logarithmic scale. The y-axis represents the percentage of correctly tagged test words

An experiment was carried out to determine the threshold such that most (if not all) true errors were presented to the user, while a maximal amount of assumedly correct data was not presented. As Figure 39.6 shows, a threshold confidence value of 0.06 resulted in a reduction of the number of decisions to be made by the human annotator by 28 %, while skipping a mere 1 % of errors. This shows that with a well trained tagger that is continuously retrained, increasingly larger amounts of data can be checked at the same time, missing hardly any errors.

Yet, using this method means that a fraction of the real errors are left unnoticed. In this particular case we used two additional knowledge-based post-processing methods to detect some remaining errors:

- **Checking against a blacklist.** All manually corrected material is regularly matched with a blacklist of typical errors made by the tagger, particularly on multi-word named entities (the tagger uses different tags for single-word proper nouns and multi-word named entities, so many words can be tagged as both), and high-frequency ambiguous function words such as *dat* (*that*, having the same ambiguity as in English) which the tagger sometimes tags incorrectly yet with high confidence.
- **Feeding back errors from shallow parsing modules.** Tagging errors are known to cause further errors in automatic shallow and full parsing. Applying a phrase chunker, for example, and correcting that, typically reveals PoS-tagging errors, which can be fed back as bug reports for manual correction.

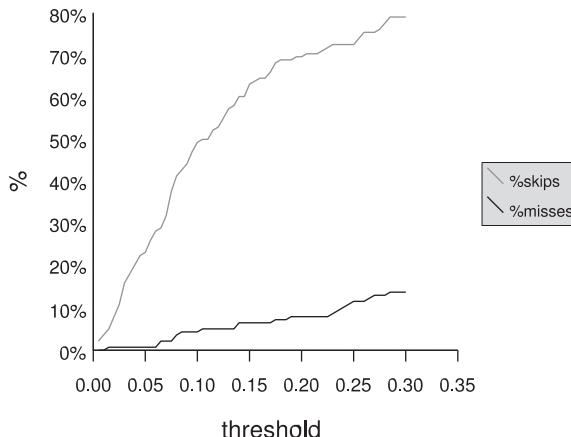


Fig. 39.6: The effect of thresholding the cases presented to the human annotator for correction on the percentage of cases skipped, and the percentage of PoS-tagging errors missed

4.2. Active learning

The goal of annotating a corpus may simply be to annotate it from beginning to end, at the highest possible quality level. Yet, when the purpose of annotation is to obtain the best possible training material for a machine-learning-based processing module performing some NLP task, annotating all words and sentences may not be necessary for the optimal result. Rather, there may be an optimal selection of a subset of the full corpus, a particular set of examples of the task at hand, for the best machine learning result. In a nutshell, the goal of active learning (Thompson/Caliiff/Mooney 1999) is to find such a subset of examples.

Active learning assumes the availability of a possibly small seed corpus that is manually annotated (corpus A), and a larger unannotated corpus (corpus B); the latter type of corpus is typically easy to obtain. A typical active learner works along the following procedure:

- A set of different machine learners is trained on the annotated seed corpus A. The amount of different learners in the set is typically three or more.
- A loop is entered which ends when no more examples are added to A, or when time is up:
 - All machine learners automatically tag corpus B;
 - Annotations are collected on which the learners disagree most. For example, with three classifiers maximal disagreement is reached when all classifiers predict a different outcome. In general, disagreement can best be quantified by computing the entropy of the outcomes, $H(O) = -\sum P(o) \log_2 P(o)$, where O is the set of outcomes, and o is an individual outcome. The selection involves some thresholding, e.g. the selection of the top 100 examples with the highest disagreement.
 - These examples with disagreement are presented to a human annotator, who decides on their actual labeling.
 - After annotation these examples are moved from B to A.
 - The machine learners are retrained on the expanded A corpus.

The end result is a corpus of examples selected from all around the larger unannotated corpus. If the task is part-of-speech tagging, the selected examples will tend to represent ambiguous words with a tricky tag in their sentential context, but without any of the sentences surrounding them originally.

Active learning is reported to yield sharper increasing learning curves than just training on increasing amounts of randomly annotated data (Thompson/Califf/Mooney 1999), meaning that indeed the examples selected during active learning are “hard cases” that need to be incorporated with more urgency in a successful processor’s training data set than the bulk of “easy cases” that abound in a random sample.

Related to active learning are methods that fully automatically attempt to select new examples from unannotated corpora, but lack the goal of beginning-to-end annotation, and any human intervention: expectation-maximization (Dempster/Laird/Rubin 1977), co-training (Blum/Mitchell 1998), and single-classifier boosting (Yarowsky 1995); for an overview, cf. Abney (2002).

4.3. Automatic annotation enhancement

Returning to the goal of producing an annotated corpus, it is also possible to use machine learning for improving already existing annotations in a corpus. Analogous to detecting disagreement between classifiers in active learning, it is informative to contrast the predictions of a classifier with the actual annotations that are already there – a mismatch may be the classifier’s fault, but it might also signify an error in the original annotation. Even if mismatches are mostly the classifier’s fault, checking the mismatches may reveal a substantial number of annotation errors.

One possible scheme is to train a classifier on the full corpus, and then re-classify the corpus with the trained classifier. This method acts as a consistency check; if performed with a k-nearest neighbor classifier, which retains perfect memory, any error must result from two or more identical examples being annotated with different labels, which in turn may point at an annotation error (or may not; the classifier may not be sensitive enough, or ambiguity in the annotations is allowed). More generally, an n -fold cross-validation experiment may be applied to the training corpus (Weiss/Kulikowski 1991). Set at the typical value of $n = 10$, a 10-fold cross-validation experiment cuts the corpus in ten disjoint portions of equal size, and runs ten experiments in which each one of the partitions acts as a test set, and a concatenation of the other nine partitions constitutes the training set. All errors made on the ten test partitions may be due to annotation errors, but with a smaller n it becomes more likely that the classifier makes the errors.

In a similar vein, machine learning may point at weak points in the annotation beyond the local level, for instance at the level of the tagset as a whole. By globally changing certain labels in the annotation layer and performing a full machine learning experiment, quantitative measurements can indicate the effect of the change on learnability, and on success in processing. Label changes should typically not be destructive, i. e. throw away information that distinguishes between two linguistic labels in a meaningful way. On the other hand, splitting certain labels into two or more different labels, when the single label appears “overloaded”, i. e., seems to label more than one linguistic function, may improve learnability and success; Ule/Veenstra (2003) demonstrate this on a German PCFG parser.

Enriching labels with more information from others that are structurally close, such as parent node labels of non-terminal nodes in a syntactic tree, has also been shown to improve parsing performance and reduce parsing complexity (Johnson 1998; Klein/Manning 2003).

5. Summary and conclusion

The relation between annotated corpora and machine learning appears a fruitful one; machine learning algorithms can learn to annotate language data as soon as a reasonable amount of example annotations are available. This way natural language processing modules can be generated automatically. The payoff for the effort invested in annotation is that machine learning can assist annotation efforts, by taking over a major part of the annotation job, and indicating on which parts of the annotation it is less confident.

Still, there are at least two mismatches between how language data behaves, and what machine learning assumes. The first is that many machine learning algorithms make the (sometimes tacit) assumption that features or feature values are distributed normally, while in language many elements have rather deviating distributions. We expand briefly on this issue in section 5.1.

The second mismatch is rooted in the assumption that training and test data are drawn out of the same source, in statistical and probability theory referred to as the “i. i. d.” (independent and identically distributed) assumption – that training and test data are independent, and at the same time identically distributed. We briefly overview why the latter is easily not the case in language data in the closing section 5.2.

5.1. There is no data like more data

Several studies in the computational linguistics literature that investigate the effect of the amount of training data report the seemingly odd phenomenon of the log-linear learning curve. As exemplified by Banko/Brill (2001) and van den Bosch/Buchholz (2002), when confronted with a doubling of the amount of training material of some natural language processing task, the performance of the machine learning algorithm trained on this doubled amount of training material often appears to increase by a constant factor, no matter the actual size of the training set. One would expect that learning curves tend to flatten even in logarithmic space until a ceiling is reached that is intrinsic to the problem. In contrast, performance on language tasks appears to simply increase constantly with every n -fold increase, until some practical limitation, such as computer memory or the size of the annotated corpus, brings an untimely end to the curve.

As one explanation, Banko/Brill (2001) point to the fact that words occur in a Zipfian distribution, or put otherwise, that the relation between their frequency and rank follows a power law. A small set of words occurs very frequently, while an endless tail of words occurs rarely. Expanding a corpus means that the tail of the Zipfian curve is charted further, but no increase will ever reach the tail’s end. A doubling of corpus size has two effects: first, many words that have been seen before, are seen again, so their observation counts improve; second, many new words are encountered for the first time. Both effects

are relatively constant with every n-fold increase of the amount of training data, and this constancy can be taken as the basis for explaining that a machine learner should be able to profit from this by improving by a constant factor too. Reasoning further, Banko/Brill (2001) come close to proclaiming the end of computational linguistics, since all that needs to be done from this viewpoint is to annotate more corpora. Although this reasoning is sound, the field has not given up yet on finding shortcuts that will push the learning curve up faster, assuming the kinds and sizes of annotated corpora we now know as realistically possible to develop.

5.2. Independent but rarely identically-distributed

Machine-learning-based NLP systems often exhibit dramatic losses in performance when the data they are applied to deviate from the type of data they were trained on. As an example, Carreras/Màrquez (2005) note an average drop of about 10 points on a scale of 1 to 100, measuring the F-score on correctly identified verb–argument relations predicted by 19 semantic role labeling systems participating in the CoNLL-2005 shared task, when test scores were computed on deviating data. All systems were trained on Penn Treebank material from the Wall Street Journal section. The best scores (up to 79 F-score points) were obtained on test data from the Wall Street Journal, while no system was able to attain an F-score above 67 F-score points on test data from the Brown corpus section of the same treebank.

It is obvious that the distribution of words in the Brown corpus (containing samples of different genres of written English) deviates from that of the Wall Street Journal corpus. More generally, any natural language processing system that is put to generic use will often encounter data of which many of the content words are unknown to the system since the text covers an unseen topic, or is of an unseen type or genre. Considering all the dimensions in corpus linguistics that are known to be relevant: style, genre, register, author, and domain, this should not come as a surprise.

NLP work at the text level, in text categorization and topic segmentation, has already acknowledged this problem (Blei/Ng/Jordan 2003; Madsen/Kauchak/Elkan 2005), and has approached it, with some success, by explicitly modeling the mixture of registers or language models (one per topic or category) from which texts or parts of texts are generated. The mixtures are modeled by Dirichlet distributions. Similar traces of a generic approach to other tasks in NLP can be found for example in the recent work by Daumé/Marcu (2006). Starting from an in-between solution, the one-time adaptation of generic NLP tools to domains, Daumé/Marcu argue in line with Blei/Ng/Jordan (2003) that a proper generic solution lies in mixtures of models. There is a clear need for approaches like this, which will hopefully change what it means for a machine-learning-based NLP system to be generally applicable.

6. Literature

- Abney, S. (2002), Bootstrapping. In: *Proceedings of the Annual Meeting of the ACL*. Philadelphia, PA, 360–367.
Aha, D. W./Kibler, D./Albert, M. (1991), Instance-based Learning Algorithms. In: *Machine Learning* 6, 37–66.

- Banko, M./Brill, E. (2001), Scaling to Very Very Large Corpora for Natural Language Disambiguation. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France. Morristown, NJ: Association for Computational Linguistics, 26–33.
- Blei, D./Ng, A./Jordan, M. (2003), Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* 3, 993–1002.
- Blum, A./Mitchell, T. (1998), Combining Labeled and Unlabeled Data with Co-training. In: *Proceedings of the 11th Annual Conference on Computational Learning Theory*. Madison, WI, 92–100.
- Brants, T. (2000), TnT – a Statistical Part-of-speech Tagger. In: *Proceedings of the 6th Applied NLP Conference, ANLP-2000, April 29–May 3, 2000*. Seattle, WA, 224–231.
- Breiman, L./Friedman, J./Ohlsen, R./Stone, C. (1984), *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.
- Brill, E. (1995), Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. In: *Computational Linguistics* 21(4), 543–565.
- Canisius, S./Bogers, T./van den Bosch, A./Geertzen, J./Tjong Kim Sang, E. (2006), Dependency Parsing by Inference over High-recall Dependency Predictions. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*. New York, NY, 176–180.
- Carreras, X./Màrquez, L./Punyakanok, V./Roth, D. (2002), Learning and Inference for Clause Identification. In: *Proceedings of the European Conference on Machine Learning*. London: Springer, 35–47.
- Carreras, X./Màrquez, L. (2005), Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ann Arbor, MI, 152–164.
- Caruana, R./Niculescu-Mizil, A. (2004), Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. In: *Proceedings of the Tenth ACM SIGKDD Conference*, Seattle, WA. New York: ACM Press, 69–78.
- Clark, P./Niblett, T. (1989), The CN2 Rule Induction Algorithm. In: *Machine Learning* 3, 261–284.
- Cohen, W. (1995), Fast Effective Rule Induction. In: *Proceedings of the 12th International Conference on Machine Learning*. Tahoe City, CA: Morgan Kaufmann, 115–123.
- Cortes, C./Vapnik, V. (1995), Support Vector Networks. In: *Machine Learning* 20, 273–297.
- Cover, T. M./Hart, P. E. (1967), Nearest Neighbor Pattern Classification. In: *Institute of Electrical and Electronics Engineers Transactions on Information Theory* 13, 21–27.
- Daelemans, W. (1999), Memory-based Language Processing. In: *Journal of Experimental and Theoretical Artificial Intelligence* 11(3), 287–296.
- Daelemans, W./Zavrel, J./Berck, P./Gillis, S. (1996), MBT: A Memory-based Part of Speech Tagger Generator. In: Ejerhed, E./Dagan, I. (eds.), *Proceedings of the Fourth Workshop on Very Large Corpora*. Copenhagen, Denmark, 14–27.
- Dasarathy, B. V. (1991), *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.
- Daumé III, H./Marcu, D. (2006), Domain Adaptation for Statistical Classifiers. In: *Journal of Artificial Intelligence Research* 26, 101–126.
- Dempster, A. P./Laird, N. M./Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. In: *Journal of the Royal Statistical Society, Series B (Methodological)* 39(1), 1–38.
- Dietterich, T. G./Bakiri, G. (1991), Error-correcting Output Codes: A General Method for Improving Multiclass Inductive Learning Programs. In: *Proceedings of AAAI-91*. Menlo Park, CA, 572–577.
- Fürnkranz, J./Widmer, G. (1994), Incremental Reduced Error Pruning. In: Cohen, W./Hirsch, H. (eds.), *Proceedings of the 11th International Conference on Machine Learning*. New Brunswick, NJ: Morgan Kaufmann, 70–77.
- Johnson, M. (1998), PCFG Models of Linguistic Tree Representations. In: *Computational Linguistics* 24(4), 613–632.

- Kingsbury, P./Palmer, M./Marcus, M. (2002), Adding Semantic Annotation to the Penn Treebank. In: *Proceedings of the Human Language Technology Conference*, San Diego, CA. Available at: http://faculty.washington.edu/fxia/courses/LING571/PropBank_HLT2002.pdf.
- Klein, D./Manning, C. (2003), Accurate Unlexicalized Parsing. In: *Proceedings of ACL-2003*. Sapporo, Japan, 423–430.
- Lafferty, J./McCallum, A./Pereira, F. (2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the 18th International Conference on Machine Learning*. Williamstown, MA, 282–289.
- Littlestone, N. (1988), Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. In: *Machine Learning* 2, 285–318.
- Madsen, R. E./Kauchak, D./Elkan, C. (2005), Modeling Word Burstiness Using the Dirichlet Distribution. In: *Proceedings of the 22nd International Conference on Machine Learning*. New York: ACM Press, 545–552.
- Marcus, M./Kim, M./MacIntyre, R./Bies, A./Ferguson, M./Katz, K./Schasberger, B. (1994), The Penn Treebank: Annotating Predicate Argument Structure. In: *Proceedings of ARPA Human Technology Workshop*. Plainsboro, NJ, 110–115.
- Marcus, M./Santorini, B./Marcinkiewicz, M. (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19(2), 313–330.
- Marsi, E./Reynaert, M./van den Bosch, A./Daelemans, W./Hoste, V. (2003), Learning to Predict Pitch Accents and Prosodic Boundaries in Dutch. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. New Brunswick, NJ, 489–496.
- Mitchell, T. (1997), *Machine Learning*. New York: McGraw-Hill.
- Oostdijk, N./Goedertier, W./Van Eynde, F./Boves, L./Martens, J. P./Moortgat, M./Baayen, H. (2002), Experiences from the Spoken Dutch Corpus Project. In: González Rodríguez, M./Paz Suárez Araujo, C. (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain, 340–347.
- Punyananok, V./Roth, D. (2001), The Use of Classifiers in Sequential Inference. In: *NIPS-13; The 2000 Conference on Advances in Neural Information Processing Systems*. Denver, CO: The MIT Press, 995–1001.
- Quinlan, J. R. (1986), Induction of Decision Trees. In: *Machine Learning* 1, 81–206.
- Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- Ratnaparkhi, A. (1996), A Maximum Entropy Part-of-speech Tagger. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, May 17–18, 1996, University of Pennsylvania*. Philadelphia, PA, 133–142.
- Ratnaparkhi, A./Reynar, J./Roukos, S. (1994), A Maximum Entropy Model for Prepositional Phrase Attachment. In: *Workshop on Human Language Technology*. Plainsboro, NJ, 250–255.
- Rissanen, J. (1983), A Universal Prior for Integers and Estimation by Minimum Description Length. In: *Annals of Statistics* 11, 416–431.
- Rosenblatt, F. (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. In: *Psychological Review* 65, 368–408.
- Rumelhart, D. E./Hinton, G. E./Williams, R. J. (1986), Learning Internal Representations by Error Propagation. In: Rumelhart, D. E./McClelland, J. L. (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: *Foundations*. Cambridge, MA: The MIT Press, 318–362.
- Shavlik, J. W./Dietterich, T. G. (eds.) (1990), *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- Thompson, C. A./Califf, M. E./Mooney, R. J. (1999), Active Learning for Natural Language Parsing and Information Extraction. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, 406–414.
- Ule, T./Veenstra, J. (2003), Iterative Treebank Refinement. In: *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*, Antwerp, Belgium. University of Antwerp. Available at: <http://www.cnts.ua.ac.be/clin2003/proc/11Ule.pdf>.

- van den Bosch, A./Buchholz, S. (2002), Shallow Parsing on the Basis of Words Only: A Case Study. In: *Proceedings of the 40th Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 433–440.
- van den Bosch, A./Schuurman, I./Vandeghinste, V. (2006), Transferring Pos-tagging and Lemmatization Tools from Spoken to Written Dutch Corpus Development. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy. Available at: http://www.ccl.kuleuven.be/~vincent/ccl/papers/LREC06_pos.pdf.
- van Eynde, F. (2004), *Part of speech tagging en lemmatisering: Protocol for the Annotators in the Spoken Dutch Corpus*. Technical report, Leuven University. Available at: http://www.ccl.kuleuven.be/Papers/POSmanual_febr2004.pdf.
- van Halteren, H./Zavrel, J./Daelemans, W. (2001), Improving Accuracy in Word Class Tagging through Combination of Machine Learning Systems. In: *Computational Linguistics* 27(2), 199–230.
- Weiss, S./Kulikowski, C. (1991), *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann.
- Yarowsky, D. (1995), Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: *Proceedings of ACL-95*. Cambridge, MA: 189–196.
- Zavrel, J./Daelemans, W. (2000), Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, 17–20.
- Zipf, G. K. (1935), *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Second paperback edition, 1968. Cambridge, MA: The MIT Press.

Antal van den Bosch, Tilburg (The Netherlands)

40. Exploratory multivariate analysis

1. Introduction
2. Exploratory multivariate analysis
3. Data
4. Exploratory multivariate methods
5. Exploratory multivariate analysis in corpus linguistics
6. Literature

1. Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional paper-based methods for search and interpretation of data have been overwhelmed by sheer volume, and a variety of computational methods have been developed in an attempt to make the deluge tractable. As such methods have been refined and new ones introduced, something over and above tractability has emerged – new and unexpected ways of understanding the data. The fact that a computer can deal with vastly larger datasets than a human is an obvious factor, but there are two others of at least equal importance. One is the ease with which data can be manipulated and reanalyzed in interesting ways with-

- van den Bosch, A./Buchholz, S. (2002), Shallow Parsing on the Basis of Words Only: A Case Study. In: *Proceedings of the 40th Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 433–440.
- van den Bosch, A./Schuurman, I./Vandeghinste, V. (2006), Transferring Pos-tagging and Lemmatization Tools from Spoken to Written Dutch Corpus Development. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy. Available at: http://www.ccl.kuleuven.be/~vincent/ccl/papers/LREC06_pos.pdf.
- van Eynde, F. (2004), *Part of speech tagging en lemmatisering: Protocol for the Annotators in the Spoken Dutch Corpus*. Technical report, Leuven University. Available at: http://www.ccl.kuleuven.be/Papers/POSmanual_febr2004.pdf.
- van Halteren, H./Zavrel, J./Daelemans, W. (2001), Improving Accuracy in Word Class Tagging through Combination of Machine Learning Systems. In: *Computational Linguistics* 27(2), 199–230.
- Weiss, S./Kulikowski, C. (1991), *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann.
- Yarowsky, D. (1995), Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: *Proceedings of ACL-95*. Cambridge, MA: 189–196.
- Zavrel, J./Daelemans, W. (2000), Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, 17–20.
- Zipf, G. K. (1935), *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Second paperback edition, 1968. Cambridge, MA: The MIT Press.

Antal van den Bosch, Tilburg (The Netherlands)

40. Exploratory multivariate analysis

1. Introduction
2. Exploratory multivariate analysis
3. Data
4. Exploratory multivariate methods
5. Exploratory multivariate analysis in corpus linguistics
6. Literature

1. Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional paper-based methods for search and interpretation of data have been overwhelmed by sheer volume, and a variety of computational methods have been developed in an attempt to make the deluge tractable. As such methods have been refined and new ones introduced, something over and above tractability has emerged – new and unexpected ways of understanding the data. The fact that a computer can deal with vastly larger datasets than a human is an obvious factor, but there are two others of at least equal importance. One is the ease with which data can be manipulated and reanalyzed in interesting ways with-

out the often prohibitive labour that this would involve using manual techniques, and the other is the extensive scope for visualization that computer graphics provide.

These developments have clear implications for corpus linguistics. On the one hand, large electronic corpora potentially exploitable by the linguist are being generated as a by-product of the many kinds of daily IT-based activity worldwide, and, on the other, more and more application-specific electronic linguistic corpora are being constructed. Effective analysis of such corpora will increasingly be tractable only by adapting the interpretative methods developed by the statistical, computational linguistics, information retrieval, data mining, and related communities.

The present article deals with one type of analytical tool: exploratory multivariate analysis. The discussion is in five main parts followed by a select bibliography. The first part is the present introduction, the second explains what is meant by exploratory multivariate analysis, the third discusses the characteristics of data and the implications of these characteristics for generation and interpretation of analytical results, the fourth gives an overview of the various exploratory analytical methods currently available, and the fifth reviews the application of exploratory multivariate analysis in corpus linguistics. The material is presented in an intuitively accessible way, avoiding formalisms as much as possible. However, in order to work with multivariate analytical methods some background in mathematics and statistics is indispensable.

2. Exploratory multivariate analysis

Observation of nature plays a fundamental role in science. In current scientific method, a hypothesis about some natural phenomenon is proposed and its adequacy assessed using data obtained from observation of the domain of inquiry. But nature is dauntingly complex, and there is no practical or indeed theoretical hope of being able to observe even a small part of it exhaustively. Instead, the researcher selects particular aspects of the domain for observation. Each selected aspect is represented by a variable, and a series of observations is conducted in which, at each observation, the values for each variable are recorded. A body of data is thereby built up on the basis of which a hypothesis can be assessed. One might choose to observe only one aspect – the height of individuals in a population, say – in which case the data consists of more or less numerous values assigned to one variable; such data is univariate. If two values are observed – say height and weight – then the data is bivariate, if three trivariate, and so on up to some arbitrary number n ; any data where n is greater than 1 is multivariate.

As the number of variables grows, so does the difficulty of understanding the data, that is, of conceptualizing the interrelationships of variables within a single data item on the one hand, and the interrelationships of complete data items on the other. Multivariate analysis is the computational use of mathematical and statistical tools for understanding these interrelationships in data.

Numerous techniques for multivariate analysis exist. They can be divided into two main categories which are often referred to as ‘exploratory’ and ‘confirmatory’. Exploratory analysis aims to discover regularities in data which can serve as the basis for the formulation of hypotheses about the domain of interest. Such techniques emphasize intuitively accessible, usually graphical representations of data structure. Confirmatory multivariate analysis attempts to determine whether or not there are significant relation-

ships between some number of selected independent variables and one or more dependent ones. These two types are complementary in that the first generates hypotheses about data, and the second tries to determine whether or not such hypotheses are valid. Exploratory analysis is naturally prior to confirmatory; this article focuses on the former.

On multivariate analysis in general, see for example Everitt/Dunn (2001); Gordon (1999); Grimm/Yarnold (1995, 2000); Hair et al. (2005); Kachigan (1991); Tinsley/Brown (2000); Tabachnick/Fidell (2006).

3. Data

Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data — measurements of such things as air temperature are. A text corpus is not the linguist's data — measurements of such things as average sentence length are. Data is constructed from observation of things in the world, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the analytical results. The importance to exploratory multivariate analysis of understanding such data issues can hardly be overstated. On the one hand, "however powerful the exploring tools, or aggressive the explorer, nothing can be discovered that is beyond the limits of the data itself" (Pyle 1999, 46). On the other, failure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. For these reasons, an account of data issues is given before moving on to exploratory multivariate methods.

3.1. Variable selection

Given that data is an interpretation of some aspect of the world, what does such an interpretation look like? It is a description of the selected aspect in terms of variables. A variable is a symbol, and as such is a physical entity with a conventional semantics, where a conventional semantics is understood as one in which the designation of a physical thing as a symbol together with the connection between the symbol and what it represents are determined by agreement within a community. The symbol 'A', for example, represents the phoneme /a/ by common assent, not because there is any necessary connection between it and what it represents. Since each variable has a conventional semantics, the set of variables chosen to describe a domain of inquiry constitutes the template in terms of which the domain is interpreted. Selection of appropriate variables is, therefore, crucial to the success of any data analysis.

Which variables are appropriate in each given case? That depends on the nature of the research. Data can only be created in relation to a research question that provides an interpretative orientation in the domain of interest. Without such an orientation, how does one know what to observe, what is important, and what is not? The fundamental principle in variable selection is that the variables must describe all and only those aspects of the domain that are relevant to the research question. In general, this is an unattainable ideal. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of

personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an optimally relevant set of variables for a research question.

3.2. Data representation

If they are to be analyzed using mathematical methods, the selected variables need to be mathematically represented. A widely used way of doing this is vector space representation (Belew 2000, 86–7; Lebart/Rajman 2000; Manning/Schütze 1999, 539–44; Pyle 1999, 202–22; Salton/Wong/Yang 1975; Salton/McGill 1983, ch. 4). A vector is a sequence of scalars indexed by the positive integers 1, 2, … n, where a scalar is a single number:

$$\mathbf{v} = \begin{bmatrix} 1.6 & 2.4 & 7.5 & 0.6 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

Fig. 40.1: A vector

A vector space is a geometrical interpretation of a vector in which

- (i) the dimensionality of the vector, that is, its index length n, defines an n-dimensional space. There are various possible types of space, but for present purposes space is taken to be the Euclidean one familiar from elementary geometry, in which the axes are straight lines at right angles to one another.
- (ii) the sequence of scalars comprising the vector specifies coordinates in the space. These coordinates are relative to the scales of the axes.
- (iii) the vector itself is a point at the specified coordinates in the space.

For example, the two components of a vector $v = [36 160]$ are the coordinates of a point in a 2-dimensional space, and those of $v = [36 160 71]$ of a point in 3-dimensional space:

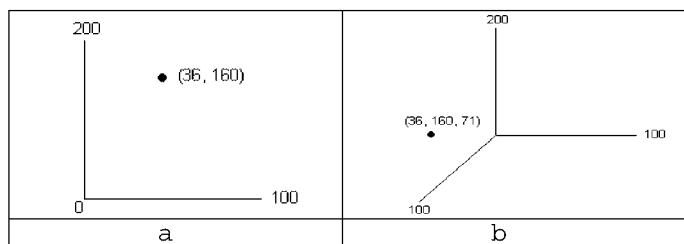


Fig. 40.2: 2 and 3-dimensional vector spaces

A length-4 vector defines a point in 4-dimensional space, and so on to any dimensionality n. Mathematically there is no problem with spaces of dimension greater than 3. The only problem lies in the possibility of visualization and intuitive understanding. As the number of variables and thus dimensions grows beyond 3, graphical representation and intuitive comprehension become impossible: who can visualize points in a 4-dimensional space, not to speak of a 40-dimensional one? It often helps to keep in mind that mathe-

matical dimension has no necessary connection with the 3 dimensions of the physical world.

Data typically consists of more or less numerous data items each of which is described in terms of the selected variables. Where vector space representation is used, each data item is described by a vector, and the data is consequently a collection of vectors. Such a collection is conveniently represented as a matrix in which the rows are the data items and the columns the variables. Thus, data consisting of m items each of which is described by n variables is represented by an $m \times n$ matrix D in which D_i (for $i = 1 \dots m$) is the i 'th data item, D_j (for $j = 1 \dots n$) is the j 'th variable, and D_{ij} the value of variable j for data item i .

	1	2	3	\dots	n
1	0.2	1.4	0.1	\dots	1.1
2	0.7	1.0	0.1	\dots	1.6
\vdots					
m	1.3	1.9	0.2	\dots	1.1

Fig. 40.3: A matrix

3.3. Variable value assignment

The semantics of each variable determines a particular interpretation of the domain of inquiry; the domain is “measured” in terms of the semantics, and that measurement constitutes the values of the variables. Measurement is fundamental in the creation of data because it makes the link between data and the world, and thus allows the results of data analysis to be applied to an understanding of the world (Pyle 1999, ch. 2).

Measurement is only possible in terms of some scale. There are various types of measurement scale, and these are discussed in the relevant textbooks (Hair et al. 2005; Pyle 1999, ch. 2), but for present purposes the main dichotomy is between numeric and non-numeric. The multivariate methods referred to in due course assume numeric measurement, and for that reason the same is assumed in what follows.

3.4. Data transformation

Once the data has been constructed, it may be necessary to transform it in various ways prior to analysis. Discussion of these in the abstract can quickly become intangible. To forestall this, a specific case is assumed: that the corpus being analyzed is a collection D of some number m of documents – of Middle English texts, say – and the research aim is to classify them on the basis of relative frequency of lexical types that they contain. The data abstracted from D is an $m \times n$ matrix Q in which

- (i) each of the rows Q_i , for $i = 1 \dots m$, represents a single data item – in this case, a document D_i .
- (ii) each column j , for $j = 1 \dots n$, is one of n variables, each representing a single lexical type that occurs at least once in D . ‘Lexical type’ is here defined as an abstraction

over a set of identical lexical tokens, ‘lexical token’ as a string of alphanumeric symbols, and ‘abstraction’ as a set label; the lexical type CAT = { $x | x = \text{‘cat’}$ }, for example. On the type-token distinction see Manning/Schütze (1999, 21–23, 124–130), Palmer (2000), and the discussion by Baroni in article 37 of this handbook.

- (iii) the matrix elements Q_{ij} contain integers representing the frequency of lexical type j in document i . Each of the m rows in Q is therefore a frequency profile for a single text.

		Variables, $j = 1..n$				
Documents $i = 1..m$		v1	v2	v3	...	v11
	d1	3	5	16	...	12
	d2	43	2	15	...	47
...				...		
		dm	29	0	27	126

Fig. 40.4: Lexical frequency data matrix

Obviously, this example is only one of many possibilities. The data items/matrix rows might be informants in a sociolinguistic or dialectological survey and the variables/matrix columns phonetic segments, or the rows might be phonetic segments and the columns phonetic features like voicing, and so on. The lexical frequency example was selected because it is generic with respect to a wide range of possible applications.

3.4.1. Adjustment for variation in document length

Documents in collections often vary in length. If the variation is substantial, the data abstracted from the collection must be adjusted to avoid distorted results. To see why, assume that all the documents are in the same language, and that, in this language, a given lexical type j has probability p_j of occurring. Then, the longer the document, the more likely it is that j will occur one or more times: if p_j is 0.01, then on average j will occur once every 100 words, twice every 200, and so on. Now, say that j occurs 10 times in two documents in D , d_i and d_k . Knowing only this, one would naturally judge that, in terms of their usage of j , the two documents are identical and that j is consequently of no use in distinguishing d_i from d_k . If, however, one also knows that d_i is 1000 words long, and d_k only 500, this is no longer the case. The frequency for d_i is what one would expect given p_j , but d_k uses j with higher-than-expected frequency, and this disparity can be used in distinguishing d_i from d_k : d_k , unlike d_i , is especially interested in what j denotes.

What is required is some way of adjusting the data matrix so that not just frequency but its significance relative to document length can be represented and thus incorporated into subsequent analysis. One approach is to transform the rows of the matrix into vectors of length 1:

$$Q_i = \frac{Q_i}{|Q_i|}$$

where $|Q_i|$ is the norm or length of row vector Q_i , defined as:

$$|Q_i| = \sqrt{Q_{i1}^2 + Q_{i2}^2 + \dots + Q_{in}^2}$$

The effect is to adjust the vector values representing document Q_i in proportion to the length of Q_i : the greater the length, the smaller the result of the division and thus the smaller the post-adjustment values in Q_i , and vice versa as the Q_i vector length shortens. This adjustment is part of a method for measuring relative proximity of document vectors in Information Retrieval called cosine normalization. Another approach is to transform the row vectors in relation to the average length of documents in the collection D:

$$Q_i = Q_i (\mu / l_{Q_i})$$

where (i) Q_i is the i 'th document's lexical frequency profile in the data matrix Q, (ii) l_{Q_i} is the total number of lexical tokens in document Q_i , and (iii) μ is the mean document length in terms of lexical tokens across all documents $d \in D$, so that:

$$\mu = \sum_{i=1 \dots m} (l_{Q_i}) / m$$

Thus, the values in each lexical frequency profile vector Q_i are multiplied by the ratio of the average number of lexical tokens per document across the collection D to the number of tokens in Q_i . The longer the document the numerically smaller the ratio, and vice versa; the effect is therefore to decrease the values in the vectors that represent long documents, and increase them in vectors that represent short ones, relative to average document length.

On transformation of data relative to document length see Belew (2000, 89–92); Lebart/Rajman (2000, 477–505); Singhal et al. (1996).

3.4.2. Sparsity minimization

Sparsity is a major issue in data analysis generally. The concept of the manifold is central to understanding why this is so. It comes from mathematical topology (Munkres 2000), a branch of pure mathematics concerned with geometrical properties; for present purposes it can be understood as the shape of data in n -dimensional space. What is the ‘shape’ of data (Pyle 1999, 84–86)? Consider a reasonably large data set of, say, 1000 3-dimensional real-valued vectors, no two of which are identical. If these vectors are plotted in 3-dimensional space, they form a cloud of points with an identifiable shape within the general space, as in Figure 40.5.

That shape is a manifold. The idea extends directly to any dimensionality, though such general spaces cannot be shown graphically. For the purposes of this discussion, therefore, a manifold is a set of vectors in n -dimensional space.

To discern the shape of a manifold, it is intuitively clear that there have to be enough data points to give it adequate definition. If, as in Figure 40.6a, there are just two points, the only reasonable manifold to propose is a line; any number of alternative manifolds are, of course, possible – the two points could come from a far more complex manifold like Figure 40.6c – but to propose this on the basis of just two points would clearly be

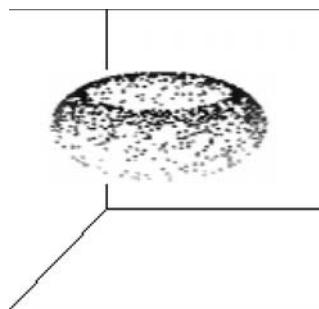


Fig. 40.5: A manifold in 3-dimensional space

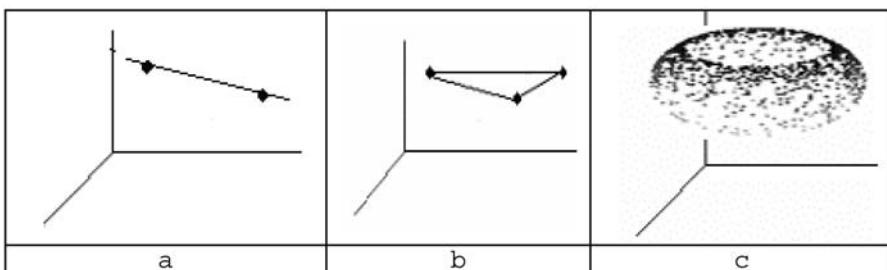


Fig. 40.6: Degrees of manifold definition

unjustified. Where there are three points a plane would be reasonable, as in Figure 40.6b. But it is only as the number of data points grows that the true shape of the manifold emerges, as in 6c. The general rule, therefore, is: the more data the better. After a certain point, increasing the amount of data becomes redundant in the sense that it simply confirms an already-clear manifold shape, but it doesn't do any harm.

In dealing with high-dimensional data, however, having too much is rarely a problem. Quite the opposite – the usual situation with high-dimensional data is that there is far too little. High-dimensional spaces are inherently sparse, and, to achieve an adequate definition of the data manifold, the amount of data required very rapidly becomes intractably large; this phenomenon is known as the ‘curse of dimensionality’. To see the problem, consider three data sets each of which contains 10 items, no two of which are identical:

- (i) Set 1 is univariate, and the single variable can take integer values in the range 1 ... 10. The ratio of data points to possible values is $10/10 = 1$, that is, the data points completely fill the data space.
- (ii) Set 2 is bivariate, and each of the two variables can take integer values in the range 1 ... 10. The ratio of data points to possible value pairs is $10/(10 \times 10) = 0.1$, that is, the data points occupy 10% of the data space.
- (iii) Set 3 is trivariate, and each of the three variables can take integer values in the range 1 ... 10. The ratio of data points to possible value triples is $10/(10 \times 10 \times 10) = 0.01$, that is, the data points occupy 1% of the data space.

And so on for increasing dimensionality: for a data set of fixed size d , the ratio of actual to possible points in the data space is d/r^n , where r is the number of different values that each variable can take (assuming for simplicity that all variables are identical in this respect). In other words, as dimensionality increases, the ratio of actual to possible points in the data space decreases at an exponential rate. In principle, therefore, a manifold consisting of some fixed number of vectors very rapidly becomes sparser as the dimensionality of the space in which it is embedded grows; to maintain its resolution at any preferred ratio, the number of vectors required must therefore grow exponentially with the dimensionality. Getting enough data becomes a serious problem even at relatively low dimensionalities, and an insuperable one soon thereafter. In practice the problem is not as severe as all this might suggest, since a typical real-world data set is not in general evenly or randomly spread around its vector space, but rather tends to be concentrated in one or more distinct regions of the space. Dimensionality nevertheless remains a potential problem for data analysis in any given application, and the moral is that dimensionality should be kept as low as possible, consistent with the need to describe the domain of inquiry adequately. For discussion of issues relating to high-dimensional data see Bishop (1995, chs.1,8); Pyle (1999, ch. 2, 355–360, 424–434); Verleysen (2003); Verleysen et al. (2003); Lee/Verleysen (2007).

Data sparsity has a particular relevance in corpus linguistics because the object of study is spoken or written natural language, and lexical distributions in samples of natural language have a characteristic shape. This shape is exemplified in a plot of lexical types in the Qur'an. The frequencies of these types were calculated, sorted into descending order of magnitude, and plotted:

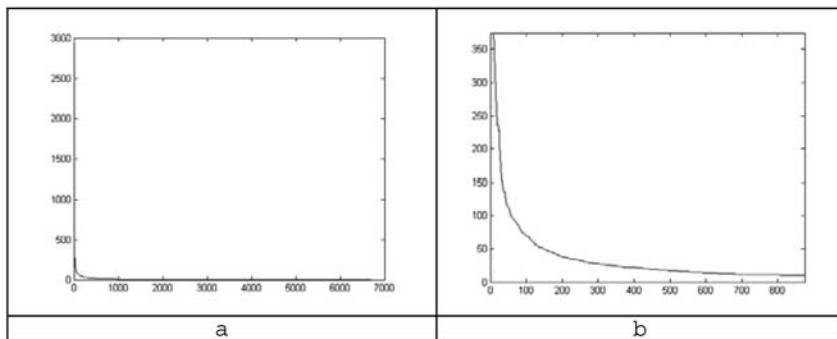


Fig. 40.7: Frequencies of lexical types in the Qur'an

Figure 40.7a is the full plot, and 40.7b is a zoomed-in region near the origin to display the shape of that region more clearly. There is a relatively small number of very frequent types, a moderate number of moderately frequent types, and a large number of very infrequent ones. This distribution is characteristic of lexical frequency distributions in natural language text generally (Baayen (2001); see also Manning/Schütze (1999, 20–29) and Baroni's discussion in article 37 of this handbook), and its shape remains pretty much constant even for natural language text corpora many orders of magnitude larger than the Qur'an: the number of occurrences of the few very frequent types continues to grow quickly as the corpus size grows and the number of occurrences of moderately

frequent types continues to grow moderately quickly, but the frequencies of the very infrequent types change hardly at all – instead, more and more types are added to the list. It is therefore clear that, in corpus linguistics studies where lexical frequency plays a role, the data will in general be very sparse on account of the large number of infrequent lexical type variables.

The obvious solution to sparsity is to select an optimal set of variables at the data design stage, but this is more easily said than done. As already noted, there is no algorithm for choosing an optimally relevant and therefore minimal set of variables for a research question, and therefore no way of knowing *a priori* whether the dimensionality of a given data set is as low as it can be. Because of this, a range of methods for transforming data matrices so as to reduce their dimensionality has been developed, and, in cases where the data is sparse, application of one or more of these methods can very substantially improve analytical results by giving the manifold better definition.

a) Stemming

Fundamental to the morphologies of many languages is the process whereby prefixes and suffixes are attached to lexical stems, and/or the lexical stems themselves are mutated in some way, in order to mark syntactic function or some modification to the primitive semantic denotation of the stem. Document collections written in such languages typically contain more or less numerous morphological variants of primitive lexical stems. Such variants can be considered to be equivalent for purposes of text analysis and information retrieval; stemming is the reduction of morphological variants to their common primitive stem.

Where lexical frequency plays a role in data creation, stemming offers scope for substantial dimensionality reduction. If a lexical type is defined as the set of identical alphabetic strings, then each variant of a given stem is treated as a distinct lexical type and assigned a column in the data matrix. If, however, all the morphological variants of a stem are collapsed into an equivalence class which then constitutes the lexical type, so that, for example, the type $CAT = \{x | x = \text{a morphological variant of 'cat'}\}$ such as ‘cats’, ‘catty’, ‘cattery’ and so on, the number of types and thus columns of the frequency matrix can be more or less substantially reduced, depending on the morphological characteristics of the language in question. The frequency of a lexical type so defined in the frequency matrix is then the sum of the frequencies of the aggregated variants.

At first glance it might seem that creation of such equivalence classes loses information, and that this loss is bound adversely to affect the validity of analyses based on the data. Just the opposite is true, however. If lexical types are regarded as sets of identical tokens, then each type is represented as a separate variable column in the data matrix, and all columns are treated equally in the analytical methods cited later. The implication is that morphologically related tokens are treated exactly the same as unrelated ones. In other words, there is no distinction between the semantic distances among morphological variants of a single stem on the one hand, and those between unrelated stems on the other – the semantic difference between ‘administer’ and ‘administration’ is taken to be the same as that between ‘administer’ and ‘cow’. If, as here, the aim is to classify documents on the basis of their lexical semantics, this is bound to distort the data and thus the analytical results based on it. Creation of equivalence classes based on morphological relatedness eliminates this distortion.

On stemming algorithms and their application in computational text processing, see Frakes (1992) and Hull (1996).

b) Variable selection

A seminal principle in Information Retrieval, extensively confirmed by empirical results, is that not all lexical types in a document collection are equally useful in document classification (for example, Belew 2000, ch. 3; van Rijsbergen 1979, ch. 2; Salton/McGill 1983, ch. 3). Various ways of identifying relatively more useful variables exist, and this section gives an overview of some of the most often used ones. The focus is on lexical frequency in document collections, but the techniques are straightforwardly applicable to other kinds of data, and thus to a wide range of analyses in corpus linguistics.

(i) Dimensionality reduction based on lexical type frequency

Luhn, one of the founders of modern Information Retrieval, proposed that the relative frequency of lexical types in a document collection is a fundamental criterion for classifying documents relative to one another (discussed in Belew 2000, 76 ff.; van Rijsbergen 1979, 15 ff; Salton/McGill 1983, 60–63). The intuition underlying this is simple: if an author uses a word repeatedly in a text, then the text is more likely to be about what the word denotes than it is to be about the denotation of a word that is infrequently used; documents with similar lexical frequency profiles are classified together and distinguished from those whose profiles are different. Luhn also observed, however, that the usefulness of a lexical type for document classification does not increase monotonically with frequency, and, more specifically, that very frequent types on the one hand and very infrequent ones on the other are less useful for the purpose than medium frequency ones. He therefore proposed that both very infrequent and very frequent words be discarded. Substantial dimensionality reduction can be achieved in this way, but Luhn did not provide any clear criteria for determining upper and lower frequency thresholds, and there is consequently the ever-present danger that too many or too few types will be eliminated, thus compromising classification based on the set of retained variables.

(ii) Dimensionality reduction based on variance

As we saw in the foregoing discussion of data, any variable x is an interpretation of some aspect of the world, and a value assigned to x is a measurement of the world in terms of that interpretation. If x is to describe more than one object – the heights of 1000 people, say – then it must take values characteristic of each person. Unless all 1000 people are exactly the same height, these values will vary. This possibility of variation gives x its descriptive utility: a constant value for x says that what x represents in the world does not change, moderate variation in the value says that that aspect of the world changes only a little, and widely differing values that it changes substantially. In general, therefore, the possibility of variation in the values assigned to variables is fundamental to the ability of variables to represent reality.

Classification of documents or of anything else therefore depends on there being variation in their characteristics. When the objects to be classified are described by variables, then the variables are only useful for the purpose if there is significant variation in the values that they take. If, for example, a large random collection of people was described by variables like height, weight, and income, there would be substantial variation in values for each of them, and they could legitimately be used to classify the people in the sample. On the other hand, a variable like ‘has nose’ would be effectively useless, since, with very few exceptions, everyone has a nose – there would be almost no varia-

tion in the boolean value 1 for this variable. In any classification exercise, therefore, one is looking for variables with substantial variation in their values, and can disregard variables with little or no variation.

Mathematically, the degree of variation in the values of a variable is described by its variance. The variance of a set of variable values is the average deviation of those values from their mean. Assume a set of n values $\{x_1, x_2 \dots x_n\}$ assigned to a variable x . The mean of these values μ is $(x_1 + x_2 + \dots + x_n)/n$. The amount by which any given value x_i differs from μ is then $x_i - \mu$. The average difference from μ across all values is therefore $\sum_{i=1 \dots n} (x_i - \mu)/n$. This average difference of variable values from their mean almost but not quite corresponds to the definition of variance. One more step is necessary, and it is technical rather than conceptual. Because μ is an average, some of the variable values will be greater than μ , and some will be less. Consequently, some of the differences $(x_i - \mu)$ will be positive and some negative. When all the $(x_i - \mu)$ are added up, as above, they will cancel each other out. To prevent this, the $(x_i - \mu)$ are squared. The standard definition of variance for n values $\{x_1, x_2 \dots x_n\}$ assigned to a variable x , therefore, is:

$$\sigma^2 = \left(\sum_{i=1 \dots n} (x_i - \mu)^2 \right) / n$$

Given a data matrix Q in which the rows are cases and the columns are lexical type variables describing the cases, and also that the aim is to classify the cases on the basis of the differences among them, the application of variance to dimensionality reduction is straightforward: eliminate all variables with low variance, that is, variables whose values do not vary enough for them to be useful in document classification. As with the upper and lower thresholds discussed in the preceding section, this begs the question of how low is too low, that is, of selecting a threshold.

(iii) Lexical frequency distribution

Spärck Jones (1972) proposed what was to become a standard principle in Information Retrieval: that a lexical type's usefulness is determined not by its absolute frequency across a collection, but by the pattern of variation in its frequency across the documents. To gain an intuition for this, assume a collection of documents related to the computer industry. At one end of the range are very low frequency words that, as expected, are of little or no use for document classification: a word like 'coffee' that occurs a few times in one or two documents that caution against spills into keyboards is insignificant in relation to the semantic content of the collection as a whole, and a word like 'bicycle' that occurs only once tells us only that the document in which it appears is unique on that criterion. At the other end of the range, a word like 'computer' and its morphological variants is likely to be both very frequent across the collection and to occur in most if not all the documents, and as such is a poor criterion for classifying documents despite its high absolute frequency: if all the documents are about computers, being about computers is not a useful distinguishing criterion. In short, lexical frequency on its own is not a reliable classification criterion. The most useful lexical types are those whose occurrences are both relatively frequent and not, like 'computer', uniformly spread across all collection documents but rather occur in clumps, such that a relatively few documents

contain most or all the occurrences, and the rest of the collection few or none; ‘debug’, for example, can be expected to occur frequently in documents that are primarily about computer programming and compiler design, but only infrequently if at all in those about, say, word processing. On this criterion, lexical types are selected in accordance with their ‘clumpiness’ of occurrence across documents in a collection.

Three methods used in Information Retrieval for determining clumpiness in data are:

- (i) TF.IDF (‘Term Frequency \times Inverse Document Frequency’): Belew (2000, 84–5); Buckley (1993); Robertson (2004); Roberston/Spärck Jones (2004); Salton/McGill (1983, 63); Spärck-Jones (1972).
- (ii) Signal-noise ratio: Belew (2000, 83–4); Salton/McGill (1983, 63–6).
- (iii) Poisson term distribution: Belew (2000, 73 ff.); van Rijsbergen (1979, 27–9); Church/Gayle (1995a, 1995b).

Space constraints do not permit these to be described here, and the reader is referred to the cited references.

c) Variable redefinition

Dimensionality reduction can be achieved by replacing the variables that have been chosen to describe the domain of interest with different variables that describe the domain as well as, or almost as well as, the originals, but are fewer in number.

We have seen that a data set of n -dimensional vectors defines a manifold in n -dimensional space. In such a space, it is possible in principle to have manifolds whose dimensionality is k , where $k < n$. Consider the 3-dimensional data set in Figure 40.8a.

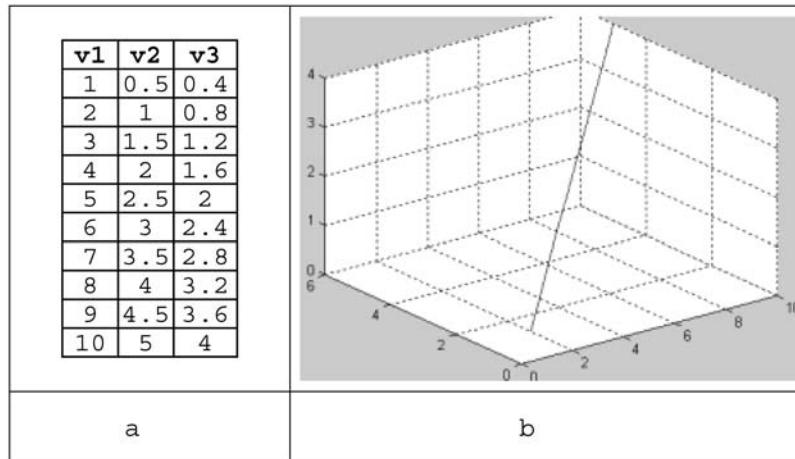


Fig. 40.8: A 1-dimensional manifold in 3-dimensional space

Plotting this data in 3-dimensional space (Figure 40.8b) shows it to describe a line. But that line can be redescribed in 2 dimensions.

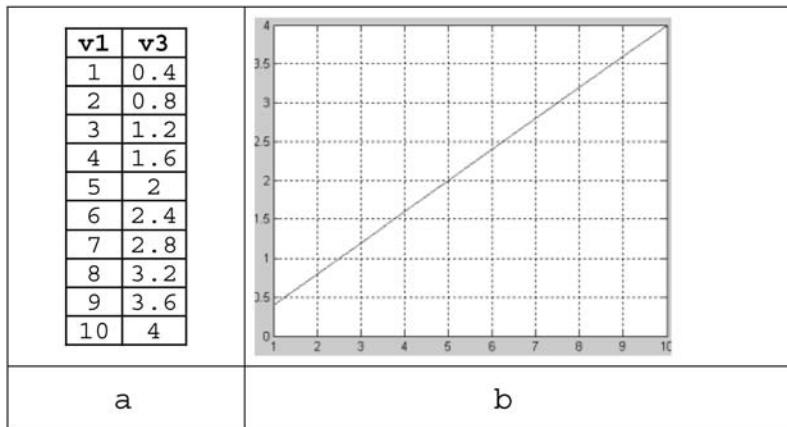


Fig. 40.9: A 1-dimensional manifold in 2-dimensional space

In fact, the line can be redescribed in 1 dimension – its length, 10.63 – by its distance from 0 on the real-number line:

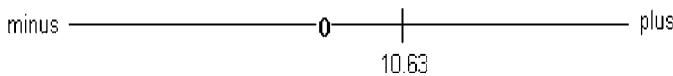


Fig. 40.10: A 1-dimensional manifold in 1-dimensional space

Consider another example – a plane in 3-dimensional space:

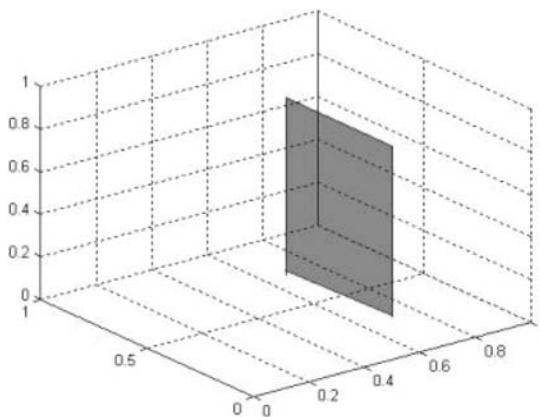


Fig. 40.11: A 2-dimensional manifold in 3-dimensional space

This plane can be redescribed in 2-dimensional space.

And, as usual, this concept extends straightforwardly to any dimensionality.

In general, therefore, a line can be described in one dimension, 2 dimensions, 3 dimensions, or any number of dimensions one likes. Essentially, though, it is a 1-dimensional object; its ‘intrinsic dimensionality’ (Verleysen 2003; Lee/Verleysen 2007) is 1. The minimum number of dimensions required to describe a line is 1; higher-dimensional descrip-

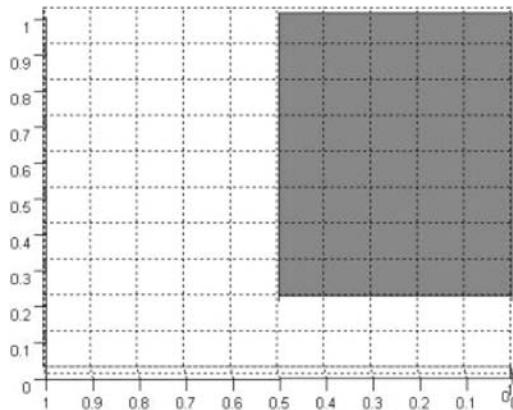


Fig. 40.12: A 2-dimensional manifold in 2-dimensional space

tions are possible but redundant. A plane was described in 2 and 3 dimensions. Could it also, like a line, be described in 1 dimension? No: the intrinsic dimensionality of a plane is 2 – the corresponding data set must be 2-dimensional at least, giving the coordinates of the points that describe it.

The concept of intrinsic dimensionality applies straightforwardly to dimensionality reduction. The informational content of data is conceptualized as a k -dimensional manifold in the n -dimensional space defined by the data variables. Where $k = n$, that is, where the intrinsic dimensionality of the data corresponds to the number of data variables, no dimensionality reduction is possible without significant loss of information. However, the foregoing discussion of data creation noted that, when describing a domain of interest, selection of variables is at the discretion of the researcher. It is therefore possible that the selection of variables in any given application will be suboptimal in the sense that there is redundancy among the variables, that is, that they overlap with one another in terms of the information they represent about the domain; where there is a significant amount of redundancy, it is possible in principle to represent this information using a smaller number of variables, thus reducing the dimensionality of the data. In such a case, the aim of dimensionality reduction of data is to discover its intrinsic dimensionality k , for $k < n$, and to redescribe its informational content in terms of those k dimensions.

The most often used variable redefinition method is principal component analysis (PCA), on which see Jolliffe (2002); briefer accounts are in Bishop (1995, 310–314); Everitt/Dunn (2001, ch. 3); Grimm/Yarnold (1995, 99–134); Hair et al. (2005); Oakes (1998, 96–108); Tabachnick/Fidell (2006); Webb (2002, 319–44); Woods/Fletcher/Hughes (1986, ch. 15). PCA is a particular case of Singular Value Decomposition (SVD) on which see Lebart/Rajman (2000) and Manning/Schütze (1999, 554–566). Both PCA and SVD are linear methods; nonlinear variable definition methods are described in Diamantaras/Kung (1996); Bishop (1995, 314–319); Pyle (1999, 355–383; Lee/Verleysen 2007).

3.4.3. Data linearization

In physical systems there is a fundamental distinction between linear and nonlinear behaviour. To get an intuition for what is involved, and why the distinction is important,

here is an experiment. Kick a ball and note how far it goes. Kick it again, but this time twice as hard, and once again note how far it goes. The natural expectation is that it will go twice as far, and this expectation is fulfilled. This is linear behaviour: the effect is proportional to the cause. But take the experiment further. Kick the ball in a series, each time twice as hard as the time before: k, 2k, 4k, 8k and so on. If it goes 10 metres for k, and 20 metres for 2k, will it also go 40 metres for 4k, and 80 metres for 8k? No. As it is kicked harder and harder, it goes faster and further. Air resistance becomes a factor at higher speeds, and so does rolling resistance. The ball might only go 78 metres for an 8k kick, and 150 metres for a 16k kick, etc. Eventually, the kick will be so hard that the ball bursts and goes hardly any distance at all. This is nonlinear behaviour: it is the breakdown of proportionality between cause and effect in physical systems, and it can generate a variety of complex and often unexpected – including chaotic – behaviours. In nature there are few truly linear systems. Nonlinearity pervades the physical world (Bertuglia 2005), and, because it does, data manifolds that describe the world are likely to contain nonlinearities. Figure 40.13a shows a linear relationship between two variables x and y , and Figure 40.13b a nonlinear one:

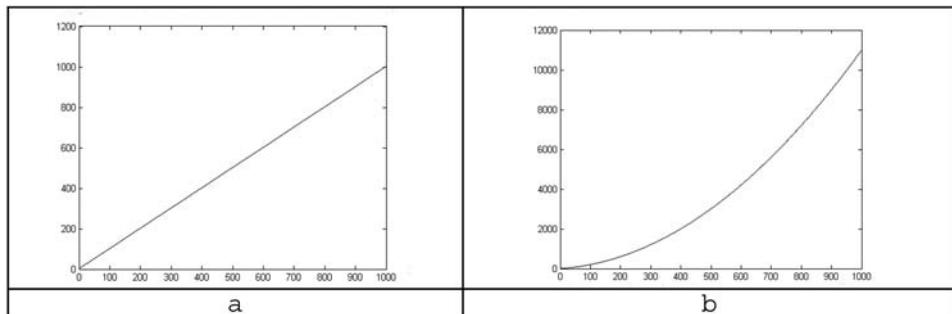


Fig. 40.13: Linear and nonlinear 1-dimensional manifolds

In the linear case there is an invariant proportionality between x and y , and that invariance is represented by a straight line; in the nonlinear case, the relationship between x and y varies with different values of x , and that variance is represented by a curved line. In 3 dimensions, linear data might generate a plane (Figure 40.14a) and nonlinear data a curved surface (Figure 40.14b).

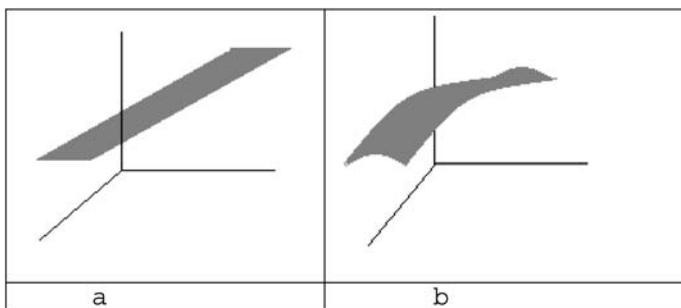


Fig. 40.14: Linear and nonlinear 2-dimensional manifolds

In general, linear manifolds are lines and planes, and nonlinear ones curves and curved surfaces; these cannot be shown graphically for higher dimensionalities. Nonlinear manifolds can range from fairly simple curves, as above, to highly complex ones.

The first step is to determine whether or not a given matrix in fact contains significant nonlinearity. This seems obvious, but, for high-dimensional data, it is not always or even usually straightforward. In the light of the foregoing observation that nonlinearity pervades the natural world, the strong suspicion must be that the generating process is nonlinear, but this is not certain. Even if the generating process is known to be nonlinear, moreover, there is no guarantee that every data set it generates will contain nonlinearities. This sounds paradoxical, but consider the shape of the familiar nonlinear logistic function, which models a range of natural processes.

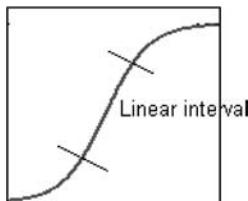


Fig. 40.15: Graph of logistic function

Though it is nonlinear globally, there is a relatively large interval that is linear or near-linear; if the data of interest happens to come from that interval of output values, then it is linear even though it was generated by a nonlinear process. *A priori* reasoning cannot, in short, establish whether or not a data set contains significant nonlinearities. Only direct examination of the data will establish this. The usual method is to plot pairs of variables and then examine the plots for deviation from linearity, but where the number of variables is large this quickly becomes burdensome (Hair et al. 2005; Tabachnick/Fidell 2006; Moisl 2007). One alternative is to linearize the data matrix (for example Croft/Davison/Hargreaves 1992, 350–64). The nonlinearities may, however, themselves be of interest, in which case linearization throws the baby out with the bath water. Another alternative is to use an analytical method that can accommodate nonlinearities, on which more below.

4. Exploratory multivariate methods

Exploratory methods are essentially variations on a theme: cluster analysis. Cluster analysis aims to identify and graphically to represent nonrandomness in the distribution of vectors in n -dimensional space. Spatial regularities in the graphical representations are interpreted as reflecting regularities in the natural process that generated the data, and support hypotheses about the characteristics of the process. In Figure 40.16a, for example, the vectors are spread more or less uniformly in 2-dimensional space; there are some local concentrations, but these are not clearly defined and it is difficult to infer anything about the process that generated the data other than that it appears to be broadly random. In Figure 40.16b, on the other hand, there are clearly defined concen-

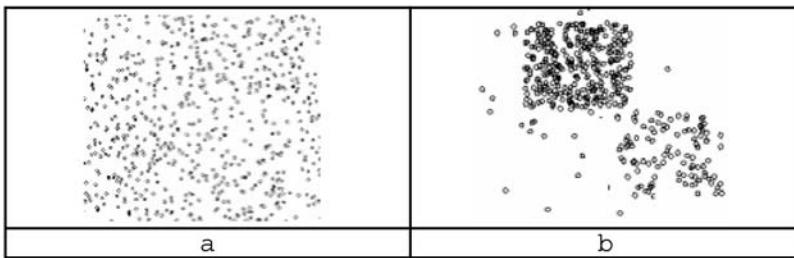


Fig. 40.16: Random and nonrandom data

trations of vectors such that two groups of points are spatially relatively close in each group, and spatially relatively far from each other, which suggests that the generating process is strongly nonrandom.

In 2 or 3 dimensions, such distributions can be plotted and interpreted by eye. In higher dimensions this is no longer possible, however; the various cluster analysis methods are just different ways of representing nonrandom structure in higher-dimensional data graphically in 2 or 3-dimensional space.

There is an extensive range of cluster analysis methods together with a large associated literature: for example Arabie/Hubert/de Soete (1992); Duda/Hart/Stork (2001, ch. 10); Everitt/Dunn (2001); Everitt/Landau/Leese (2001); Gordon (1999); Gore (2000); Grimm/Yarnold (1995, 2000); Hair et al. (2005); Jain/Dubes (1988); Jain/Murty/Flynn (1999); Kachigan (1991); Manning/Schütze (1999, ch. 14); Oakes (1998, ch. 3); Tan/Steinbach/Kumar (2006, ch. 8–9); Tinsley/Brown (2000); Tabachnick/Fidell (2006); Webb (2002, ch. 10); Woods/Fletcher/Hughes (1986, ch. 14). There is no space to describe individual methods in detail here, so what follows gives an overview in three parts. The first part introduces basic issues in cluster analysis, the second cites some commonly used methods, and the third issues a caution about using those methods.

a) Basic issues

The most important thing to realize about cluster analysis is that there is no single “best” method; see for example Everitt/Landau/Leese (2001, ch. 8); Tan/Steinbach/Kumar (2006, 639–42). In any particular application, selection of one or more methods must be informed by a variety of considerations, three of the most important of which are:

(i) How much is known about the cluster structure of the data?

In cluster analysis there is a distinction between methods which make no a priori assumptions about the structure of given data and attempt to discover clusters purely on the basis of the data’s characteristics, and those which presuppose that the data has a cluster structure and requires specification of the number of clusters in advance of analysis. If little or nothing is known about the cluster structure, then one of the former methods is appropriate, but if there is a reasonable degree of certainty about its structure then one of the latter type of method, such as k-means clustering or kernel-based adaptive algorithms, can be used (for a survey of these methods, see Webb 2002). The present discussion is concerned with exploratory analysis, and as such is henceforth concerned only with methods that make no a priori assumptions about data.

(ii) Is the data linear or nonlinear?

The selected method or methods must be compatible with the data being analyzed. For continuous-valued numerical data such as that being discussed here, the main criterion for compatibility is whether the data manifold is linear or not. Data that contains significant nonlinearity must be analyzed using a nonlinear clustering method; use of a linear method in such a case misrepresents the structure of the data to greater or lesser degrees, depending on the nature of the nonlinearity. What does it mean for a method to be linear or nonlinear? Assume a curved manifold in n -dimensional space. What is the distance d_{ij} between any two points i and j on that manifold? A linear method measures that distance as a straight line joining the points, ignoring the manifold's curvature, whereas a nonlinear method measures the distance along the surface of the manifold, thereby taking account of the curvature. Depending on the amount of curvature, the difference between the two measures can be significant and can therefore significantly affect analysis based on it. An example is the distance between two points A and B on the perimeter of the circle in Figure 40.17: the linear distance between them is a chord drawn through the interior, and the nonlinear one the length of the perimeter segment between the points as indicated by the arc in the figure:

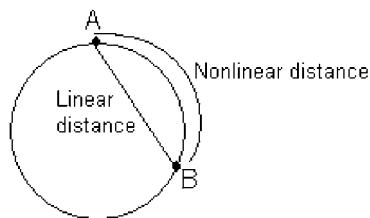


Fig. 40.17: Linear and nonlinear distance

Consider, for example, the problem of discovering a classification for the data in Figure 40.18a. The data space must be partitioned such that all the points in the left-hand cluster fall into one partition, and all the points in the right-hand cluster into another. Linear methods are, by definition, limited to doing this using straight lines or surfaces; in this case, that is sufficient.

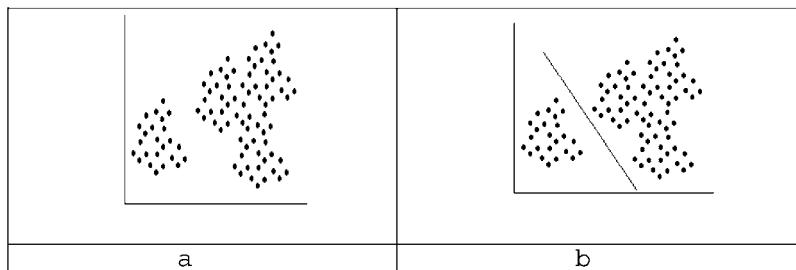


Fig. 40.18: Linearly separable clusters

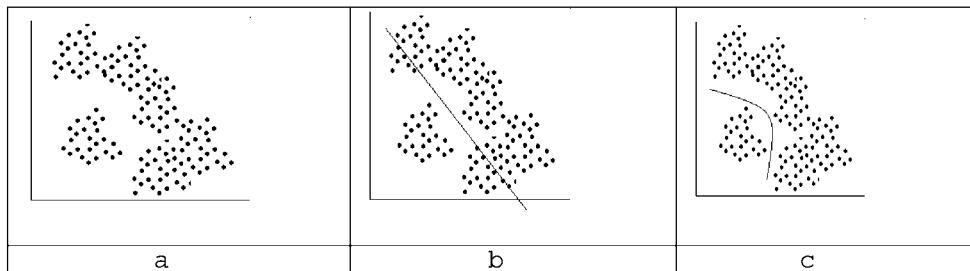


Fig. 40.19: Nonlinearly separable clusters

For the data in Figure 40.19a, however, there is no straight line that can separate the two clusters without misclassifying some of the points, as in 19b. What is required for correct classification is a method for finding a nonlinear partition, as in 19c.

(iii) Is a hierarchical or nonhierarchical analysis required?

The fundamental aim of exploratory analysis is to generate hypotheses about some domain of inquiry, and it may be that, in any particular case, some methods provide representations of structure that do this more usefully than others. The main distinction among methods in this regard is between those that generate hierarchically ordered clusters, and those that do not and are therefore described as nonhierarchical. Nonhierarchical methods generate graphical representations in 2 or 3-dimensional space such that, given a suitable measure of proximity, vectors which are spatially or topologically relatively close to one another in high-dimensional space are spatially or topologically close to one another in their 2 or 3-dimensional representation, and vectors which are relatively far from one another in high-dimensional space are clearly separated, either by relative spatial distance or by some other graphical means, resulting – in the case of nonrandom data – in a configuration of well defined clusters. Figures 40.18 and 40.19 are a 2-dimensional example; a 3-dimensional one might look like Figure 40.20:

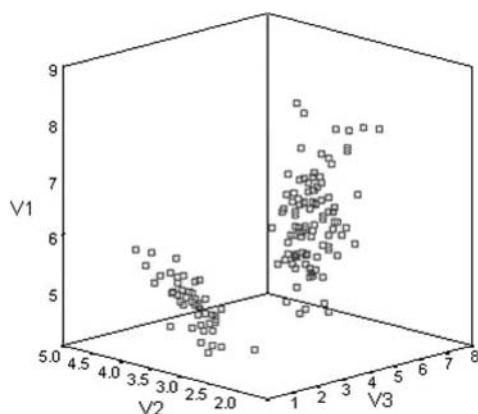


Fig. 40.20: Clusters in 3-dimensional space

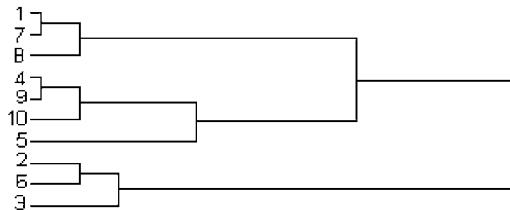


Fig. 40.21: A cluster dendrogram

Hierarchical methods, on the other hand, represent proximity structure in high-dimensional data not as spatial clusters but as ‘dendograms’:

A dendrogram is simply a tree of the kind linguists are familiar with from sentence structure analysis. It is shown horizontally rather than in the vertical orientation that is more usual in linguistics in order to make it more readily representable on a page, and the labels at the ‘leaves’ are not lexical tokens but labels for the vectors in the data set – ‘1’ is the first vector, ‘2’ the second, and so on. Like a linguistic phrase structure tree, a dendrogram shows constituency structure: in this tree, vectors 1 and 7 constitute a ‘phrase’. Combined with vector 8, they form a superordinate ‘phrase’, which itself combines with (4,9,10,5) to form an even higher-level ‘phrase’, and so on. Unlike a linguistic phrase structure tree, however, this one represents not grammatical constituency but vector proximity in n -dimensional space: vectors 1 and 7 are relatively very close and both of them are quite close to 8; vectors 4 and 9 are relatively close and both are quite close to 10; and so on. And, again unlike grammatical phrase structure trees, the lengths of the branches linking ‘phrases’ represent relative degrees of proximity: that the lines for linking 4 and 9 are relatively very short indicates that the corresponding vectors are close in n -dimensional space, but the relatively long lines between (1, 7, 8) and (4, 9, 10, 5) indicate considerable distance. In the light of this, the cluster interpretation of the above tree is straightforward: there are two main clusters: (1, 7, 8, 4, 9, 10, 5) and (2, 6, 3); within each of the two main clusters there are subclusters (1, 7, 8) and (4, 9, 10, 5), and (2, 6) and (3); and so on.

b) Cluster analysis methods

This section lists some widely-used exploratory cluster analysis methods. It is not even nearly exhaustive; an extensive range of methods is available, for which the reader is referred to the literature on cluster analysis cited earlier in this section.

(i) Linear methods

Hierarchical linear methods comprise a group of closely related algorithms which define proximity in n -dimensional space, the nature of a cluster, and clustering algorithms in a variety of ways: see for example Duda/Hart/Stork (2001, ch. 10); Everitt/Landau/Leese (2001); Everitt/Dunn (2001, ch. 6); Gordon (1999, 69–109); Gore (2000); Hair et al. (2005); Jain/Murty/Flynn (1999, 275–279); Kachigan (1991, 261–270); Oakes (1998, 110–120).

Nonhierarchical linear methods include PCA and SVD in that, if dimensionality is reduced to 2 or 3, the data vectors can be displayed, and any clusters visually identified using conventional plotting tools. Another widely-used linear nonhierarchical method is Multidimensional Scaling: Borg/Groenen (2005); Davison/Sireci (2000); Everitt/Dunn

(2001, ch. 5); Gordon (1999, 157–167); Grimm/Yarnold (1995, 137–168); Hair et al. (2005); Kachigan (1991, 271–278); Woods/Fletcher/Hughes (1986, 262–265).

(ii) Nonlinear methods

In parallel with linear PCA and SVD, nonlinear variable redefinition methods can be used for cluster analysis if dimensionality is reduced to 2 or 3 (Bishop 1995, 314–319; Pyle 1999, 358–383; Diamantaras/Kung 1996; Duda/Hart/Stork 2001, 569–570; Grimes 2006). Beyond these, there is a good range of methods, such as Isomap (Tenenbaum/de Silva/Langford 2000), Locally Linear Embedding (Roweis/Saul 2000, and the very widely used Kaski/Kangas/Kohonen 1998; Oja/Kaski/Kohonen 2001) Self-Organizing Map (Kohonen 2001).

c) Caution

Different methods can and often do generate different results when applied to the same data. This is partly because the methods make explicit or implicit assumptions about what constitutes a cluster and how clusters so defined can be algorithmically identified, and partly because they depend to greater or lesser degrees on parameter values that are user-specified, often on a heuristic basis. It is not obvious which method and/or combination of parameter values is to be preferred in any specific application, or why. This leads to an obvious question: what are these clustering methods really telling us about the structure of the data they describe – how reliable, in other words, are they, and are they in fact of any use at all if they cannot be relied on to reveal the true structure of the data?

In the literature there are two main approaches to an answer. One is to attempt to establish the validity of cluster results using numerical measures (Everitt/Landau/Leese 2001, ch. 8; Duda/Hart/Stork 2001, 557–559; Tan/Steinbach/Kumar 2006, 532–555). The other approach is to apply a variety of different clustering methods to the same data and to compare the results: a clear convergence on one particular cluster structure is held to support the validity of that structure with respect to the data. And, of course, the two approaches can be used in combination.

5. Exploratory multivariate analysis in corpus linguistics

Any collection of written or spoken language potentially comes within the remit of corpus linguistics, and as such “corpus linguistics” can include not only the traditional subdisciplines of linguistics proper such as phonology, morphology, and so on, but also the philological work that comes under the heading “humanities computing” as well as information retrieval and data mining from full-text collections. To keep the length of this section within reasonable bounds, therefore, there is no attempt at exhaustiveness. The aim is rather to provide a selection of references that is representative of the applications of exploratory multivariate methods to language corpora.

- Language classification: Kita (1999).
- Phonetics & phonology: Berdan (1978); Miller/Nicely (1955); Shepard (1972); Jassem/Lobacz (1995).
- Morphology: Oakes/Taylor (1994).
- Syntax: Gries (2001); Gamallo/Agustini/Lopes (2005).

- Lexical semantics & word-sense disambiguation: Yarowsky (2000); Stevenson/Wilks (2003); Pedersen (2006); Watters (2002); Landauer/Foltz/Laham (1998); Zernik (1991).
- Dialectology: Babitch/LeBrun (1989); Chambers/Trudgill (1998, 135–148); Heeringa/Nerbonne (2001); Kessler (1995); Kleiweg/Nerbonne/Bosveld (2004); Nerbonne/Heeringa (2001).
- Sociolinguistics: Chambers (2003); Horvath (1985); Jones-Sargent (1983); Moisl/Jones (2005); Moisl/Maguire/Allen (2006); Sankoff et al. (1989).
- Language register / textual genre variation: Biber and his co-workers have published extensively on this topic. See Biber's discussion in article 38 of this handbook, which also contains a full bibliography.
- Text classification: Lebart/Rajman (2000); Willett (1988); Manning/Schütze (1999, ch. 16). Included here also is the large amount of work in Information Retrieval and Data Mining: see for example Belew (2000); Salton/McGill (1983); van Rijsbergen (1979); Strzalkowski (1999); Tan/Steinbach/Kumar (2006); Webb (2002).
- Stylometry: Hoover (2003); Ledger (1995); Linmans (1998); McEnery/Oakes (2000); Mealand (1995); Temple (1996). See also Oakes' discussion in article 50 of this handbook.

6. Literature

- Arabie, P./Hubert, L./de Soete, G. (eds.) (1992), *Clustering and Classification*. River Edge, New Jersey: World Scientific Press.
- Baayen, R. H. (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.
- Babitch, R./LeBrun, E. (1989), Dialectometry as Computerized Agglomerative Hierarchical Classification Analysis. In: *Journal of English Linguistics* 22, 83–90.
- Belew, R. (2000), *Finding Out About: A Cognitive Perspective in Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.
- Berdan, R. (1978), Multidimensional Analysis of Vowel Variation. In: Sankoff, D. (ed.), *Linguistic Variation. Models and Methods*. New York: Academic Press, 149–160.
- Bertuglia, C. (2005), *Nonlinearity, Chaos, and Complexity: The dynamics of Natural and Social Systems*. Oxford: Oxford University Press.
- Bishop, C. (1995), *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.
- Borg, I./Groenen, P. (2005), *Modern Multidimensional Scaling: Theory and Applications*. 2nd ed. Berlin: Springer.
- Buckley, C. (1993), The Importance of Proper Weighting Methods. In: Bates, M. (ed.), *Human Language Technology*. San Mateo, CA: Morgan Kaufmann.
- Chambers, J. (2003), *Sociolinguistic Theory. Linguistic Variation and its Social Significance*. 2nd ed. Oxford: Blackwell Publishers.
- Chambers, J./Trudgill, P. (1998), *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.
- Church, K./Gale, W. (1995a), Poisson Mixtures. In: *Natural Language Engineering* 1, 163–190.
- Church, K./Gale, W. (1995b), Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. In: *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, MA, 121–130.
- Croft, A./Davison, R./Hargreaves, M. (1992), *Engineering Mathematics*. Wokingham: Addison-Wesley.
- Dale, R./Moisl, H./Somers, H. (eds.) (2000), *Handbook of Natural Language Processing*. New York: Marcel Dekker.

- Davison, M./Sirci, S. (2000), Multidimensional Scaling. In: Tinsley/Brown 2000, 323–352.
- Diamantaras, K./Kung, S. (1996), *Principal Component Neural Networks: Theory and Applications*. New Jersey: Wiley Interscience.
- Duda, R./Hart, P./Stork, D. (2001) *Pattern Classification*. 2nd ed. New York: Wiley Interscience.
- Everitt, B./Landau, S./Leese, M. (2001), *Cluster Analysis*. 4th ed. London: Arnold.
- Everitt, B./Dunn, G. (2001), *Applied Multivariate Data Analysis*. 2nd ed. London: Arnold.
- Frakes, W. (1992), Stemming Algorithms. In: Frakes, W./Baeza-Yates, R. (eds.), *Information Retrieval: Data Structures & Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 131–160.
- Gamallo, P./Agustini, A./Lopes, G. (2005), Clustering Syntactic Positions with Similar Semantic Requirements. In: *Computational Linguistics* 31, 107–146.
- Gordon, A. (1999), *Classification*. 2nd ed. London: Chapman & Hall.
- Gore, P. (2000), Cluster Analysis. In: Tinlsey/Brown 2000, 297–321.
- Gries, S. Th. (2001), Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited. In: *Journal of Quantitative Linguistics* 8, 33–50.
- Grimes, C. (2006), *Nonlinear Dimensionality Reduction*. London: Chapman & Hall/CRC.
- Grimm, L./Yarnold, P. (eds.) (1995), *Reading and Understanding Multivariate Statistics*. Washington, DC: American Psychological Association.
- Grimm, L./Yarnold, P. (eds.) (2000), *Reading and Understanding More Multivariate Statistics*. Washington, DC: American Psychological Association.
- Hair, J./Black, B./Babin, B./Anderson, R./Tatham, R. (2005), *Multivariate Data Analysis*. 6th ed. London: Prentice-Hall International.
- Heeringa, W./Nerbonne, J. (2001), Dialect Areas and Dialect Continua. In: *Language Variation and Change* 13, 375–400.
- Hoover, D. (2003), Multivariate Analysis and the Study of Style Variation. In: *Literary and Linguistic Computing* 18, 341–360.
- Horvath, B. (1985), *Variation in Australian English*. Cambridge: Cambridge University Press.
- Hull, D. (1996), Stemming Algorithms: A Case Study for Detailed Evaluation. In: *Journal of the American Society for Information Science*, 47(1), 70–84.
- Jain, A./Dubes, R. (1988), *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.
- Jain, A./Murty, M./Flynn, P. (1999), Data Clustering: A Review. In: *ACM Computing Surveys* 31, 264–323.
- Jassem, W./Lobacz, P. (1995), Multidimensional Scaling and its Applications in a Perceptual Analysis of Polish Consonants. In: *Journal of Quantitative Linguistics* 2, 105–124.
- Jolliffe, I. (2002), *Principal Component Analysis*. 2nd ed. Berlin: Springer.
- Jones-Sargent, V. (1983) *Tyne Bytes: A Computerized Sociolinguistic Study of Tyneside English*. Frankfurt: P. Lang.
- Kachigan, S. (1991), *Multivariate Statistical Analysis. A Conceptual Introduction*. New York: Radius Press.
- Kaski, S./Kangas, J./Kohonen, T. (1998), Bibliography of Self-Organizing Map (SOM) Papers: 1981–1997. In: *Neural Computing Surveys* 1, 102–350.
- Kessler, B. (1995), Computational Dialectology in Irish Gaelic. In: *Seventh Conference of the European Chapter of the Association for Computational Linguistics*. University College Dublin, March 1995, 60–66.
- Kita, K. (1999), Automatic Clustering of Languages Based on Probabilistic Models. In: *Journal of Quantitative Linguistics* 6, 167–171.
- Kleweg, P./Nerbonne, J./Bosveld, L. (2004), Geographic Projection of Cluster Composites. In: Blackwell, A./Marriott, K./Shimojima, A. (eds.), *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004. Cambridge, UK, March 2004*. Berlin: Springer, 392–394.
- Kohonen, T. (2001), *Self-Organizing Maps*. 3rd ed. Berlin: Springer.
- Landauer, T./Foltz, P./Laham, D. (1998), Introduction to Latent Semantic Analysis. In: *Discourse Processes* 25, 259–284.

- Lebart, L./Rajman, M. (2000), Computing Similarity. In: Dale/Moisl/Sommers 2000, 477–505.
- Ledger, G. (1995), An Exploration of Differences in the Pauline Epistles Using Multivariate Statistical Analysis. In: *Literary and Linguistic Computing* 10, 85–97.
- Lee, J./Verleyen, M. (2007), *Nonlinear Dimensionality Reduction*. Berlin: Springer.
- Linmans, A. (1998), Correspondence Analysis of the Synoptic Gospels. In: *Literary and Linguistic Computing* 13, 1–13.
- Manning, C. D./Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Mealand, D. (1997), Measuring Genre Differences in Mark with Correspondence Analysis. In: *Literary and Linguistic Computing* 12, 227–245.
- McEnery, T./Oakes, M. (2000), Authorship Identification and Computational Stylometry. In: Dale/Moisl/Sommers 2000, 545–562.
- Miller, G./Nicely, P. (1955), An Analysis of Perceptual Confusion among English Consonants. In: *Journal of the Acoustic Society of America* 27, 338–352.
- Moisl, H. (2007), Data Nonlinearity in Exploratory Multivariate Analysis of Language Corpora. In: Nerbonne, J./Ellison, M./Kondrak, G. (eds.) *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Association for Computational Linguistics, 93–100.
- Moisl, H./Jones, V. (2005), Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A Comparison of Methods. In: *Literary and Linguistic Computing* 20, 125–146.
- Moisl, H./Maguire, W./Allen, W. (2006), Phonetic Variation in Tyneside: Exploratory Multivariate Analysis of the Newcastle Electronic Corpus of Tyneside English. In: Hinskens, F. (ed.), *Language Variation. European Perspectives*. Amsterdam: Meertens Institute, 127–141.
- Munkres, J. (2000), *Topology*. 2nd ed. New Jersey: Pearson Education International.
- Nerbonne J./Heeringa W. (2001), Computational Comparison and Classification of Dialects. In: *Dialectologia et Geolinguistica* 9, 69–83.
- Oakes, M. (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Oakes, M./Taylor, M. (1994), Morphological Analysis in Vocabulary Selection for the Ringdoc Pharmacological Database. In: Barahona, P./Veloso, M./Bryant, J. (eds.), *Proceedings of the 12th International Congress of the European Federation for Medical Informatics (1994)*. Lisbon, Portugal, 523–528.
- Oja, M./Kaski, S./Kohonen, T. (2001), Bibliography of Self-Organizing Map (SOM) Papers: 1998–2001. In: *Neural Computing Surveys* 3, 1–156.
- Palmer, D. (2000), Tokenisation and Sentence Segmentation. In: Dale/Moisl/Sommers 2000, 11–35.
- Pedersen, T. (2006), Unsupervised Corpus Based Methods for WSD. In: Agirre, E./Edmonds, P. (eds.), *Word Sense Disambiguation: Algorithms and Applications*. Berlin: Springer, 133–166.
- Pyle, D. (1999), *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.
- Robertson, S. (2004), Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. In: *Journal of Documentation* 60, 503–520.
- Robertson, S./Spärck Jones, K. (2004), IDF Term Weighting and IR Research Lessons. In: *Journal of Documentation* 60, 521–523.
- Roweis, S./Saul, L. (2000), Nonlinear Dimensionality Reduction by Locally Linear Embedding. In: *Science* 290, 2323–2326.
- Salton, G./Wong, A./Yang, C. (1975), A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18, 613–620.
- Salton, G./McGill, M. (1983), *Introduction to Modern Information Retrieval*. Auckland: McGraw-Hill.
- Sankoff, D./Cedergren, H./Kemp, W./Thibault, P./Vincent, D. (1989), Montreal French: Language, Class, and Ideology. In: Fasold, R./Schiffrin, D. (eds.), *Language Change and Variation*. Amsterdam: John Benjamins, 107–118.
- Shepard, R. (1972), Psychological Representation of Speech Sounds. In: David, E./Denes, P. (eds.), *Human Communication. A Unified View*. London: McGraw-Hill, 67–113.

- Singhal, A./Salton, G./Mitra, M./Buckley, C. (1996), Document Length Normalization. In: *Information Processing & Management* 32(5), 619–633.
- Spärck Jones, K. (1972), Exhaustivity and Specificity. In: *Journal of Documentation* 28, 11–21.
- Stevenson, M./Wilks, Y. (2003), Word-sense Disambiguation. In: Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 249–65.
- Strzalkowski, T. (1999), *Natural Language Information Retrieval*. Dordrecht: Kluwer.
- Tabachnick, B./Fidell, L. (2006), *Using Multivariate Statistics*. 5th ed. Boston: Allyn and Bacon.
- Tan, P./Steinbach, M./Kumar, V. (2006), *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.
- Temple, J. (1996), A Multivariate Synthesis of Published Platonic Stylistic Data. In: *Literary and Linguistic Computing* 11, 67–75.
- Tenenbaum, J./de Silva, V./Langford, J. (2000), A Global Geometric Framework for Nonlinear Dimensionality Reduction. In: *Science* 290, 2319–2323.
- Tinsley, H./Brown, S. (2000), *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. New York: Academic Press.
- van Rijsbergen, C. (1979), *Information Retrieval*. 2nd ed. London: Butterworths.
- Verleysen, M. (2003), Learning High-dimensional Data. In: Ablameyko, S./Goras, L./Gori, M./Piuri, V. (eds.), *Limitations and Future Trends in Neural Computation*. Amsterdam: IOS Press, 141–162.
- Verleysen, M./François, D./Simon, G./Wertz, V. (2003), On the Effects of Dimensionality on Data Analysis with Neural Networks. In: Mira, J. (ed.), *International Work-conference on Artificial and Natural Neural Networks*. Mao, Menorca (Spain), June 3–6, 2003, 105–112.
- Vesanto, J./Alhoniemi, E. (2000), Clustering of the Self-Organizing Map. In: *IEEE Transactions on Neural Networks* 11(3), 586–600.
- Watters, P. (2002), Discriminating English Word Senses Using Cluster Analysis. In: *Journal of Quantitative Linguistics* 9, 77–86.
- Webb, A. (2002), *Statistical Pattern Recognition*. 2nd ed. New Jersey: John Wiley & Sons.
- Willett, P. (1988), Recent Trends in Hierarchic Document Clustering: A Critical Review. In: *Information Processing and Management* 24(5), 577–597.
- Woods, A./Fletcher, P./Hughes, A. (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Yarowsky, D. (2000), Word-sense Disambiguation. In: Dale/Moisl/Sommers 2000, 629–654.
- Zernik, U. (1991), Train 1 vs Train 2: Tagging Word Sense in a Corpus. In: Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale NJ: Lawrence Erlbaum Associates, 97–112.

Hermann Moisl, Newcastle (UK)

41. Corpus linguistics in morphology: Morphological productivity

1. Morphological productivity
2. Theoretical frameworks
3. Measuring productivity
4. Forces shaping productivity
5. Concluding remarks
6. Literature

- Singhal, A./Salton, G./Mitra, M./Buckley, C. (1996), Document Length Normalization. In: *Information Processing & Management* 32(5), 619–633.
- Spärck Jones, K. (1972), Exhaustivity and Specificity. In: *Journal of Documentation* 28, 11–21.
- Stevenson, M./Wilks, Y. (2003), Word-sense Disambiguation. In: Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 249–65.
- Strzalkowski, T. (1999), *Natural Language Information Retrieval*. Dordrecht: Kluwer.
- Tabachnick, B./Fidell, L. (2006), *Using Multivariate Statistics*. 5th ed. Boston: Allyn and Bacon.
- Tan, P./Steinbach, M./Kumar, V. (2006), *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.
- Temple, J. (1996), A Multivariate Synthesis of Published Platonic Stylistic Data. In: *Literary and Linguistic Computing* 11, 67–75.
- Tenenbaum, J./de Silva, V./Langford, J. (2000), A Global Geometric Framework for Nonlinear Dimensionality Reduction. In: *Science* 290, 2319–2323.
- Tinsley, H./Brown, S. (2000), *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. New York: Academic Press.
- van Rijsbergen, C. (1979), *Information Retrieval*. 2nd ed. London: Butterworths.
- Verleysen, M. (2003), Learning High-dimensional Data. In: Ablameyko, S./Goras, L./Gori, M./Piuri, V. (eds.), *Limitations and Future Trends in Neural Computation*. Amsterdam: IOS Press, 141–162.
- Verleysen, M./François, D./Simon, G./Wertz, V. (2003), On the Effects of Dimensionality on Data Analysis with Neural Networks. In: Mira, J. (ed.), *International Work-conference on Artificial and Natural Neural Networks*. Mao, Menorca (Spain), June 3–6, 2003, 105–112.
- Vesanto, J./Alhoniemi, E. (2000), Clustering of the Self-Organizing Map. In: *IEEE Transactions on Neural Networks* 11(3), 586–600.
- Watters, P. (2002), Discriminating English Word Senses Using Cluster Analysis. In: *Journal of Quantitative Linguistics* 9, 77–86.
- Webb, A. (2002), *Statistical Pattern Recognition*. 2nd ed. New Jersey: John Wiley & Sons.
- Willett, P. (1988), Recent Trends in Hierarchic Document Clustering: A Critical Review. In: *Information Processing and Management* 24(5), 577–597.
- Woods, A./Fletcher, P./Hughes, A. (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Yarowsky, D. (2000), Word-sense Disambiguation. In: Dale/Moisl/Sommers 2000, 629–654.
- Zernik, U. (1991), Train 1 vs Train 2: Tagging Word Sense in a Corpus. In: Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale NJ: Lawrence Erlbaum Associates, 97–112.

Hermann Moisl, Newcastle (UK)

41. Corpus linguistics in morphology: Morphological productivity

1. Morphological productivity
2. Theoretical frameworks
3. Measuring productivity
4. Forces shaping productivity
5. Concluding remarks
6. Literature

1. Morphological productivity

The vocabulary of English and most other languages contains many words that have internal structure (see also article 25). Words such as STRANGENESS, WEAKNESS, and SOFTNESS contain the formative NESS, which is usually found to the right of adjectives. Words ending in NESS are almost always abstract nouns. We refer to the sets of words sharing aspects of form structure and aspects of meaning as morphological categories.

Some morphological categories have a fixed or declining membership, while others have a growing membership. Categories with fixed or declining membership are said to be unproductive, categories with growing membership are described as productive.

Morphological categories differ tremendously in size. Some contain only a few words (e.g., the category of words in TH such as WARMTH and STRENGTH), others may have tens of thousands of members (e.g., nominal compounding). A large morphological category and its associated morphological rule, therefore, are also described as more productive than a small category and its rule. The importance of productivity for studies of the lexicon and lexical processing is witnessed by the fact that no finite lexicon will suffice to process unseen text: Computational tools cannot do without taking productive word formation into account.

A first key question in productivity research is what conditions need to be met for a rule to be productive in these ways. A second key question is whether a rule is ever totally unproductive, i.e., whether productivity is in essence a graded phenomenon. A related issue is how the degree of productivity of a morphological category might be measured. A third set of questions addresses how productivity changes through time, and how affixes are used across registers by different social groups. A final issue is the relation between productivity and processing constraints in the mental lexicon.

How these questions are answered, and the relevance of corpora for guiding research, however, depends crucially on what is viewed as the goal of morphological theory.

2. Theoretical frameworks

The goal of morphological theory as defined in early generative morphology is to account for what complex words are possible words. In this approach, productive and unproductive rules have very different properties. First, productive rules are seen as the true rules, they are dynamic, and crucial for producing and understanding the associated complex words. Unproductive rules, by contrast, are viewed as only describing the structure of complex words that have been committed to memory. Unproductive rules have been described as redundancy rules (Jackendoff 1975), redundant in the sense that they describe (and account for) existing structure but do not play an active role in production or comprehension. Second, regularity is taken to be a necessary condition for productivity, but not a sufficient condition – there are unproductive morphological categories with fully regular (redundancy) rules (Dressler 2003). Third, morphological rules are viewed as part of an internally consistent module of the grammar that characterizes the knowledge of an ideal speaker in a homogeneous speech community. This knowledge of the ideal speaker is assumed to be an adequate characterization of the knowledge of

actual speakers. By implication, what real speakers actually say must provide an imperfect window on their true morphological competence, a window that is distorted by intervening pragmatic, sociolinguistic and stylistic variables. Consequently, corpora are largely irrelevant. In the words of Dressler (2003, 54), “statistic approaches [...] are of little relevance in itself, because they refer to language norm and to individual performances. In fact, all corpora data are performance data which reflect the realisation of linguistic norms and thus only indirectly the realisation of the corpus producers’ competence of the system of potentialities.”

A series of findings challenges this classical view of morphology and productivity. The central role of gradedness in morphology that has emerged from a wide range of recent studies (see Hay/Baayen 2005, for a review) casts doubt on the usefulness of an absolute distinction between productive and unproductive rules. The productivity of probabilistic paradigmatic morphology invalidates the assumption that productivity crucially depends on regularity as described by straightforward syntagmatic rules, and reinforces the importance of schemas, constructions, and local generalizations (Albright/Hayes 2003; Baayen 2003; Bybee 2001; Dabrowska 2004). In this emerging new theory, morphological productivity can be understood as resulting from a great many factors such as the individual language user’s experience with the words of her language, her phenomenal memory capacities, her conversational skills, her command of the stylistic registers available in her language community, her knowledge of other languages, her communicative needs, her personal language habits and those of the people with which she interacts.

There are many ways in which corpora contribute to making progress towards the highly ambitious goal of understanding morphological productivity in its full complexity. Corpora allow researchers to explore how productivity varies across registers, written versus spoken language, social and geographical space, and even time. Corpus-derived measures play an increasingly important role in research on lexical processing in the mental lexicon, and have proved essential for developing rigorous and falsifiable models for processing constraints on productivity. A first step in this direction was provided by the development of statistical measures of productivity.

3. Measuring productivity

Several corpus-based measures are now available for gauging different aspects of productivity (Baayen 1992, 1993; Baayen/Renouf 1996).

3.1. Mathematical formalizations of productivity

A first measure of productivity focuses on the size of the morphological category. A category with many members is more productive in the sense that it has produced many complex words that are useful to the language community. A rule that is highly productive in this sense is like a successful company selling a product that has a large share of the market. Such a rule has a high REALIZED PRODUCTIVITY. Realized productivity is similar to profitability in the sense of Corbin (1987; see also Bauer 2001,

49), but restricted to ‘past achievement’. In Baayen (1993), it is referred to as extent of use. The realized productivity of a morphological category C is estimated by the type count $V(C, N)$ of its members in a corpus with N tokens.

A second measure of productivity assesses the rate at which a morphological category is expanding and attracting new members. A category that is expanding at a higher rate is more productive than a category that is expanding at a lower rate, or that is not expanding at all. A rule that is highly productive in this sense is like a company that is expanding on the market (independently of whether that company has or does not have a large share of the market). Such a rule has a high EXPANDING PRODUCTIVITY. Expanding productivity is similar to Corbin’s profitability, but oriented to what is expected for the near future. This aspect of productivity is estimated by means of the number of words $V(1, C, N)$ in morphological category C that occur only once in a corpus of N tokens, the hapax legomena. Let $V(1, N)$ denote the total number of hapax legomena in the corpus. The ratio $P^* = V(1, C, N)/V(1, N)$ is an estimate of the contribution of morphological category C to the growth rate of the total vocabulary. This measure is referred to as the hapax-conditioned degree of productivity (Baayen 1993).

A company may have a large share of the market, but if there are hardly any prospective buyers left because the market is saturated, it is nevertheless in danger of going out of business. A third measure of productivity gauges the extent to which the market for a category is saturated. A rule with a low risk of saturation has greater potential for expansion, and hence a greater POTENTIAL PRODUCTIVITY. The potential productivity of a rule is estimated by its hapax legomena in the corpus divided by the total number of its tokens $N(C)$ in the corpus: $P = V(1, C, N)/N(C)$. This ratio, known as the category-conditioned degree of productivity (Baayen 1993), estimates the growth rate of the vocabulary of the morphological category itself.

The horizontal axis displays corpus size in tokens (N), the vertical axis displays the number of types observed as the corpus size is increased. Solid lines represent two growth curves. The extent of use is the highest point of a curve. Potential productivity is defined as the slope of the tangent to the curve at its endpoint (dashed line). The dotted line illustrates that potential productivity depends on N, and decreases with increasing N.

All three measures are defined with respect to the statistical properties of word frequency distributions (Baayen 2001). A corpus providing a synchronic sample of the language can be viewed as a text that is to be processed from beginning to end. For each successive word token read, we note the total number of different word types observed thus far. When the number of types is plotted against the number of tokens, curves such as shown in Figure 41.1 are obtained. The measure for potential productivity (P) represents the rate at which the vocabulary is increasing at the end of the curve. It is mathematically equivalent to the slope of the tangent to the curve at its endpoint.

A corpus can also be viewed as a (simplified) model of diachrony, as through life, new samples of text are continuously added to one’s cumulative experience. In this case, the statistical theory of vocabulary growth curves presents the simplest possible model of how past experience combines with expectations for the near future. However, since most corpora present synchronic slices of adult language use, they are not well suited for studying diachronic change, not for the individual speaker, nor for communities of speakers. In psycholinguistics, the absence of diachronic corpora representing language input from birth to old age is acutely felt, and has led to experimental measures gauging age of acquisition, beginning with the study by Carroll/White (1973). Estimates of age

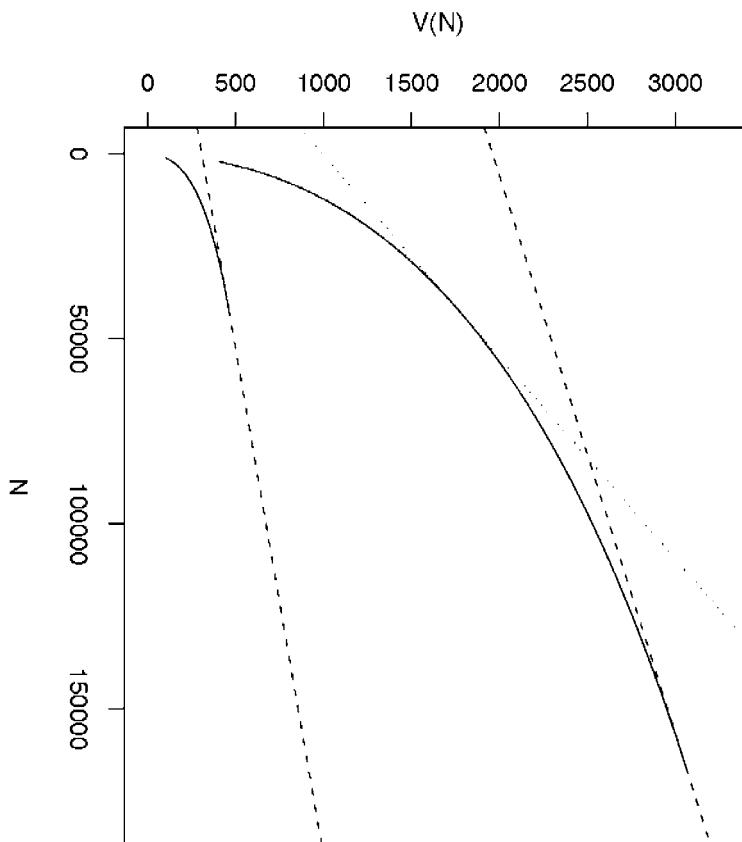


Fig. 41.1: The dynamics of vocabulary growth

of acquisition often have superior predictivity for lexical processing compared to synchronic frequency counts, and suggest that corpora sampling speech from different stages will be important resources for (corpus) linguistics as well.

The growth rate as measured with the P statistic is based on probability theory. An alternative computational method for estimating the growth rate is based on the deleted estimation method of Jelinek/Mercer (1985). Nishimoto (2003) shows that similar productivity rankings are obtained when potential productivity is estimated with this technique.

All these productivity measures are based on a partition of the vocabulary into morphological categories. As a consequence, only those words in which a given rule was the last to apply are taken into account. For instance, HELPFULNESS is counted only once, namely, as a member of the morphological category of words in NESS. It is not counted as a member of the category of words in FUL. If one were to assign HELPFULNESS to both morphological categories, the observations in the two categories would no longer be independent, and the statistical tests for comparing degrees of productivity would no longer be valid. On the other hand, it certainly does make sense to consider HELPFULNESS as instantiating a particular use of a member of the morphological

category of FUL. When lack of statistical independence is not a problem, for instance, when the focus of interest is on ranking affixes by their degree of productivity, words with a given affix that fall outside of the corresponding morphological category proper can be taken into account as well. Gaeta/Ricca (2005) have shown that similar rankings are obtained for counts excluding and counts including words that have undergone further word formation. This suggests that measures that are based strictly on the morphological category itself provide a good approximation that has, on the one hand, the advantage of ease of extraction from (unanalyzed) corpora, and on the other hand, the advantage of allowing further statistical testing.

3.2. Interpretation and validation

These three productivity measures are statistical formalizations of the intuitive notion of productivity. To what extent are these measures linguistically interpretable, and to what extent have they been validated?

3.2.1. Realized productivity

First consider the simple type count that provides a corpus-based estimate of realized productivity. According to Bybee (2001), the productivity of a word formation schema is largely determined by its type frequency. For instance, the English past tense in ED is realized on thousands of verbs, whereas irregular schemas such as that exemplified by KEEP/KEPT and SLEEP/SLEPT, pertain to only small numbers of verbs. The regular past tense schema has a much greater realized productivity than the irregular past tense schemata. The importance of type frequency emerges even more clearly when it is pitted against token frequency. Productive categories are characterized by the presence of large numbers of low-frequency forms, whereas unproductive categories tend to contain many high-frequency forms, unsurprisingly, as a high token frequency protects irregular forms against regularization and helps explain their continued existence. Assessing the productivity of a schema in terms of token frequency would therefore be counterproductive. For instance, Baayen/Moscoso del Prado Martín (2005) studied 1600 monomorphemic English verbs, of which 146 were irregular and 1454 regular. The summed frequency of all irregular verbs, 1793949, exceeds the summed frequency of the much larger set of regular verbs (732552) by a factor 2.5 (counts based on Baayen/Piepenbrock/Gulikers 2005, which is based on a corpus of 18.5 million tokens). An assessment of the productivity of the past tense in English using token frequency would suggest that vocalic alternation would be more productive than suffixation with ED, contrary to fact.

Type frequency, however, provides only a first approximation of the productivity of a schema. First of all, it does not take into account the degrees of similarity between the words that are governed by that schema. Much improved estimates of the productivity of a particular schema are obtained with analogical models (e.g., Skousen 1989; Albright/Hayes 2003). Second, type-based counts do not do justice to the lower weight of low-frequency words that speakers may not know well or not know at all. In this respect, analogical models also offer further precision (see Skousen 1989, for token-weighted

analogy). Third, type-based counts work reasonably well in a fixed domain like past tense inflection in English, where morphological alternatives carrying the same function are compared. Type-based counts fare less well, however, when very different morphological categories are compared. For instance, Dutch has several suffixes for creating nouns denoting female agents. The most productive of these suffixes is STER, as in VERPLEEG-STER, ‘female nurse’ (compare VERPLEEG-ER, ‘male nurse’). Dutch also has a verb-forming prefix VER (as in VER-PLEEG-EN, ‘to nurse’), that is described as less productive. Even though STER is judged intuitively to be more productive than VER, the type frequency of VER (985) is much higher than the type frequency of STER (370) in counts based on a corpus of 42 million words. This shows that there are aspects of productivity that are not well-represented by a category’s realized productivity. A morphological category may have a high realized productivity, but a high realized productivity does not imply that its expanding productivity or its potential productivity will be high as well.

3.2.2. Expanding productivity

The relative rate at which a category is expanding provides a first complement to the type-based estimate of productivity. This measure generates productivity rankings that provide reasonable reflections of linguists’ overall intuitions about degrees of productivity (e.g., Baayen 1993). Recall that the degree to which a category is expanding can be estimated by considering that category’s contribution to the growth rate of the vocabulary in a corpus. In practise, differences in expanding productivity can be gauged simply by comparing counts of hapax legomina.

Gaeta/Ricca (2005) propose an alternative measure that is also based on the count of hapax legomina. They argue that counts of hapax legomina should be compared when equal numbers of tokens have been sampled for each of the morphological categories involved. With reference to Figure 41.1, they propose to measure the growth rate for the endpoint of the short curve and for the same number of tokens for the upper curve. This amounts to comparing the slopes of the lower dashed line and the upper dotted line. It turns out that their ‘variable corpus’ productivity measure is mathematically and empirically closely related to the present measure for expanding productivity. Both lead to plausible productivity rankings of derivational categories in Italian.

Hapax legomina should not be confused with neologisms. An ideal lexicographic measure of expanding productivity would specify the rate at which neologisms with a given morphological structure are added to the vocabulary of the language community, calibrated with respect to the rate at which the community’s vocabulary as a whole is expanding. For the problems involved with approaches based on neologisms, see section 4.1.2. Corpus-based counts of hapax legomina provide an indirect way of estimating the rate at which a morphological category enriches the vocabulary. For small corpora, the number of neologisms among the hapax legomina will be small. As the corpus size increases, the number of neologisms in the corpus increases as well. Crucially, these neologisms are found primarily among the hapax legomina, and to a lesser extent among the words occurring twice, three times, etc. (Baayen/Renouf 1996; Plag 2003). Nevertheless, even in a large corpus, there may be words among the hapax legomina that have been well-established in the language for centuries. This is not a problem as long as it is

kept in mind that the hapax legomena are not a goal in itself, they only function as a tool for a statistical estimation method aimed at gauging the rate of expansion of morphological categories.

The degree to which a morphological category is expanding captures an important aspect of what it is to be productive. But there is a further aspect of productivity that is not properly captured by the count of types nor by the count of hapaxes. Recall that the Dutch suffix STER is in some sense more productive than the Dutch prefix VER. A type-based comparison failed to bring this difference to light, and a count of hapax legomena (274 for VER, 161 for STER) fails to do so as well. The problem with these two counts is that they do not do justice to the intuition that it is easy to think of new well-formed words in STER, but very hard to think of well-formed neologisms in VER.

3.2.3. Potential productivity

The ratio of hapax legomena in a given morphological category to the total number of tokens in that category, the category's potential productivity, assigns VER a productivity index of 0.001 and STER a much higher index of 0.031. This measure for a category's potential productivity indicates correctly that it is easier to think of a neologism in STER than of a neologism in VER.

The potential productivity measure is highly sensitive to markedness relations. The unmarked suffix for creating agent nouns in Dutch is ER (GEEF-ER, 'giver'), its marked counterpart is STER (GEEF-STER, 'female giver'). Unmarked ER has the greater realized productivity as well as the greater expanding productivity, but marked STER has the greater potential productivity.

What the potential productivity measure highlights is that productivity can be a self-defeating process, in the sense that once an affix has saturated the onomasiological market, it has no potential for further expansion. Unmarked ER has saturated its market to a much greater extent than has STER. As a consequence, STER can freely attach to a great many verbs where it has not been used before, due to the reluctance of Dutch speakers to explicitly mark the gender of agents.

An experimental study validating the potential productivity measure is Baayen (1994b). Following Anshen/Aronoff (1988), subjects were asked to generate within 5 minutes as many words with a specified affix as they could think of. Exactly as predicted, subjects produced many more neologisms in STER than in ER or VER. Further validation of this measure has been provided by Hay (2003), who showed that it is correlated with measures for the parseability of the complex words in the morphological category (see section 4.2.2.). Furthermore, Wurm/Aycock/Baayen (manuscript submitted for publication) observed that the potential productivity measure is predictive for visual lexical decision latencies.

The potential productivity measure is also sensitive to the compositionality of the words in the morphological category, albeit indirectly. Words with less compositional meanings typically tend to be high-frequency words. Since the token frequencies of the words in the morphological category contribute to the denominator of the potential productivity measure, the presence of opaque words will tend to lead to lower estimates of potential productivity.

A closely related measure for potential productivity is the ratio I of the estimated size of the category S in an infinitely large corpus and the observed number of types in a corpus of size N : $I = S/V(N)$. This ratio quantifies the extent to which the attested types exhaust the possible types. Rough estimates of S are provided by statistical models for word frequency distributions, such as the finite Zipf-Mandelbrot model developed in Evert (2004). Affixes with a high potential productivity also tend to have a high I (see, e.g., Baayen, 1994b), indicating that many more types could be formed than are actually attested in the corpus.

The validity of these productivity measures hinges on the availability of correct input data. Generally, string-based searches of affixes in corpora produce highly polluted word frequency distributions that may seriously distort the true pattern in the data. Manual inspection and correction, although time-consuming, is as crucial for productivity research (Evert/Lüdeling 2001) as it is for research on syntax (see articles 42 and 43).

4. Forces shaping productivity

Traditional approaches to morphological productivity have invested in finding structural explanations for degrees of productivity. One influential view originating from early structuralism (see Schultink 1961) is that the degree of productivity is inversely proportional to the number of grammatical restrictions on that rule. However, it is difficult to see how the quantitative effects of the very different kinds of structural constraints would have to be weighted. As pointed out by Bauer (2001, 143), “words are only formed as and when there is a need for them, and such a need cannot be reduced to formal terms”.

If structural constraints as such do not directly drive morphological productivity (see section 4.2.3. for indirect effects, however), research on morphological productivity should be directed towards other factors. There are two clusters of such factors that play a demonstrable role, one cluster pertaining to societal factors, the other to the role of processing constraints in the mental lexicons of individual speakers.

4.1. Productivity in the speech community

It is well-known that there is consistent variation in how speakers with different backgrounds and varying communicative goals make use of the morphological and grammatical constructions offered by their language. Biber (1988), for instance, provided detailed evidence that the linguistic resources employed in speech differ from those in writing, and that within each of these communicative modalities, further systematic differences differentiate the styles of more specific text types. Contemporary work in stylometry (e.g., Burrows 1992) showed, furthermore, that individual writers develop their own characteristic speech habits, not only in the selection of their topics, but, crucially, in which grammatical resources offered by the language authors typically tend to use (see also articles 38 and 50).

This work in corpus linguistics has had little impact on productivity research in theoretical morphology. Bauer’s monograph (2001) reveals no awareness of the possibility that morphological categories might be more productive in some registers than in others,

and the potential consequences of such stylistic forces for the weight of structural constraints in explanations of productivity. However, the little work that has been done in this area shows unambiguously that, unsurprisingly, different genres recruit different morphological categories to very different degrees.

A further complication in productivity research is that the needs of speech communities and groups of specialists within these speech communities (Clark 1998) change over time. In modern technological societies, the ever increasing rate of scientific and technological progress leads to a proliferation of new techniques, concepts and products that require names. How productive affixes are used, and the rate at which new words appear through the years will depend on whether discourse is studied from a domain with rapid innovation or with slow or little innovation. In what follows, we first consider productivity in relation to register variation. Next, we consider productivity from the perspective of the society and its changing needs.

4.1.1. Register and productivity

Plag/Dalton-Puffer/Baayen (1999) showed, using the British National Corpus, that the degree of productivity of a suffix may differ depending on whether it is used in written language, formal spoken language, or informal spoken language. Most derivational suffixes were observed to be more productive in written than in spoken language, with the exception of WISE. Furthermore, the productivity rankings of affixes changed from one register to the other. For instance, NESS emerged as more productive than ABLE in written language, but in spontaneous conversations, ABLE was slightly more productive than NESS. The different degrees of productivity observed for spoken and written language are expected, given the very different conditions under which oral and written language are produced. Oral language tends to be produced on the fly, written language tends to go through several rounds of revision before it appears in print. Furthermore, oral language is anchored in the physical context, where prosody, gesture, gaze, and common ground provide very different constraints on communication compared to written language, where sentence and discourse structure have to bear the full burden of communication. Thus, the greater productivity of NESS in written discourse observed by Plag/Dalton-Puffer/Baayen (1999) may be due to the possibility to use this suffix to refer to states of affairs previously introduced into the discourse (Kastovsky 1986; Baayen/Neijt 1997).

Different registers tend to be used for communication about very different kinds of topics. The suffix ITY, for instance, is more productive in scientific and technical discourse, and the suffix ITIS appears predominantly in medical discourse and occasionally in non-medical texts in certain journalistic registers (Lüdeling/Evert 2005, see also Clark 1998). A study addressing differences in potential productivity across various registers of written English using methods from stylometry is Baayen (1994a). Figure 41.2 visualizes the correlational structure for selected affixes and texts using principal components analysis. Germanic affixes are found predominantly in the right half of the plot, Latinate affixes occur more to the left. The texts that have a preference for the Germanic affixes are, for instance, the stories for children by Lewis Carroll and Frank Baum. The Latinate affixes, by contrast, are most productive in officialese, Startrek novels, and the scientific prose of William James. Register variation challenges theoretical approaches

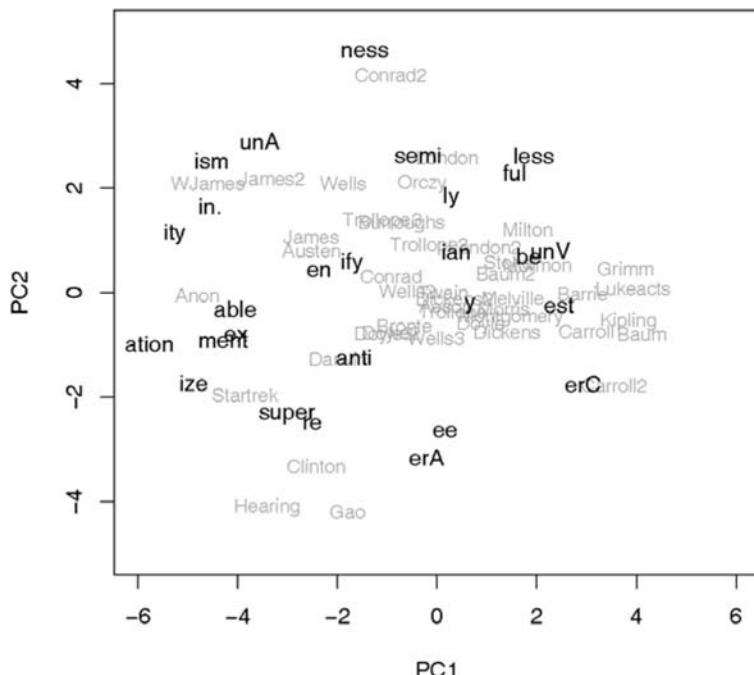


Fig. 41.2: Selected affixes in the space spanned by the first two dimensions resulting from a principal components analysis of the correlational structure of their potential productivity in four different kinds of texts (stories for children, officialese, literary texts, and religious texts). Individual texts are shown in grey (after Baayen 1994a)

that seek to ground the productivity of morphological categories purely in structural constraints, since these constraints do not vary with register. Register variation was also observed by Baayen/Neijt (1997) for the Dutch suffix HEID, the translation equivalent of English NESS. They studied a newspaper corpus, and found that HEID was not productive at all the articles on economics, and most productive in the sections on literature and art.

4.1.2. Productivity through time

Languages change as time passes by. New morphological categories may come into existence and established categories may fade away, while other categories always remain peripheral, drifting along the tides of fashion (e.g., non-medical ITIS, Lüdeling/Evert 2005, or COD and FAUX in British English, Renouf/Baayen 1997).

Traditionally, the diachronic aspects of productivity have been studied by means of dictionaries. Especially the *Oxford English Dictionary* (OED) has proved to be a useful source of information, as it provides dates for first and last mentions (Neuhaus 1973; Anshen/Aronoff 1999; Bolozky 1999). However, the use of dictionaries brings along several methodological problems. Especially for the older stages of the language, the sam-

pling is – unavoidably – sparse. As a consequence, a word may have been in use long before it is first observed in the historical record. Conversely, for recent developments, the sheer volume of text available both in print and on the internet prohibits exhaustive description. Furthermore, dictionaries provide little control over variation in productivity due to register. Finally, words with a new onomasiological function are of lexicographical interest and are relatively easy to detect, while words with referential functions (in the sense of Kastovsky 1986) tend to go unnoticed.

Diachronic studies of language change, however, need not be based on dictionaries (see article 4). Ellegråd (1953), for instance, is a classic example of a corpus-based study long before electronic corpora were available. The diachronic study of German nominal ER by Meibauer/Guttropf/Scherer (2004) is similar in spirit. They carefully extracted samples from four centuries of German newspapers, which were digitized and analyzed semi-automatically. This procedure is exemplary, in that it controls to a large extent for register variation.

In general, the use of corpora sampling texts at different points in time is not without its own share of methodological pitfalls, unfortunately. Data sparseness remains a problem, especially for the older stages of the language. Furthermore, any imbalance in the materials included invariably leads to diachronic artefacts, such as the sudden increase in the use of medical ITIS in the first decade of the 20th century observed by Lüdeling/Evert (2005), which they were able to trace to the inclusion of an encyclopedia published in 1906.

What is clear from diachronic studies of productivity is that probabilistic models that assume a fixed population of possible words from which successively more words are sampled as time proceeds are fundamentally flawed. Studies based on the OED show bursts of productivity around 1600 and around 1850 (Neuhaus 1973; Anshen/Aronoff 1999), but changes may even be going on at much shorter time scales of just a few years (Baayen/Renouf 1996; see also article 52 for corpora in the study of recent language change).

The study of Meibauer/Guttropf/Scherer (2004) offers an interesting perspective on the increasing realized and potential productivity of nouns in ER in German newspapers. Through time, ER is attached more often to complex base words, it is increasingly more productive as a deverbal suffix, and its primary function has become to denote persons. Meibauer and colleagues point out that a similar development characterizes the micro-level of children acquiring German, and they offer several explanations of why this similarity might arise. Possibly, the linguistic development reflects, at least in part, the expanding cognitive, social and intellectual skills of both the child and of the increasingly complex society of speakers of German as it developed over 400 years.

The historical record also reveals that morphological categories may cease to be productive. Anshen/Aronoff (1997, 1999) discuss English OF and AT (which dropped out of use within a short period of time) and MENT (which showed a much more gradual decline). Keune et al. (2005) provide further detail on the decline of productivity using a corpus of spoken Dutch. They show how loss of productivity is reflected in the reduction of the acoustic realizations of the highest-frequency members of the category. Words such as NATUUR-LIJK (literally ‘nature-like’, but with the opaque meaning ‘of course’) can be reduced to TUUK in spontaneous conversations, which shows these words are in the process of becoming monomorphemic.

4.2. Productivity and processing constraints in the mental lexicon

Productivity is subject not only to societal forces, but also to cognitive constraints governing lexical processing in the mental lexicon of the individual. Corpus-based surveys of actual use and corpus-based estimates of a wide range of lexical, sublexical, and supralexical probabilities play a crucial role in psycholinguistic research on productivity.

4.2.1. Productive and unproductive: An absolute distinction?

Many researchers view productivity as a diagnostic that can be used “to determine which patterns are fossilized, and which represent viable schemas accessible to speakers” (Bybee 2001, 13). The implication is that productive morphology would be in some sense cognitively more real than unproductive morphology, and that it would make sense to make a principled distinction between being productive (to a greater or lesser degree) and being totally unproductive. Recent results in mental lexicon research argue against such an absolute split between live rules and fossilized residues.

First, it has become clear that complex words leave traces in lexical memory, irrespective of whether they are regular or irregular (see Hay/Baayen 2005). The frequency with which a complex word is used may even co-determine the fine acoustic detail of its constituents (Pluymaekers/Ernestus/Baayen, to appear). Consequently, a distinction between totally unproductive rules or schemas on the one hand, and productive rules or schemas on the other hand, would imply that the stored exemplars of unproductive schemas would not be available for generalization, while the stored exemplars of productive rules would allow generalization. This is not only implausible, but also contradicts the well-documented finding that the schemas of irregular verb classes can serve as attractors in both synchrony and diachrony (Bybee/Slobin 1982).

Second, the same kind of gang effects that characterize the unproductive schemas for the irregular past tense in English have been shown to be active for the regular past tense in ED as well (Albright/Hayes 2003). The strength of such gang effects, and not regularity or default status as such, has also been shown to predict the productivity of case inflection in Polish (Dabrowska 2004). Given that exactly the same analogical mechanisms underly both the irregular and the regular past tense, the difference between the unproductive irregular and the productive regular forms is a matter of degree and not a matter of fundamentally different cognitive principles (see also Pothos 2005, but Pinker 1997; Anshen/Aronoff 1999; Dressler 2003, for the opposite position).

It might be argued that the ‘semi-productivity’ of certain irregular inflections contrasts with the complete lack of productivity for a derivational suffix like TH in English WARMTH and STRENGTH, which is the textbook example of an unproductive suffix (e.g., Bauer 2001, 206). Nevertheless, new words in TH are occasionally used, as illustrated by the following text: “The combination of high-altitude and low-latitude gives Harare high diurnal temperature swings (hot days and cool nights). The team developed a strategy to capture night-time coolth and store it for release the following day. This is achieved by blowing night air over thermal mass stored below the verandah’s ...” (<http://www.arup.com/insite/features/printpages/harare.htm>, observed in 2001). When considered out of context, COOLTH seems odd, jocular, perhaps literary or even pretentious. Yet inspection of how it is actually used in context reveals that COOLTH fills an ono-

masiological niche by supplying a word denoting a quantity of the physical property referenced by the adjective COOL that fits with the existing words in TH that express similar quantities (WARMTH, STRENGTH, LENGTH, and WIDTH). COOLTH is a possible word of English, but it is at the same time a very low-probability word of English, not because speakers of English are unaware of the structural similarity of WARMTH, STRENGTH, LENGTH, and WIDTH, but because the probability that TH can be used to satisfy a sensible onomasiological need is extremely low. For a similar example from Dutch, see Keune et al. (2005).

4.2.2. Processing constraints

We have seen that productivity is co-determined by register as well as by the onomasiological needs in a language community, and that the same cognitive principles underlie both unproductive and productive rules: exemplar-driven analogical generalization. Productivity is further restricted by processing constraints.

Recall the prominence of the count of hapax legomena in the measures for expanding and potential productivity. From a processing perspective, the hapax legomena represent the formations with the weakest traces in lexical memory. Consequently, it is for these words that comprehension and production is most likely to benefit from rule-driven processes. Conversely, when a morphological category comprises predominantly high-frequency words, strong memory traces for these words exist that decrease the functional load for production and comprehension through rule-driven processes. Hence, the importance of rules for the lexical processing of complex words will be larger for morphological categories with many low-frequency words.

There are two ways in which the consequences of lexical processing for productivity can be made more precise, as shown by Hay (2003).

First, the frequency of the base relative to that of the derived word should be taken into account. Many complex words have a lower frequency than their base word (e.g., ILLIBERAL). But there are also complex words that are more frequent than their bases (e.g., ILLEGIBLE). The greater the frequency of the complex word compared to that of its base, the greater the likelihood will be that its own memory trace will play a role during lexical access. Conversely, rule-driven processing will be more important for formations with memory traces that are much weaker than those of their constituents.

Effects of relative frequency (defined here as the frequency of the base divided by the frequency of the derivative) have been observed both in production and in comprehension. In English, t-deletion is more likely for words with a low relative frequency such as SWIFTLY (221/268) than for words with a high relative frequency such as SOFTLY (1464/440) (Hay 2001). SWIF(T)LY, in other words, is in the process of becoming independent of its base word, its simplified phonotactics indicate it is becoming more like a monomorphemic word in speech production.

In comprehension, relative frequency is an indicator of the relative importance of decompositional processing. As shown in Hay/Baayen (2002), morphological processing in reading can be understood in terms of lexical competition between the base and the derivative, such that the derivative (the whole) has a small headstart over the base (one of its parts). In their model, words with a high relative frequency are accessed primarily

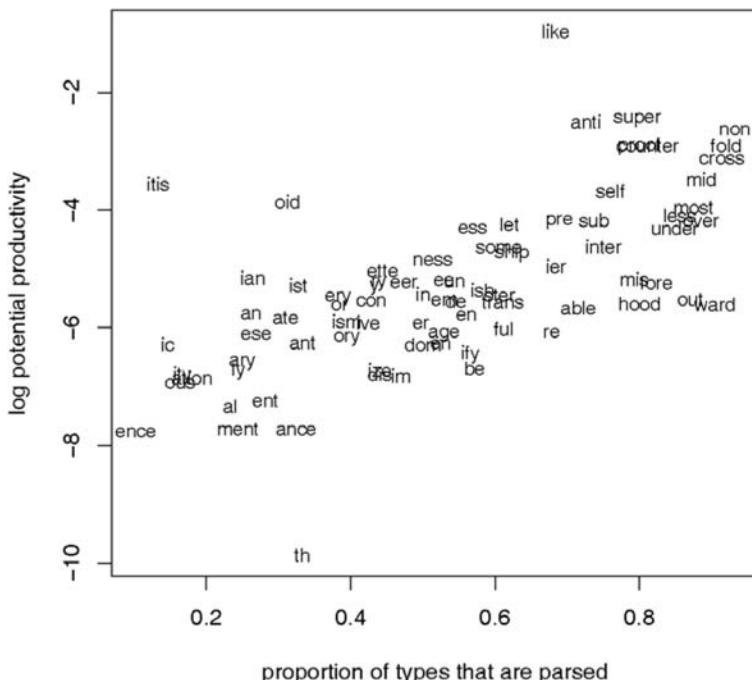


Fig. 41.3: Log potential productivity as a function of the proportion of types that are parsed in the model of Hay/Baayen (2002)

through their base, while words with a low relative frequency are accessed primarily on the basis of their own memory traces.

It turns out that relative frequency predicts potential productivity. Hay (2003) showed that the proportion of types in a morphological category for which the base frequency exceeds the frequency of the derivative is positively correlated with the logarithmic transform of its potential productivity (P). Hay/Baayen (2002) showed that estimates of the proportion of types that are parsed (according to a computational model of morphological processing in reading) likewise predict potential productivity (see Figure 41.3). The advantage of this modeling approach is that the perceptual advantage of the whole over its parts (see also Hay/Baayen 2005) as well as the frequency and length of the affix are taken into account. It is remarkable that 41 % of the variance in the (logged) P measure of 80 English morphological categories can be accounted for by relative frequency alone. One way of interpreting relative frequency as a predictor of productivity is that it gauges the experience with the rule in production and comprehension. A greater relative frequency implies stronger memory traces for the rule itself, and hence an intrinsically increased potential for producing and understanding new words.

A second important processing constraint concerns the derivative's junctural phonotactics, the probability of the sequence of sounds spanning the juncture between its parts. Low-probability sequences such as NH in INHUMANE create boundaries that are highly unlikely to occur within morphemes, and hence provide probabilistic information about morphological complexity. For speech production, low-probability sequences re-

quire more articulatory planning whereas high-probability sequences may benefit from automatized gestural scores. Hence, an affix is more likely to be an independent unit in speech production if it tends to create low-probability junctural phonotactics. In comprehension, the constituents are easier to parse out for words with low-probability junctures; for experimental evidence see Seidenberg (1987), Hay (2003) and Bertram/Pollatsek/Hyönä (2004). Given that complex words with low-probability junctures are easier to parse, affixes that create words with low-probability junctures should be more productive. Hay (2003) and Hay/Baayen (2003) observed just this: Several measures of junctural phonotactics (derived from the token frequencies of 11,383 English monomorphemic words in a corpus of 18 million words) correlated significantly with all three above-mentioned measures of productivity. For instance, the junctural probability averaged over all words in the morphological category explained some 14% of the variance in (log) potential productivity.

4.2.3. Conspiracies

Relative frequency and junctural phonotactics are involved in two correlational conspiracies.

The first conspiracy concerns the strong intercorrelations of all measures of productivity, measures for junctural phonotactics, measures for relative frequency and parsing, and lexical statistical measures such as Shannon's Entropy. Hay/Baayen (2003) observed, using principal components analysis, that this correlational structure has two orthogonal dimensions of variation. The first dimension represents the tight intercorrelations between measures that gauge how affixes are used against the backdrop of the corpus as anchor point for normalisation. Realized productivity, expanding productivity, entropy (information load), the count of formations with low-probability junctural phonotactics, and the estimates of the number of types parsed all enter into strong positive correlations. These measures quantify aspects of the past and present usefulness of an affix. This dimension of variation is probably most closely linked to its onomasiological usefulness in society, its referential functionality (Kastovsky 1986; Baayen/Neijt 1997), and register variation.

The second dimension unifies measures that are normalized with respect to the individual morphological categories. Potential productivity, the estimated proportion of types in the category that are parsed, and the frequency of the base (averaged over the types in the category) enter into strong positive correlations, and reveal strong negative correlations with the frequency of the derivative (averaged over the types in the category) and with the probability of the juncture (similarly averaged). This dimension gauges the strength of the rule in terms of the proportion of words in the corresponding morphological category that are accessed in comprehension and production primarily through that rule rather than through the memory traces of the derivatives themselves.

The second conspiracy involves processing constraints, grammatical constraints, and memory constraints.

Hay/Plag (2004) showed that English suffixes can be arranged in a hierarchy so that their rank in the hierarchy is predictive for the order in which these suffixes can occur in complex words. Given that a suffix has rank i , suffixes with rank greater than i may follow that suffix in a word, while suffixes with rank lower than i will never follow it. The position of a suffix in this hierarchy is predictable from measures gauging the strength of

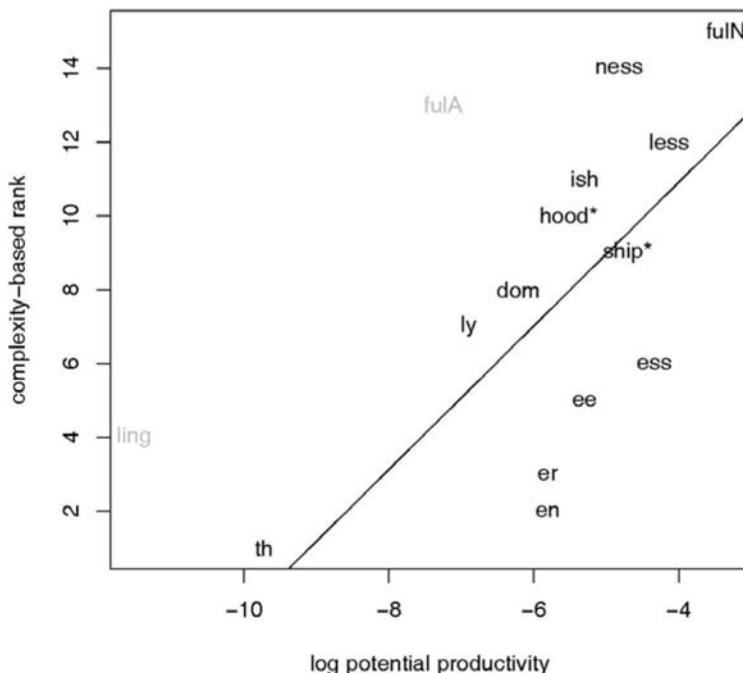


Fig. 41.4: Log potential productivity as predictor of complexity-based rank. The two words marked with an asterisk exchange rank in structure-based ordering. Affixes in grey were identified as outliers and were excluded when calculating the regression line (estimated slope: 1.95, R-squared: 0.47)

the suffix such as potential productivity and the proportion of types in the morphological category that are parsed (the second dimension identified above for the first conspiracy). This is illustrated in Figure 41.4 (based on Table III of Hay/Plag 2004) for potential productivity. As the log-transformed potential productivity increases, the word's rank increases as well. Hay (2003) argues that this hierarchy is driven by processing complexity: An affix that can be easily parsed out should not precede an affix that is more difficult to process. Hay/Plag (2004) refer to this as the hypothesis of complexity-based ordering. In other words, suffixes that are more productive, and that function more as processing units in speech production and comprehension, must follow less productive suffixes.

The hypothesis of complexity-based ordering raises two questions. The first question concerns the goal of word-formation: the creation of new words for communication. Is this goal completely subordinated to low-level processing constraints? This unlikely possibility has been ruled out by Hay/Plag (2004), who showed that grammatical restrictions (such as the required syntactic, semantic and phonological properties of the base) lead to a nearly identical hierarchy as that established on the basis of attested suffix combinations. This hierarchy can also be predicted from potential productivity and related processing measures. What seems to be at stake, then, is a conspiracy of on the one hand potential productivity and its correlated processing constraints, and on the

other hand grammatical constraints, all of which work in tandem to optimize the structure of multiply suffixed words for communication.

The second question prompted by the hypothesis of complexity-based ordering is why comprehension would benefit from less productive affixes preceding productive affixes. After all, there is no evidence that parsability constrains possible sequences of words in sentences in a similar way.

As a first step towards an answer, we note that this conspiracy of processing and grammatical factors is reminiscent of another probabilistic ordering hierarchy first discussed by Bybee (1985) for inflection. Bybee showed that more inherent inflection (tense and aspect marking, for instance) tends to be realized closer to the stem or root than more contextual inflection (person and number marking, for instance). In this inflectional hierarchy, therefore, the more (formally and semantically) predictable formatives are found to be peripheral to the less predictable and more fusional exponents, just as in derivation the less productive and often semantically less predictable suffixes are closer to the stem. What this suggests is that semantic transparency is also at issue. Especially the syntactic and semantic grammatical constraints studied by Hay/Plag (2004), which guarantee a minimal level of transparency, point in this direction.

A further step towards an answer can be made by a more careful consideration of the role of frequency in morphological processing. Hay/Baayen (2002), following Hay (2003), assume that relative frequency primarily affects low-level processes at the level of form. However, word frequency is more strongly correlated with measures of a word's meaning than with measures of a word's form (Baayen/Feldman/Schreuder in press), and it seems likely that a measure such as relative frequency (and derived measures such as parsing ratios) also reflects the relative complexities of compositional processes at the level of semantics. This may help explain why a conspiracy of processing constraints and grammatical constraints can exist.

Finally, it has been observed that productivity is inversely proportional to the likelihood of serving as input for further word formation. More specifically, Krott/Schreuder/Baayen (1999) reported that a greater potential productivity enters into a negative correlation with the proportion of the derivatives in the morphological category that serve as input to further word formation. Frequent, short words with less productive affixes are more likely to produce morphological offspring than more ephemeral complex words. This is exactly as expected from an onomasiological perspective, as well as from a processing perspective: More frequent words are more readily available in lexical memory as input for not only syntactic but also for morphological processing. This is, of course, the other side of the coin of complexity based ordering, but it extends beyond affix ordering to derivatives in compounds.

To conclude, potential productivity is part of a correlational conspiracy of different factors: low-level perceptual factors (as evidenced by junctural phonotactics and relative frequency), factors pertaining to morphological processing at the levels of form and meaning (as evidenced by relative frequency and selectional restrictions), and factors arising at the interface of memory and onomasiological needs.

5. Concluding remarks

What, then, is morphological productivity? Many theoretical morphologists have attempted to define productivity as a property of the language system (e.g., Schultink 1961; Bauer 2001; Dressler 2003). Unfortunately, these definitions and the underlying

theories have not led to models with predictive power for degrees of productivity. At the same time, traditional research has been dismissive of the potential relevance of system-external factors. Bauer's (2001, 211) definition of productivity is telling, in that it states that the extent to which a morphological category is actually used "may be subject unpredictably to extra-systemic factors". Contrary to what Bauer suggests, recent research has shown not only that the effects of 'extra-systemic' factors are truly predictive for productivity, but also that the 'intra-systemic' factors are part of a much larger system of interacting factors.

Exciting new insights have been obtained precisely by combining historical, stylistic, onomasiological, and cognitive factors in a quantitative and hence falsifiable empirical research paradigm. Without corpora, these insights would never have been obtained, prediction – the goal of scientific inquiry – would have remained out of reach, and productivity research would never have emerged from the quagmire of studies providing overviews and syntheses of previous studies based on idiosyncratic, small, and non-representative data. However, much still remains to be done, and even new fields await exploration, such as the role of sociolinguistic variables or the role of word formation in communal lexicons (Clark 1998). In short, in order to come to a full understanding of the challenging phenomenon of morphological productivity, a truly interdisciplinary data-driven research effort is required.

6. Literature

- Albright, Adam/Hayes, Bruce (2003), Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. In: *Cognition* 90, 119–161.
- Anshen, Frank/Aronoff, Mark (1981), Morphological Productivity and Morphological Transparency. In: *The Canadian Journal of Linguistics* 26, 63–72.
- Anshen, Frank/Aronoff, Mark (1988), Producing Morphologically Complex Words. In: *Linguistics* 26, 641–655.
- Anshen, Frank/Aronoff, Mark (1997), Morphology in Real Time. In: Booij, Geert E./van Marle, Jaap (eds.), *Yearbook of Morphology 1996*. Dordrecht: Kluwer Academic Publishers, 9–12.
- Anshen, Frank/Aronoff, Mark (1999), Using Dictionaries to Study the Mental Lexicon. In: *Brain and Language* 68, 16–26.
- Baayen, R. Harald (1992), Quantitative Aspects of Morphological Productivity. In: Booij, Geert E./van Marle, Jaap (eds.), *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers, 109–149.
- Baayen, R. Harald (1993), On Frequency, Transparency, and Productivity. In: Booij, Geert E./van Marle, Jaap (eds.), *Yearbook of Morphology 1992*. Dordrecht: Kluwer Academic Publishers, 181–208.
- Baayen, R. Harald (1994a), Derivational Productivity and Text Typology. In: *Journal of Quantitative Linguistics* 1, 16–34.
- Baayen, R. Harald (1994b), Productivity in Language Production. In: *Language and Cognitive Processes* 9, 447–469.
- Baayen, R. Harald (2001), *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. Harald (2003), Probabilistic Approaches to Morphology. In: Bod, Rens/Hay, Jennifer B./Jannedy, Stefanie (eds.), *Probability Theory in Linguistics*. Cambridge: The MIT Press, 229–287.
- Baayen, R. Harald/Feldman, Laura/Schreuder, Robert (in press), Morphological Influences on the Recognition of Monosyllabic Monomorphemic Words. In: *Journal of Memory and Language*.

- Baayen, R. Harald/Moscoso del Prado Martín, Fermín (2005), Semantic Density and Past-tense Formation in Three Germanic Languages. In: *Language* 81, 666–698.
- Baayen, R. Harald/Neijt, Anneke (1997), Productivity in Context: A Case Study of a Dutch Suffix. In: *Linguistics* 35, 565–587.
- Baayen, R. Harald/Piepenbrock, Richard/Gulikers, Leon (1995), *The CELEX Lexical Database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.
- Baayen, R. Harald/Renouf, Antoinette (1996), Chronicling The Times: *Productive Lexical Innovations in an English Newspaper*. In: *Language* 72, 69–96.
- Bauer, Laurie (2001), *Morphological Productivity*. Cambridge: Cambridge University Press.
- Bertram, Raymond/Pollatsek, Alexander/Hyönä, Jukka (2004), Morphological Parsing and the Use of Segmentation Cues in Reading Finnish Compounds. In: *Journal of Memory and Language* 51, 325–345.
- Biber, Douglas (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Bolozky, Shmuel (1999), *Measuring Productivity in Word Formation*. Leiden: Brill.
- Burrows, John F. (1992), Computers and the Study of Literature. In: Butler, C. S. (ed.), *Computers and Written Texts*. Oxford: Blackwell, 167–204.
- Bybee, Joan L. (1985), *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: Benjamins.
- Bybee, Joan L. (2001), *Phonology and Language Use*. Cambridge: Cambridge University Press.
- Bybee, Joan L./Slobin, Dan I. (1982), Rules and Schemas in the Development and Use of the English Past Tense. In: *Language* 58, 265–289.
- Carroll, John B./White, Margaret N. (1973), Age of Acquisition Norms for 220 Picturable Nouns. In: *Journal of Verbal Learning and Verbal Behavior* 12, 563–576.
- Clark, Herbert H. (1998), Communal Lexicons. In: Malmkjaer, K./Williams, J. (eds.), *Context in Language Learning and Language Understanding*. Cambridge: Cambridge University Press, 63–87.
- Corbin, Danielle (1987), *Morphologie derivationelle et structuration du lexique*. Tübingen: Niemeyer.
- Cowie, Claire (2003), Uncommon Terminations: Proscription and Morphological Productivity. In: *Italian Journal of Linguistics* 15, 99–130.
- Dabrowska, Ewa (2004), Rules or Schemas? Evidence from Polish. In: *Language and Cognitive Processes* 19, 225–271.
- Dalton-Puffer, Christiane/Cowie, Claire (2002), Diachronic Word-formation and Studying Changes in Productivity over Time. Theoretical and Methodological Considerations. In: Diaz Vera, Javier E. (ed.), *A Changing World of Words. Studies in the English Historical Lexicography, Lexicology and Semantics*. (Consterus New Series 141.) Amsterdam/New York: Rodopi, 410–437.
- Dressler, Wolfgang (2003), Degrees of Grammatical Productivity in Inflectional Morphology. In: *Italian Journal of Linguistics* 15, 31–62.
- Ellegård, Alvar (1953), *The Auxiliary Do: The Establishment and Regulation of its Use in English*. Stockholm: Almqvist & Wiksell.
- Evert, Stefan (2004), A Simple LNRE Model for Random Character Sequences. In: Purnelle, G./Fairon, C./Dister, A. (eds.), *Le Poids des Mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*. Louvain-la-Neuve: UCL, 411–422.
- Evert, Stefan/Lüdeling, Anke (2001), Measuring Morphological Productivity: Is Automatic Preprocessing Sufficient? In: Rayson, Paul/Wilson, Andrew/McEnery, Tony/Hardie, Andrew/Khoja, Shereen (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, 167–175.
- Gaeta, Davide/Ricca, Livio (2005), Productivity in Italian Word Formation: A Variable Corpus Approach. In: *Linguistics* 44, 57–89.
- Hay, Jennifer B. (2001), Lexical Frequency in Morphology: Is Everything Relative? In: *Linguistics* 39, 1041–1070.
- Hay, Jennifer B. (2003), *Causes and Consequences of Word Structure*. New York: Routledge.

- Hay, Jennifer B./Baayen, R. Harald (2002), Parsing and Productivity. In: Booij, Geert E./van Marle, Jaap (eds.), *Yearbook of Morphology 2001*. Dordrecht: Kluwer Academic Publishers, 203–235.
- Hay, Jennifer B./Baayen, R. Harald (2003), Phonotactics, Parsing and Productivity. In: *Italian Journal of Linguistics* 15, 99–130.
- Hay, Jennifer B./Baayen, R. Harald (2005), Shifting Paradigms: Gradient Structure in Morphology. In: *Trends in Cognitive Sciences* 9, 342–348.
- Hay, Jennifer B./Plag, Ingo (2004), What Constrains Possible Suffix Combinations? On the Interaction of Grammatical and Processing Restrictions in Derivational Morphology. In: *Natural Language and Linguistic Theory* 22, 565–596.
- Jackendoff, Ray S. (1975), Morphological and Semantic Regularities in the Lexicon. In: *Language* 51, 639–671.
- Jelinek, Frederick/Mercer, Robert L. (1985), Probability Distribution Estimation for Sparse Data. In: *IBM Technical Disclosure Bulletin* 28, 2591–2594.
- Kastovsky, Dieter (1986), The Problem of Productivity in Word Formation. In: *Linguistics* 24, 585–600.
- Keune, Karen/Ernestus, Mirjam/Van Hout, Roeland/Baayen, R. Harald (in press), Variation in Dutch: From Written MOGELIJK to Spoken MOK. In: *Corpus Linguistics and Linguistic Theory*.
- Krott, Andrea/Schreuder, Robert/Baayen, R. Harald (1999), Complex Words in Complex Words. In: *Linguistics* 37, 905–926.
- Lüdeling, Anke/Evert, Stefan (2005), The Emergence of Productive Non-medical *-itis*. Corpus Evidence and Qualitative Analysis. In: Kepser, Stephan/Reis, M. (eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin/New York: Mouton de Gruyter.
- Meibauer, Jörg/Guttropf, Anja/Scherer, Carmen (2004), Dynamic Aspects of German -er-Nominals: A Probe into the Interrelation of Language Change and Language Acquisition. In: *Linguistics* 42, 155–193.
- Neuhaus, Hans J. (1973), Zur Theorie der Produktivität von Wortbildungssystemen. In: Cate, A. P./Jordens, Peter (eds.), *Linguistische Perspektiven. Referate des VII Linguistischen Kolloquiums Nijmegen 1972*. Tübingen: Niemeyer, 305–317.
- Nishimoto, Eiji (2003), Measuring and Comparing the Productivity of Mandarin Chinese Suffixes. In: *Computational Linguistics and Chinese Language Processing* 8, 49–76.
- Pinker, Steven (1997), *Words and Rules: The Ingredients of Language*. London: Weidenfeld and Nicolson.
- Plag, Ingo (2003), *Word-formation in English*. Cambridge: Cambridge University Press.
- Plag, Ingo/Dalton-Puffer, Christiane/Baayen, R. Harald (1999), Productivity and Register. In: *English Language and Linguistics* 3, 209–228.
- Pluymaekers, Mark/Ernestus, Mirjam/Baayen, R. Harald (to appear), Frequency and Acoustic Length: The Case of Derivational Affixes in Dutch. In: *Journal of the Acoustical Society of America*.
- Pothos, Emmanuel M. (2005), The Rules versus Similarity Distinction. In: *Behavioral and Brain Sciences* 28, 1–49.
- Scherer, Carmen (2005), *Wortbildungswandel und Produktivität. Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.
- Schultink, Henk (1961), Produktiviteit als morfologisch fenomeen. In: *Forum der Letteren* 2, 110–125.
- Seidenberg, Mark (1987), Sublexical Structures in Visual Word Recognition: Access Units or Orthographic Redundancy. In: Coltheart, M. (ed.), *Attention and Performance XII*. Hove: Lawrence Erlbaum Associates Hove, 245–264.
- Skousen, Royal (1989), *Analogical Modeling of Language*. Dordrecht: Kluwer.
- Wurm, H. Lee/Aycock, Joanna/Baayen, R. Harald (manuscript submitted for publication), Lexical Dynamics for Low-frequency Complex Words: A Regression Study across Tasks and Modalities.

42. Corpora and syntax

1. Introduction
2. Treebank-based case studies
3. Summary
4. Literature

1. Introduction

Syntactic analysis connects empirical observations about language with theoretical generalizations and explanations. Depending on the perspective of the framework or individual researcher, syntactic research has emphasized the empirical or the theoretical aspect of the enterprise; but independent of the philosophical dispute between empiricism and rationalism about the nature of the connection between data and knowledge (cf., e.g., Markie 2004), it is clear that neither aspect exists entirely without the other: observation of data is shaped by prior experience and current research questions, and data is needed for establishing or falsifying a theory. Leaving the philosophical dispute aside, we can thus ask how one can obtain data that is relevant for a particular theoretical issue. We address this question in this article by discussing how electronic corpora can be used in support of the creation and falsification of syntactic theories.

1.1. On the use and limits of corpora for syntactic research

Text corpora have always been used by philologists, historical linguists, and lexicographers; but over the past decades, the availability of large electronic corpora annotated with morphological and syntactic information has significantly extended the possible uses of corpora for syntactic research. (Cf. articles 1 and 3 in this volume for the historical context, and articles 13, 23–32 and 34 for a discussion of corpus annotation.) Before we turn to exploring these interesting possibilities, let us mention some relevant issues and limitations that arise when considering the use of corpora for syntactic research:

First, annotated electronic corpora exist only for very few of the world's languages; for example, the Linguistic Data Consortium (LDC, <http://www.ldc.upenn.edu>) lists corpora for 39 languages, a small fraction of the around 6000 living languages (cf. Crystal 1997, 287). Traditional fieldwork with informants will thus remain the most important methodology for obtaining data from most living languages, at least as a first step (cf., e.g., the Open Language Archives Community, <http://www.language-archives.org/>).

Second, for the languages for which electronic corpora have been compiled and annotated, one needs to keep in mind that even the largest corpora can only represent a finite subset of a language's infinite potential. Given Zipf's law that the frequency of use of the n^{th} most frequently used word (or other phenomenon) in a corpus is inversely proportional to n , even the largest corpus will appear small for linguistic research. In consequence, to address questions involving parts of a language that happen not to occur in a corpus, syntactic research will also have to make use of handcrafted examples.

Finally, one needs to distinguish how data is *obtained* from how data is *evaluated*. Data exemplifying some theoretically interesting pattern can, e.g., be obtained by handcrafting examples, by searching in corpora, or by eliciting data from informants. Data can be evaluated on many dimensions in various qualitative and quantitative ways, e.g., through psycholinguistic experiments, introspection, neuroimaging, or analysis of corpus frequency. Often the evaluation method is independent of how the data to be evaluated was obtained; for example, while it is traditional in generative linguistics to handcraft examples and evaluate them introspectively, it is equally possible to search for interesting examples in corpora and evaluate those introspectively. Other evaluation methodologies, such as quantitative corpus analysis, are dependent on how the data was obtained given that such an analysis relies on representative corpora, a full understanding of the corpus query language and query tool to ensure that the relevant data set is obtained with high precision and recall, and typically a large corpus size to obtain statistically significant results. While in this article we focus on obtaining corpus data, assuming a traditional qualitative syntactic analysis, Stefanowitsch (2005) shows that the often-cited generative linguistic arguments against a quantitative corpus analysis are questionable, and article 36 in this volume provides a detailed discussion of statistical methods for corpus exploitation.

We turn to the question why it is particularly attractive to make use of corpus searches for syntactic research. To study a syntactic phenomenon, one needs to reduce examples to whatever properties are relevant for the linguistic issues being researched and to vary selected properties in order to explore the grammatical correlations. This is a complex undertaking that assumes an understanding of what properties can play a role for a given linguistic issue – which often is far from clear, as illustrated by the fact that supposedly syntactic effects in recent years have turned out to be explainable by long-overlooked contextual properties (cf., e.g., De Kuthy/Meurers 2003).

Corpus data obtained by searching for a linguistically relevant pattern exhibits a wide variation of known and unknown parameters and can include information on the context, as needed for exploring the interaction of constraints from syntax and formal pragmatics. When searching for a particular pattern in a corpus, it is thus possible to observe the theoretically interesting pattern within sentences that exhibit a wide variation of lexical, syntactic, semantic, and contextual properties; this makes it possible to obtain a better picture of which of these properties are relevant for a given phenomenon. The fact that corpus examples generally are natural and contextualized can also be helpful whenever examples are to be evaluated through introspection.

Having situated and motivated the use of corpora for syntactic research, we are now ready to address the question how data exhibiting theoretically interesting patterns can be found, and what corpora and annotations are needed to support searching for such patterns. After discussing some basic issues in the next section, we turn to a series of small case studies involving corpora with full syntactic annotation in section 2.

1.2. Basics of syntactically motivated corpus searches

In using corpora for syntactic research, we want to find instances of some pattern of linguistic relevance in order to explore, support, or refute a linguistic claim involving that pattern. Syntactic research, at the fundamental empirical level, observes words, their

form, order and cooccurrence in a sentence. The patterns of interest in syntactic research are, however, typically described in terms of generalizations and abstractions over the form and order of words (or groups thereof, for those syntactic paradigms that assume a notion of constituency). This raises the question how a syntactic pattern of interest can be characterized in terms of the properties of a particular corpus and its annotation.

Unannotated corpora, precision and recall of queries

The most basic kind of corpus consists of plain text; tokenized, but without linguistic annotations or segmentation. Using such corpora for linguistic research is essentially like using a basic search engine on the web, and indeed the web has gained a significant popularity as an enormous, searchable text repository (cf., e.g., Kilgarriff/Grefenstette 2004; Lüdeling/Evert/Baroni 2007; and article 18). The use of such unannotated corpora for syntactic research requires formulating queries which explicitly list lexical possibilities and spell out entire paradigms given that no generalizations or abstractions can directly be referred to in the query. Since it is complex and often simply impossible to extensionally encode a general syntactic pattern, one has to approximate the intended pattern to be searched, which results in decreased precision and recall.

Precision here measures how many correct matches (vs. false positives) the search for a particular syntactic pattern returns, and *recall* reports how many of the relevant examples in the corpus were found by the search. From our linguistic perspective, a search with low recall for a particular language pattern means that many instances of the pattern of interest are missed. It can still be sufficient for finding examples counter-exemplifying a particular claim, but for empirically grounding a linguistic theory, the partial empirical blindness caused by searches with low recall is a problem. Searching for a pattern with low precision, on the other hand, means that the search results will contain many false positives that one needs to weed through, generally by hand, in order to find the pattern instances one was actually interested in – which in practice might or might not be feasible.

The utility and caveats of annotation

To be able to query more abstract linguistic patterns directly, one can make use of corpora that are annotated with the relevant (or related) linguistic abstractions. Meurers (2005) presents five case studies using a sentence segmented and part-of-speech (POS) annotated newspaper corpus to explore syntactic issues and address claims from the linguistic literature. Meurers discusses how increasingly complex syntactic patterns can be expressed in terms of the properties available in such a corpus. Given the increased availability of corpora with more complex syntactic annotation, the case studies in the present article will focus on the use of *treebanks* for syntactic research, which we turn to in section 2, after discussing some general issues that are relevant in the context of using annotated corpora.

Compared to working with unannotated corpora, some of the mentioned complexity resulting from approximating patterns extensionally and the resulting loss in precision/recall can be avoided by searching in corpora with relevant linguistic annotations. At the same time, the move to using annotated corpora also opens the door to a new problem that can negatively impact precision and recall of queries: errors in the annotation. Even the so-called gold-standard POS or syntactic annotation currently available contains a significant number of errors (cf. van Halteren 2000; Květoň/Oliva

2002; Dickinson/Meurers 2003, 2005; Dickinson 2005, and references cited therein). For example, the POS assignment in the widely used Wall Street Journal corpus (WSJ, Marcus/Santorini/Marcinkiewicz 1993) has an estimated 3% error rate. Such annotation errors can result from shortcomings of the annotation scheme, its documentation, or the failure of the human annotators or correctors to apply the annotation guidelines correctly and consistently throughout the corpus. The effect of even a couple of percent of annotation errors on the use of such corpora for syntactic research should not be underestimated. Given Zipf's law, a syntactic pattern of interest can easily have only few occurrences in a corpus. In addition, an error rate such as the 3% mentioned for the WSJ above, is not evenly distributed over all annotation distinctions; instead, certain tokens are unambiguous or trivial to annotate, whereas other distinctions are very difficult to make (and to make consistently). The latter will thus exhibit an error rate many times higher than that of the corpus as a whole. In sum, the annotation errors present in current gold-standard corpora can seriously impact precision/recall of a query relying on distinctions which happen not to be made reliably in the corpus annotation.

A related point concerns the fact that large corpora, traditionally those with one million tokens or more, for practical reasons can only be annotated automatically; and even the annotation of smaller corpora typically arises from a semi-automatic annotation process, where human annotation or correction is based on the output of automatic taggers and parsers. As a result, the fact that current NLP technology cannot reliably make certain distinctions, such as the resolution of argument/adjunct or attachment ambiguities, means that these distinctions will often be incorrect in the annotated corpora or excluded from the annotation scheme to begin with (as seen by the prevalence of flat syntactic annotation in currently available treebanks).

Turning from errors in the application (or the definition or the documentation) of an annotation scheme to the foundation of the annotation itself, one needs to keep in mind that the annotation schemes used are the result of linguistic theorizing and insight. Of course, current syntactic research frequently questions the established analyses, and a particular set of data might be interesting precisely because the delineation of a phenomenon and/or its analysis are not yet adequately understood. For example, a corpus annotated based on the traditional syntactic assumption that German only allows a single constituent to be fronted naturally would not produce many results for a query referring to this annotation when searching for examples where more than one constituent has been fronted. In a sense, writing queries referring to corpus annotation instead of the corpus data itself is much like writing a travel book based on someone else's photos instead of visiting the place oneself – with all the pros and cons that this entails.

Finally, the use of corpora with structural annotation requires the use of a more sophisticated query language in order to refer to the various linguistic properties and the dominance and precedence relations encoded by the annotation. The case studies we turn to in the next section make use of the TIGERSearch tool (Lezius 2002), and its query language will be introduced there. The core components our discussion is based on should carry over to most other query languages designed for syntactically annotated corpora (cf., e.g., Pito 1994; Randall 2000; Rohde 2001; McKelvie 2001; Kepser 2003; Kallmeyer/Steiner 2003; Carletta et al. 2006). But before ending this section, let us mention an interesting, somewhat different approach to querying syntactic corpora: the Linguist's Search Engine (Resnik/Elkiss 2005, <http://lse.umiacs.umd.edu/>), an approach that is exemplified with two linguistic case studies in Resnik et al. (2005). The basic idea of

the Linguist's Search Engine is that a query is created by processing and generalizing an example. A parser processes an instance of the pattern one is interested in and the resulting parse tree can be manipulated to obtain a general pattern. That pattern is then used as a query to search in a corpus that has been processed with the same parser. Note that this setup has the interesting property that errors made by the parser do not have to be a problem given that both the initial instance of the search pattern and the corpus are processed with the same tool; the purpose of the parser is not to provide the ultimate linguistic analysis but to provide a link from the instance used to create the search pattern to other instances of that pattern in the corpus.

2. Treebank-based case studies

Following the discussion of the general issues involved, we now turn to three linguistic case studies exemplifying the use of a treebank for syntactic research. We discuss three phenomena of general interest for the architecture of grammar and show that a thorough empirical base is important both for constructing new linguistic analyses and for constructing arguments to support or refute existing theories. We focus on the question how to find the relevant data in corpora and organize the discussion based on an increasing complexity of the query that is needed to obtain the desired types of examples.

The TIGER corpus

The case studies are based on the TIGER Corpus (v.1, Brants et al. 2004) and the query tool TIGERSearch (Lezius 2002, <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/>). The TIGER corpus is a German newspaper corpus consisting of roughly 700,000 tokens (40,000 sentences), taken from the *Frankfurter Rundschau*, a national German newspaper. It was semi-automatically annotated with part-of-speech information (using the Stuttgart-Tübingen-Tagset; Schiller/Teufel/Thielen 1995) and syntactic information. The syntactic annotation consists of tree structures with node and edge labels. The trees focus on encoding the argument structure and are relatively flat, e. g., in a prepositional phrase, the preposition, the determiner and the noun are directly dominated by a PP node. The nodes encode the syntactic category (e. g., NP, PP), and the edge labels are used to encode grammatical functions (e. g., subject, object). There are no empty terminal nodes; instead the annotation scheme allows for discontinuous constituents. For instance, the extraposed relative clause *der lacht* in (1) is annotated as directly dominated by an NP node that also directly dominates the determiner *ein* and the noun *Mann*, but excludes the intervening verb *kommt*.

- (1) *Ein Mann kommt, der lacht.*
 a man comes who laughs

The query language of TIGERSearch

The query language consists of two levels: nodes and relations. Nodes can be described by Boolean expressions over feature-value pairs. For instance, the query

[word="suche" & pos="VVFIN"]

finds all words with the orthography *süche* and the part-of-speech tag for finite verbs (VVFIN). As values of features, one finds the categories distinguished by the annotation scheme (in double quotes) or so-called types (without quotes), which is an abbreviation; e.g., in the TIGER corpus the type *noun* abbreviates the POS-tags for proper noun and common noun.

Relations between two or more nodes can specify constraints on immediate precedence (.), immediate dominance (>), immediate dominance with edge label L (>L), left corner of a phrase (>@l), as well as the derived node relations of general dominance (>*) or siblings (\$). For example, the following query would find an NP dominating a sentence functioning as a relative clause, such as the one we saw in (1):

```
[cat="NP"] >RC [cat="S"]
```

In addition, a small set of special predicates can be used to describe nodes; for example, the expression discontinuous(#n) requires that the terminal yield of the node #n is not continuous.

Boolean expressions (without negation) over node relations can be used to form complex descriptions. For example, sentences that contain an S node dominating a particle (PTKVZ) and a finite verb satisfy the following query:

```
([cat="S"] > [pos="PTKVZ"]) & ([cat="S"] > [pos="VVFIN"])
```

Variables are used to express coreference of nodes or feature values. For example, the query above also returns sentences that contain two separate S nodes, one dominating the finite verb and the other one the verbal particle. One can use a variable to state that the same S node is supposed to dominate both nodes:

```
(#n:[cat="S"] > [pos="PTKVZ"]) & (#n > [pos="VVFIN"])
```

2.1. Case 1: Extraposition, complex NPs and subjacency

Turning to the first case study, Chomsky (1986, 40; among others) argues that the trace *t* in (2) cannot be the source of the extraposition and explains this by the principle of subjacency, which says that only one Barrier may be crossed by such movement.

(2) [NP Many books [PP with [stories t]] t'] were sold [that I wanted to read].

Grewendorf (1988, 281), Haider (1996, 261), and Rohrer (1996, 103) assume that subjacency also plays a role for extraposition in German. But if one substitutes the noun *books* in (2) in a way that reduces attachment ambiguities, one can obtain parallel German examples which are grammatical (S. Müller 2004):

(3) *weil viele Schallplatten mit Geschichten verkauft wurden, die ich noch lesen wollte*
because many records with stories sold were that I yet read wanted
‘because many records with stories that I wanted to read were sold.’

This sentence describes a situation where the speaker goes to a record shop and for certain audio book records there, he realizes he wants to read those stories.

In general, there seems to be no upper limit on the number of phrase nodes that may be crossed by dislocation to the right. Example (4) shows that relative clauses can be extraposed from a deeply embedded NP, and (5) shows the same for a complement clause.

- (4) *Karl hat mir [eine Kopie [einer Fälschung [des Bildes [einer Frau t_i]]]] gegeben, [die schon lange tot ist] $_i$.*
 Karl has me a copy of a forgery of a picture of a woman given who already long dead is
 ‘Karl gave me a copy of a forged picture of a woman who’s long been dead.’
- (5) *Ich habe [von [dem Versuch [eines Beweises [der Vermutung t_i]]]] gehört, daß es Zahlen gibt, die die folgenden Bedingungen erfüllen $_i$.*
 I have of the attempt of a proof of the assumption heard that it numbers gives that the following conditions satisfy
 ‘I have heard of the attempt to prove the assumption that there are numbers for which the following conditions hold.’

How can we find more examples to empirically explore this issue? Even with an unannotated corpus, examples with such extraposed complement clauses can be found by looking for sentences that contain a complementizer and a noun selecting a clausal complement. The precision of such searches is quite low, though, since in many of the matches the complement clause is not extraposed.

Using a syntactically annotated corpus one can formulate a more precise query that includes the requirement that the complement clause be extraposed. For our TIGER setup, we can express the query as follows:

```
#xp:[cat="NP"] >OC [] &
[cat="NP"|"PP"] > #xp &
discontinuous(#xp)
```

The three lines of the query have the following meaning:

1. Search for a node of category NP; use the variable #xp to refer to it. The #xp immediately dominates a node functioning as an object clause (OC).
2. The #xp is immediately dominated by a node that is an NP or a PP. (Note that immediate dominance is sufficient here since NPs in the TIGER corpus are annotated as flat structures, i. e., the determiner and the noun are sisters in a local tree; for PPs the preposition can also be found in the same local tree.)
3. The #xp is discontinuous, in which case the object clause is typically extraposed (but any other discontinuous realization of the #xp would also be matched).

Running this query on the TIGER corpus finds examples such as the one in (6).

- (6) [...] *die Erfindung der Guillotine könnte [NP die Folge [NP eines the invention of the guillotine can the consequence of a verzweifelten Versuches des gleichnamigen Doktors] gewesen sein, [seine desperate attempt of the homonymous doctor been is his*

Patienten ein für allemal von Kopfschmerzen infolge schlechter Kissen zu patients once for all of headache due to bad pillows to befreien]. free

‘The invention of the guillotine may have been the consequence of a desperate attempt of a doctor by the same name to, once and for all, free his patients of headaches caused by bad pillows.’

It is straightforward to modify this query to find extraposed relative clauses: the labeled dominance constraint $>OC$ in the first line of the query has to be replaced by $>RC$. To find sentences with extraposed relative clauses that cross one more maximal projection, we can use the following query:

```
#xp:[cat="NP"] >RC [] &
discontinuous(#xp) &
#yp:[cat=("NP"|"PP")] > #xp &
[cat=("NP"|"PP")] > #yp
```

Here, the additional maximal projection between the topmost NP or PP and the #xp is the node called #yp, which is required to be an NP or PP node itself. This query finds sentences such as the one in (7).

- (7) *Der 43jährige will nach eigener Darstellung damit [NP den Weg [PP für the 43 year old will after own account thereby the way to [NP eine Diskussion [PP über [NP den künftigen Kurs [NP der stärksten a discussion of the future direction of the strongest Oppositionsgruppierung]]]]]] freimachen, [die aber mit 10,4 Prozent der opposition party clear which however with 10,4 percent of the Stimmen bei der Wahl im Oktober weit hinter den Erwartungen votes at the elections in October far behind the expectations zurückgeblieben war]. remained was*
- ‘By his own account, the 43 year old thereby wants to clear the way to a discussion of the future direction of the strongest opposition party, which had, however, fallen far behind the expectations by receiving only 10.4 percent of the votes at the elections in October.’

The specification with regard to #yp ensures that the extraposition crosses more than one NP or PP node.

Based on corpus examples such as these, which we take to be well-formed ordinary sentences of German, one can conclude that subjacency or related constraints such as the Complex NP Constraint of Ross (1967) do not universally hold for movement to the right.

2.2. Case 2: The structure of the German clause and particle verbs

The second case study addresses the frequently made claim that particles of particle verbs cannot be fronted in German (cf. Müller 2002b, for an overview). The empirical issue has been used to define the class of particle verbs (Zifonun 1999, 212), and it has

played an important role in a number of syntactic arguments. For instance, Haider (1990) claimed that verb traces cannot be a part of the fronted projection, since if they were, one would expect sentence like (8) to be grammatical.

- (8) * [Ein Buch auf *t_i*] *schlug_i* *Hans*.
 a book open (PARTICLE) beat Hans
 'Hans opened a book.'

Turning to corpus searches intended to explore the empirical side of this issue, if one wants to use an unannotated corpus, one can try to look for fronted particles by searching for a particle that is separated by a space from its corresponding verb. According to orthographic conventions this would be the way to write particle and verb if the particle is fronted and the finite verb is in second position. But this requires spelling out all possible particle verbs and it clearly is questionable to rely on orthographic conventions for finding cases that supposedly do not exist at all.

Based on a syntactically annotated corpus, such as the TIGER corpus used in this study, we can formulate the following query:

[pos="PTKVZ"] . [pos=finite]

The query looks for a word with part-of-speech PTKVZ (separated verbal particle) followed by a finite verb (the type finite is an abbreviation for the finite auxiliary, modal, and main verbs tags). This query yields 36 sentences for the TIGER corpus, including sentences of the kind we are looking for (9), but also verb-final sentences like (10), which are irrelevant for our issue. See Müller (2002a, 271–272) on the status of *feststehen* and Müller (2002a, chapter 6.1.2) for more fronting examples from other corpora.

- (9) a. *Fest steht, daß dort 580 der insgesamt 4650 Arbeitsplätze wegfallen.*
 solid stands that there 580 of the in total 4650 jobs are cut
 'It is certain, that 580 of the 4650 jobs are cut.'
 b. *Verloren ging dabei endgültig das Selbstverständnis der Einheimischen.*
 lost went there.at finally the self-understanding the natives
 'Due to this the way the natives saw themselves got finally lost.'
- (10) *dem Anfang der neunziger Jahre Hohn und Spott zuteil wurde*
 who beginning of the nineties year derision and sneer part of become
 'who was derided at the beginning of the nineties'

To exclude such verb-final sentences, we can extend the query in the following way:

```
#s:[cat="S"] > #part:[pos="PTKVZ"] &
#part.[pos=finite] &
#s>@l#leftcorner &
#leftcorner:[pos= ! (prorel | prointer | conjunction)]
```

This query searches for a sentence that dominates a verbal particle which is adjacent to a finite verb. The additional constraints rule out certain clause types (relative clauses, embedded interrogative clauses, and subordinated clauses) that are verb-final and thus are not interesting in the present context. The operator >@l is used to find the leftmost terminal symbol in a tree. The last three conjuncts of the query above state that the

leftmost terminal must not be a relative pronoun, an interrogative pronoun, or a conjunction. This query results in a set of examples, all of which are relevant for the question under discussion (i. e., the query has a 100 % precision).

In sum, searching for fronted particles in a syntactically annotated corpus provides a range of examples showcasing this supposedly impossible pattern.

2.3. Case 3: Fronting as a constituent test

The third case study will lead us to the most complex query – and to the limits of what can be found in currently available corpora. German is a so-called verb-second language and a generally accepted empirical generalization is that only one constituent can appear in front of the finite verb in declarative main clauses (Erdmann 1886, ch. 2.4; Paul 1919, 69, 77). The strongest claim found in the literature is that the ability of material to appear in front of the finite verb is both sufficient and necessary for constituenthood (cf., e. g., Bubmann 1983, 446).

However, as discussed in Müller (2003), there are well-formed example sentences such as those in (11), from the national German newspaper *taz*, 16.01.2003, 6 and 03.04.2003, 9 respectively, which according to other constituent tests include more than one constituent in front of the finite verb.

- (11) a. [Gar nichts mehr] [mit dem Tabakkonzern] hat Jan Philipp
 nothing at all more with the tobacco company has Jan Philipp
 Reemtsma zu tun
 Reemtsma to do
 ‘Jan Philipp Reemtsma has nothing at all to do with the tobacco combine
 any more.’
- b. [Mit ihm] [auf der Anklagebank] sitzen zwei 18-Jährige,
 with him on the dock sit two 18 year olds
 ‘Two 18 year olds are in the dock with him ...’

Müller (2005) proposes that such examples can be analyzed by assuming an empty verbal element as the head of the fronted projection, which therefore can only include dependents of that verb. To explore and test this proposal, we want to search for the pattern in the TIGER corpus and write the following pattern:

```
#s:[cat="S"] >HD #fin:[pos=finite] &
#s >@l #sleftcorner &
#s > #vf1 &
#vf1 >@l #sleftcorner &
#vf1 >@r #vf1rightcorner &
#s > #vf2 &
#vf2 >@l #vf2leftcorner &
#vf2 >@r #vf2rightcorner &
#vf1rightcorner . #vf2leftcorner &
#vf2rightcorner .* #fin
```

This query searches for a node #s with the category S that dominates a finite verb #fin. The node #s has the left periphery #sleftcorner and immediately dominates a node #vf1 which also has the left periphery #sleftcorner. This ensures that the node #vf1 starts at the same position as #s. The right corner of #vf1 is #vf1rightcorner. The query asks for a second node that is also dominated by #s, namely #vf2. The node #vf2 has to be adjacent to #vf1, which is ensured by the constraint that the node at the right corner of #vf1 (i.e., #vf1rightcorner) immediately precedes the node at the left corner of #vf2 (i.e., #vf2leftcorner). Note that this precedence constraint cannot be encoded directly by a statement like #vf1 . #vf2, since the precedence operator . compares the left corners of two nodes, which would restrict the #vf1 node to nodes with exactly one word. Since there are sentences with more than two constituents in front of the finite verb, we do not require that the right edge of #vf2 is immediately adjacent to #fin, but in the last line instead require that the right edge of #vf2 (i.e., #vf2rightcorner) is placed somewhere to the left of the finite verb (#fin).

Unfortunately, this query returns several classes of false positives: it admits verb-final sentences containing relative or interrogative pronouns and some other constituent before the verb. The search results include sentences with complex coordinations of relative or interrogative sentences in which the relative phrase part is not part of the conjunction. Finally, the query also returns examples with adverbials such as the one in (12).

- (12) *Hier wiederum mangelt es an Opferbereitschaft.*
 here again lacks it of readiness to make sacrifice
 ‘There is an insufficient readiness to make sacrifices here.’

Such examples have been analyzed differently in the literature and do not constitute evidence for multiple frontings.

Extending the query to eliminate these three classes of false positives results in a rather complex query, which returns six results, two of which are given in (13):

- (13) a. [Am schwersten] [mit der Selbtkritik] tat sich Jürgen Kocka.
 at the heaviest with the self-criticism did self Jürgen Kocka
 ‘Jürgen Kocka had the most difficulties with self-criticism.’
 b. [Negativ] [auf den Gewinn] wirkten sich vor allem
 negative on the profit have an effect self before all
Wechselkursschwankungen aus.
 exchange rate variations PART
 ‘In particular exchange rate variations had a negative effect on the profit.’

While such examples are illustrative of the phenomenon, a set of six corpus examples is not sufficient to study and reach an understanding of the restrictions and properties of the phenomenon.

We conclude that, as a consequence of Zipf’s law, many infrequent but theoretically relevant phenomena can only be found in very large corpora, which given their size cannot be manually annotated or corrected. While searching for the constituency issue discussed in this section requires full syntactic annotation with reliable attachment disambiguation, for other rare phenomena large automatically annotated corpora, such as the 200-million-token “Tübingen Partially Parsed Corpus of Written German” (TüPP-D/Z; F. H. Müller 2004, Ule 2004) can be an interesting option.

3. Summary

Following an introduction characterizing the context of using corpora in syntactic research, we investigated how unannotated and annotated corpora can be searched to find data exemplifying patterns of interest to theoretical syntax. Based on three case studies making use of a syntactically annotated newspaper corpus, we illustrated that searching for relevant corpus examples can serve as an important component of empirically grounded syntactic research.

4. Literature

- Brants, S./Dipper, S./Eisenberg, P./Hansen-Schirra, S./König, E./Lezius, W./Rohrer, C./Smith, G./Uszkoreit, H. (2004), TIGER: Linguistic Interpretation of a German Corpus. In: *Research on Language and Computation* 2(4), 597–620.
- Bußmann, H. (1983), *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kröner Verlag.
- Carletta, J./Evert, S./Heid, U./Kilgour, J. (2006), The NITE XML Toolkit: Data Model and Query Language. In: *Language Resources and Evaluation* 39(4), 313–334. Available at: <http://www.ltg.ed.ac.uk/NITE/papers/NXT-LREJ.web-version.ps>.
- Chomsky, N. (1986), *Barriers*. Cambridge, MA/London, UK: The MIT Press.
- Crystal, D. (1997), *The Cambridge Encyclopedia of Language*. 2nd edition. Cambridge: Cambridge University Press.
- De Kuthy, K./Meurers, W. D. (2003), The Secret Life of Focus Exponents, and What it Tells us about Fronted Verbal Projections. In: Müller, S. (ed.), *Proceedings of the Tenth Int. Conference on HPSG*. Stanford, CA: CSLI Publications, 97–110. Available at: <http://ling.osu.edu/~dm/papers/dekuthy-meurers-hpsg03.html>.
- Dickinson, M. (2005), Error Detection and Correction in Annotated Corpora. PhD thesis, Department of Linguistics, Ohio State University.
- Dickinson, M./Meurers, W. D. (2003), Detecting Errors in Part-of-speech Annotation. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, 107–114. Available at: <http://www.aclweb.org/anthology/E/E03/E03-1068>.
- Dickinson, M./Meurers, W. D. (2005), Detecting Errors in Discontinuous Structural Annotation. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Ann Arbor, MI, 322–329. Available at: <http://www.aclweb.org/anthology/P/P05/P05-1040>.
- Erdmann, O. (1886 [1985]), *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung*, vol. 1. Stuttgart: Verlag der J. G. Cotta'schen Buchhandlung. Reprint: Hildesheim, Georg Olms Verlag.
- Grewendorf, G. (1988), *Aspekte der deutschen Syntax. Eine Rektions-Bindungs-Analyse*. Tübingen: Gunter Narr Verlag.
- Haider, H. (1990), Topicalization and Other Puzzles of German Syntax. In: Grewendorf, G./Sternefeld, W. (eds.), *Scrambling and Barriers*. Amsterdam/Philadelphia: Benjamins, 93–112.
- Haider, H. (1996), Downright Down to the Right. In: Lutz, U./Pafel, J. (eds.), *On Extraction and Extrapolation in German*. Amsterdam: Benjamins, 245–271.
- Kallmeyer, L./Steiner, I. (2003), Querying Treebanks of Spontaneous Speech with VIQTORYA. In: *Traitement Automatique des Langues* 43(2), 155–179.
- Kepser, S. (2003), Finite Structure Query: A Tool for Querying Syntactically Annotated Corpora. In: *EACL '03: Proceedings of the Tenth Conference on European Chapter of the Association for*

- Computational Linguistics*. Morristown, NJ: Association for Computational Linguistics, 179–186.
- Kilgarriff, A./Grefenstette, G. (2004), Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics* 29(3), 333–348.
- Květoň, P./Oliva, K. (2002), Achieving an Almost Correct PoS-tagged Corpus. In: Sojka, P./Kopeček, I./Pala, K. (eds.), *TSD 2002*. Heidelberg: Springer, 19–26.
- Lezius, W. (2002), Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. PhD thesis, IMS, Universität Stuttgart. Appeared as *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)* 8(4). Available at: <http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/dissezius.pdf>.
- Lüdeling, A./Evert, S./Baroni, M. (2007), Using Web Data for Linguistic Purposes. In: Hundt, M./Nesselhauf, N./Biewer, C. (eds.), *Corpus Linguistics and the Web*. (Language and Computers – Studies in Practical Linguistics 59.) Amsterdam/New York: Rodopi, 7–24. Available at: <http://purl.org/stefan.evert/PUB/LuedelingEvertBaroni2005.pdf>.
- Marcus, M./Santorini, B./Marcinkiewicz, M. A. (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19(2), 313–330. Available at: <http://www.aclweb.org/anthology/J/J93/J93-2004>.
- Markie, P. (2004), Rationalism vs. Empiricism. In: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. Available at: <http://plato.stanford.edu/archives/fall2004/entries/rationalism-empiricism/>.
- McKelvie, D. (2001), *XMLQUERY 1.5 Manual*. Technical report/Web page. University of Edinburgh. Available at: <http://www.cogsci.ed.ac.uk/~dmck/xmlstuff/xmlquery/index.html>.
- Meurers, W. D. (2005), On the Use of Electronic Corpora for Theoretical Linguistics. Case Studies from the Syntax of German. In: *Lingua* 115(11), 1619–1639. Available at: <http://ling.osu.edu/~dm/papers/meurers-03.html>.
- Müller, F. H. (2004), *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen. Available at: <http://www.sfb441.uni-tuebingen.de/a1/Publikationen/stylebook-04.pdf>.
- Müller, S. (1999), *Deutsche Syntax deklarativ. Head-driven Phrase Structure Grammar für das Deutsche*. Tübingen: Max Niemeyer Verlag. Available at: <http://hpsg.fu-berlin.de/~stefan/Pub/hpsg.html>.
- Müller, S. (2002a), *Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German*. (Studies in Constraint-based Lexicalism 13.) Stanford: CSLI Publications.
- Müller, S. (2002b), Syntax or Morphology: German Particle Verbs Revisited. In: Dehé, N./Jackendoff, R. S./McIntyre, A./Urban, S. (eds.), *Verb-particle Explorations*. Berlin/New York: Mouton de Gruyter, 119–139. Available at: <http://hpsg.fu-berlin.de/~stefan/Pub/syn-morph-part.html>.
- Müller, S. (2003), Mehrfache Vorfeldbesetzung. In: *Deutsche Sprache* 31(1), 29–62. Available at: <http://hpsg.fu-berlin.de/~stefan/Pub/mehr-vf-ds.html>.
- Müller, S. (2004), Complex NPs, Subjacency, and Extraposition. In: *Snippets* 8, 10–11. Available at: <http://hpsg.fu-berlin.de/~stefan/Pub/subjacency.html>.
- Müller, S. (2005), Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. In: *Linguistische Beiträge* 203, 297–330. Available at: <http://hpsg.fu-berlin.de/~stefan/Pub/mehr-vf-lb.html>.
- Paul, H. (1919 [1968]), *Deutsche Grammatik. Teil IV: Syntax*, vol. 3. Halle an der Saale: Max Niemeyer Verlag. 2nd unchanged edition, Tübingen: Max Niemeyer Verlag.
- Pito, R. (1994), *TGREPDOC. Manual Page for tgrep*. Available at: <http://mccawley.cogsci.uiuc.edu/corpora/tgrep.pdf>.
- Randall, B. (2000), *CorpusSearch User's Manual*. University of Pennsylvania. Technical report/Web page. Available at: <http://www.ling.upenn.edu/mideng/ppcme2dir>.
- Resnik, P./Elkiss, A. (2005), The Linguist's Search Engine: An Overview. In: *Proceedings of the ACL-05 Interactive Poster and Demonstration Sessions*. Ann Arbor, MI, 33–36. Available at: <http://www.aclweb.org/anthology/P/P05/P05-3009>.

- Resnik, P./Elkiss, A./Lau, E./Taylor, H. (2005), The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine. In: *Proceedings of the 31st Meeting of the Berkeley Linguistics Society (BLS-31)*. Berkeley, CA: Berkeley Linguistics Society, 265–276.
- Rohde, D. (2001), *Tgrep2. The Next-generation Search Engine for Parse Trees*. Version 1.02. Technical report/Web page, Carnegie Mellon University. Available at: <http://www-2.cs.cmu.edu/~dr/Tgrep2/>.
- Rohrer, C. (1996), Fakultativ kohärente Infinitkonstruktionen im Deutschen und deren Behandlung in der Lexikalisch Funktionalen Grammatik. In: Harras, G./Bierwisch, M. (eds.), *Wenn die Semantik arbeitet. Klaus Baumgärtner zum 65. Geburtstag*. Tübingen: Max Niemeyer Verlag, 89–108.
- Ross, J. R. (1967), Constraints on Variables in Syntax. PhD thesis, MIT, Cambridge, MA. Appeared as Ross, J. R. (1986), *Infinite Syntax*. Norwood, NJ: Ablex Publishing Corporation.
- Schiller, A./Teufel, S./Thielen, C. (1995), *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical report, IMS-CL, Univ. Stuttgart and SFS, Univ. Tübingen. Available at: ftp://www.ims.uni-stuttgart.de/pub/corpora/stts_guide.ps.gz.
- Stefanowitsch, A. (2005), New York, Dayton (Ohio), and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 1(2), 295–301.
- Ule, T. (2004), *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen. Available at: <http://www.sfs.uni-tuebingen.de/tupp/dz/markupmanual.pdf>.
- van Halteren, H. (2000), The Detection of Inconsistency in Manually Tagged Text. In: Abeillé, A./Brants, T./Uszkoreit, H. (eds.), *Proceedings of LINC-00*. Luxembourg, 48–55.
- Zifonun, G. (1999), Wenn mit alleine im Mittelfeld erscheint: Verbpartikeln und ihre Doppelgänger im Deutschen und Englischen. In: Wegener, H. (ed.), *Deutsch kontrastiv. Typologisch-vergleichende Untersuchungen zur deutschen Grammatik*. Tübingen: Stauffenburg Verlag, 211–234.

*W. Detmar Meurers, Tübingen (Germany)
and Stefan Müller, Berlin (Germany)*

43. Corpora and grammar

1. Introduction
2. Structure-sensitive collocates
3. Collocational frameworks and grammar patterns
4. Colligates
5. Collostructional analysis
6. Outlook and desiderata
7. Final remarks: Association measures vs. raw frequencies
8. Literature

1. Introduction

The study of grammar is a relatively recent activity in corpus linguistics: for a long time, the word (more specifically, the orthographic word form) was the primary unit of investigation. As a consequence, the majority of corpus-linguistic studies have dealt with

- Resnik, P./Elkiss, A./Lau, E./Taylor, H. (2005), The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine. In: *Proceedings of the 31st Meeting of the Berkeley Linguistics Society (BLS-31)*. Berkeley, CA: Berkeley Linguistics Society, 265–276.
- Rohde, D. (2001), *Tgrep2. The Next-generation Search Engine for Parse Trees*. Version 1.02. Technical report/Web page, Carnegie Mellon University. Available at: <http://www-2.cs.cmu.edu/~dr/Tgrep2/>.
- Rohrer, C. (1996), Fakultativ kohärente Infinitkonstruktionen im Deutschen und deren Behandlung in der Lexikalisch Funktionalen Grammatik. In: Harras, G./Bierwisch, M. (eds.), *Wenn die Semantik arbeitet. Klaus Baumgärtner zum 65. Geburtstag*. Tübingen: Max Niemeyer Verlag, 89–108.
- Ross, J. R. (1967), Constraints on Variables in Syntax. PhD thesis, MIT, Cambridge, MA. Appeared as Ross, J. R. (1986), *Infinite Syntax*. Norwood, NJ: Ablex Publishing Corporation.
- Schiller, A./Teufel, S./Thielen, C. (1995), *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical report, IMS-CL, Univ. Stuttgart and SFS, Univ. Tübingen. Available at: ftp://www.ims.uni-stuttgart.de/pub/corpora/stts_guide.ps.gz.
- Stefanowitsch, A. (2005), New York, Dayton (Ohio), and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 1(2), 295–301.
- Ule, T. (2004), *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Sonderforschungsbereich 441, Seminar für Sprachwissenschaft, Universität Tübingen. Available at: <http://www.sfs.uni-tuebingen.de/tupp/dz/markupmanual.pdf>.
- van Halteren, H. (2000), The Detection of Inconsistency in Manually Tagged Text. In: Abeillé, A./Brants, T./Uszkoreit, H. (eds.), *Proceedings of LINC-00*. Luxembourg, 48–55.
- Zifonun, G. (1999), Wenn mit alleine im Mittelfeld erscheint: Verbpartikeln und ihre Doppelgänger im Deutschen und Englischen. In: Wegener, H. (ed.), *Deutsch kontrastiv. Typologisch-vergleichende Untersuchungen zur deutschen Grammatik*. Tübingen: Stauffenburg Verlag, 211–234.

*W. Detmar Meurers, Tübingen (Germany)
and Stefan Müller, Berlin (Germany)*

43. Corpora and grammar

1. Introduction
2. Structure-sensitive collocates
3. Collocational frameworks and grammar patterns
4. Colligates
5. Collostructional analysis
6. Outlook and desiderata
7. Final remarks: Association measures vs. raw frequencies
8. Literature

1. Introduction

The study of grammar is a relatively recent activity in corpus linguistics: for a long time, the word (more specifically, the orthographic word form) was the primary unit of investigation. As a consequence, the majority of corpus-linguistic studies have dealt with

lexical issues (see article 58). The reason for this bias, which to some degree exists to this day, is mainly a methodological one: corpora are accessed via word forms, making them a natural choice for a focal point around which observations are made and theories are built. However, advances in automatic tagging and parsing (see article 13 and article 28) as well as the arrival of reasonably-sized corpora containing detailed manual or semi-manual grammatical annotation (see article 34) have increasingly enabled corpus linguists to shift their attention towards genuinely grammatical issues. This reorientation has not, for the most part, led researchers to discard the study of lexis. On the contrary, much of the recent quantitatively oriented corpus-based research of grammatical phenomena has been centrally concerned with the relationship between lexis and grammar. This article focuses on this line of investigation (see article 42 for more qualitative uses of corpora in the study of grammar).

2. Structure-sensitive collocates

A first, albeit indirect step toward the corpus-based investigation of grammar and its interaction with the lexicon is taken in a variant of collocational analysis that retrieves collocates on the basis of their part of speech and/or their syntactic relation to the node word rather than on the basis of their linear position. For example, a researcher may retrieve the adjectival collocates of a particular noun, the nominal collocates in subject position of a particular verb, etc. This method is primarily aimed at removing some of the noise of purely linear collocational techniques and thus achieves greater precision. As an example, consider Table 43.1, which shows the fifteen most frequent collocates directly preceding *time* (the most frequent noun in the BNC World corpus) as well as the fifteen most frequent *adjectival* collocates in the same position.

Tab. 43.1: Most frequent left collocates of *time* in a one-percent *n*-th line sample of the BNC World

All parts of speech		Adjectives only	
	F		F
<i>the</i>	266	<i>long</i>	38
<i>first</i>	104	<i>good</i>	11
<i>this</i>	96	<i>spare</i>	7
<i>of</i>	72	<i>little</i>	6
<i>same</i>	67	<i>present</i>	6
<i>a</i>	65	<i>whole</i>	5
<i>that</i>	49	<i>short</i>	5
<i>in</i>	44	<i>right</i>	4
<i>some</i>	39	<i>sufficient</i>	3
<i>long</i>	38	<i>best</i>	3
<i>to</i>	26	<i>appropriate</i>	3
<i>any</i>	25	<i>reasonable</i>	3
<i>last</i>	25	<i>real</i>	3
<i>every</i>	23	<i>different</i>	3
<i>no</i>	21	<i>particular</i>	3

Tab. 43.2: Most frequent [A + N] combinations in an *n*-th line one-percent sample of the BNC

[ADJ + N] combination		[ADJ + N] combination	
	F		F
<i>Prime Minister</i>	102	<i>local government</i>	29
<i>other hand</i>	65	<i>European Community</i>	26
<i>local authorities</i>	44	<i>wide range</i>	26
<i>long time</i>	42	<i>working class</i>	25
<i>Soviet Union</i>	41	<i>armed forces</i>	24
<i>other words</i>	41	<i>old man</i>	23
<i>local authority</i>	37	<i>higher education</i>	23
<i>labour party</i>	37	<i>front door</i>	22
<i>hon. friend</i>	36	<i>social security</i>	22
<i>male speaker</i>	35	<i>other things</i>	22
<i>other side</i>	34	<i>private sector</i>	21
<i>other people</i>	34	<i>other countries</i>	20
<i>young people</i>	33	<i>central government</i>	20
<i>chief executive</i>	31	<i>great deal</i>	20
<i>video-taped report</i>	30	<i>recent years</i>	20
<i>hon. gentleman</i>	29		

Retrieving only adjectival collocates removes many function words that are often not particularly informative with respect to the node word. This procedure is frequently used, and is implemented, for example, in the SARA concordancing tool (see article 33), which allows the user to restrict collocates to a particular part of speech. Clearly, grammar is still relatively peripheral in this approach, serving only to make lexical analyses more precise.

A related, but slightly more grammar-oriented approach involves choosing a grammatical frame and then investigating several or even all co-occurring words in this frame. As an example, Table 43.2 lists the 31 most frequent [ADJ + N] combinations in the BNC World edition (note that here and throughout this article we use the format [POS + POS] rather than the authors' own representational formats).

Typically, such frames are relatively specific. For example, Justeson/Katz (1995a) use the frame [ADJ + N] for the purpose of disambiguating different senses of adjectives and Justeson/Katz (1995b) use the same frame as well as other NP frames (such as [N + N], [N + P + N], etc.) for the purposes of terminology identification; Krenn (2000) and Krenn/Evert (2001) use the frame [PP + V] to investigate the success of different association measures in identifying figurative expressions and support-verb constructions in German (see Evert/Kermes 2003 and Evert 2004 for a similar research question involving [ADJ + N] frames); Gries (2003) uses [ADJ + N] to distinguish between *-ic/-ical* adjective pairs; and Wulff (2003) uses the frame [ADJ + ADJ + N] to investigate factors influencing adjective order in English.

However, such frames can also be relatively abstract, as in Justeson/Katz (1991) who use *s[... ADJ ... ADJ ...]* to identify degrees of association between antonymic adjectives or Stefanowitsch/Gries (2005), who use *s[... V ... V ...]* to identify baseline co-occurrences of verbs within a sentence.

The majority of studies taking this approach retain a focus on words or lexically filled multi-word expressions and their properties. The inclusion of syntactic information is usually still aimed at improving the precision of collocational techniques rather than at the investigation of genuinely syntactic issues (although Krenn/Evert 2001 and Wulff 2003 are exceptions to some degree). However, by using grammatical frames to retrieve collocates, the crucial role of grammar is acknowledged and grammatical information is – implicitly or explicitly – taken into account in such studies. Investigating, for example, adjective-noun combinations does not only tell us something about the adjectives and nouns involved, but also about the syntax and semantics of the adjectival modification of nouns.

3. Collocational frameworks and grammar patterns

A second, much more explicit approach to the corpus-based investigation of grammar is the notion of collocational frameworks. These are defined as discontinuous sequences of two function words with an intervening content word such as [*a + N + of*] or [*too + ADJ + to*] by Renouf/Sinclair (1991), who show that words do not usually occur randomly in a given framework but typically belong to particular semantic classes associated with the framework in question. As an example, consider Table 43.3, which shows the twenty most frequent nouns occurring in the framework [*a + N + of*] in the BNC World.

Tab. 43.3: Most frequent nouns in the framework [*a + N + of*] in an *n*-th line one-percent sample of the BNC World

Word		Word (contd.)	
	F		F
<i>lot</i>	143	<i>piece</i>	25
<i>number</i>	129	<i>kind</i>	25
<i>couple</i>	76	<i>range</i>	25
<i>series</i>	54	<i>sort</i>	20
<i>bit</i>	54	<i>sense</i>	20
<i>member</i>	46	<i>part</i>	20
<i>result</i>	44	<i>group</i>	20
<i>variety</i>	41	<i>total</i>	18
<i>matter</i>	38	<i>pair</i>	18
<i>set</i>	25	<i>form</i>	17

Many of the nouns in this framework refer to quantities, and their ranking in this framework clearly does not reflect their ranking in the corpus as a whole. In other words, this – and other – collocational frameworks attract non-random, semantically restricted classes of words, a fact that Renouf/Sinclair interpret as evidence for the fact that such frameworks are linguistic units in their own right. More generally, they argue that the existence of such frameworks provides evidence for Sinclair's (1991) 'idiom principle' (although they do not use the term in this paper). The idiom principle states that "a language user has available to him or her a large number of semi-preconstructed phrases

that constitute single choices, even though they might appear to be analyzable into segments” (Sinclair 1991, 110); it contrasts with the open-choice principle, which states that “[a]t each point where a unit is completed (a word or a phrase or a clause), a large number of choices opens up and the only restraint is grammaticalness” (Sinclair 1991, 109). In Sinclair’s view, both principles coexist, but the idiom principle is much more pervasive than traditionally assumed. Crucially in the present context, collocational frameworks demonstrate that the idiom principle is clearly not limited to what would traditionally be seen as idioms.

Collocational frameworks are probably the first attempt to genuinely investigate the relationship between grammar and lexicon on the basis of collocational methods. However, the strict definition of collocational frameworks as trigrams of the form [function word + content word + function word] limits the scope of this approach.

Partly in response to this limitation, Hunston/Francis (1999) develop the notion *grammar pattern*, which they define as “all the words and structures which are regularly associated with the word and which contribute to its meaning” (Hunston/Francis 1999, 37). They list three criteria that a structure must satisfy in order to be considered a pattern: first, it must be relatively frequent; second, it must be associated with a particular word or a semantic class of words (in other words, it must conform to some degree to the idiom principle discussed above); and, third, it must contribute a clearly identifiable meaning to expressions in which it occurs. These criteria cover not just collocational frameworks, but also a wide range of other structures, from partially lexically specified expressions like [V + possessive pronoun + way + PP/Adverbial] (as in *I made my way into the orchard*) or [V + NP + into + V-ing] (as in *They talked Lewis into becoming a Christian*) to fully abstract syntactic frames like [V + NP], [V + that] or [V + NP + NP].

As an example, consider Table 43.4, which lists all verbs occurring in the first slot of the pattern [V + from + V-ing] in the BNC World, where this pattern encodes the meaning ‘the referent of the subject is a result of the process encoded by the gerund’. This structure meets all three criteria for pattern-hood: it is limited to a small set of verbs that can be characterized semantically (they are motion verbs or change-of-state verbs), it contributes to the expressions in which it occurs (the verbs by themselves do not necessarily encode ‘result’ relations), and it is relatively frequent (it occurs 266 times in the BNC World). Note that a pattern is not just a formal string of words and grammatical categories: the string [V + from + V-ing] represents a number of distinct patterns in addition to the one in Table 43.4. For example, the most frequent pattern instantiated by this string is one encoding the meaning ‘the referent of the subject actively does not

Tab. 43.4: Verbs in the first slot of the pattern [V + from + V-ing] instantiating the meaning ‘be a result of’ in the BNC World

Word		Word (contd.)	
	F		F
<i>come</i>	151	<i>flow</i>	3
<i>result</i>	67	<i>grow</i>	2
<i>arise</i>	34	<i>originate</i>	2
<i>stem</i>	7	<i>ensue</i>	1
<i>follow</i>	5	<i>develop</i>	1
<i>emerge</i>	3		

Tab. 43.5: Nouns in the pattern [V + *from* + V-ing] instantiating the meaning ‘actively not do something’

Word		Word (contd.)	
	F		F
<i>refrain</i>	248	<i>stop</i>	4
<i>keep</i>	21	<i>flinch</i>	3
<i>abstain</i>	18	<i>shy</i>	2
<i>desist</i>	17	<i>demur</i>	1
<i>withdraw</i>	5	<i>resist</i>	1

engage in the activity encoded by the gerund’. This pattern is associated with a different set of verbs shown in Table 43.5.

It is crucial to the notion of grammar patterns that not every string of words and/or grammatical categories counts as a pattern. For example, the string [V *in* NP] with an NP that encodes a location does not count as a pattern, since it is not restricted to a particular verb or class of verbs and contributes little or no semantic information to the verb it occurs with (Hunston/Francis 1999, 73).

The idea of grammar patterns is probably the most substantial advance in the corpus-based study of grammar in recent years. Still it suffers from a number of drawbacks, most importantly, first, a lack of quantification (it remains unclear how frequent a structure must be to count as a pattern); second, a lack of systematicity in its application (the criteria for pattern-hood are not always stringently applied); and, third, a lack of exhaustiveness (often, the most frequent verbs for a given pattern are left out of consideration completely).

4. Colligates

A third corpus-based approach to grammar is based on the notion *colligation*, introduced by Firth (e.g. 1968, 182) as a term for relations between grammatical categories. In corpus linguistics, the term is typically taken to refer to the co-occurrence of words with particular grammatical categories (cf., e.g., Hoey 2000, 234). Although this notion has long been recognized in corpus linguistics, surprisingly little substantial work exists explicating and/or applying it. Where it is used, it is typically operationalized in terms of word classes occurring in a particular position relative to a node word, i.e. as collocation at the level of part-of-speech. For example, Table 43.6 lists the word classes occurring immediately to the left and to the right of the word *consequence* in a random sample of 100 concordance lines from the British National Corpus (note that *of* and the different forms of *have* have their own part-of-speech tag in the BNC).

This interpretation of Firth’s idea is very close to Renouf/Sinclair’s concept of collocational frameworks: based on the lists in Table 43.6, we could, for example, hypothesize that the word occurs in the collocational framework [*a* + N + *of*] quite frequently (which it does: it occurs nine times in the sample used in Table 43.3 above). Consequently, this version of colligation analysis suffers from the same drawbacks. Most importantly, it

Tab. 43.6: Word-class colligates of *consequence* (incl. punctuation) in a random sample from the BNC World edition

LEFT COLLIGATES		RIGHT COLLIGATES	
Word class	F	Word class	F
Indefinite Article	46	<i>of</i>	55
Adjective	24	Punctuation Mark	26
Preposition	19	Personal Pronoun	4
Cardinal Numeral	5	Adverb	3
Definite Article	4	Preposition	2
<i>of</i>	2	<i>have</i> (3Sg)	2
		Indefinite Article	1
		Conjunction	1
		Possessive Pronoun	1
		Definite Article	1
		Noun (Singular)	1
		Quotation Mark	1
		Verb (3Sg)	1
		Verb (Past Tense)	1

does not take grammar into consideration beyond the level of word class and it retains a purely linear view of grammatical structure.

Recently, however, Hoey (e. g. 1997, 2004) has developed a considerably more comprehensive understanding of colligational relationships that remedies both of these shortcomings.

First, Hoey argues for a view of grammatical categories that goes beyond the notion of word class. He includes categories at considerably more abstract levels of grammatical structure, such as definiteness. As an example, take again the word *consequence*. In the sample also used above, this word occurs in indefinite contexts in 84 cases and in definite contexts in only 16 cases. Thus, Hoey would claim that *consequence* has a colligational preference for indefiniteness that goes beyond its preference for the determiner *a* at the position immediately to its left.

Second, Hoey includes hierarchical structure under his notion of colligation. Specifically, he suggests that the association of words to particular grammatical functions like subject, object, and complement can be insightfully investigated. For example, Hoey observes that the noun *consequence* frequently occurs as part of a complement but is rarely found in object position. The sample used above confirms this observation (cf. Table 43.7). The word *consequence* colligates strongly with the grammatical function *adverbial*, followed by *subject complement* and *subject*. The function *object* is clearly avoided.

Hoey's notion of colligation is broad enough to include many studies of lexicogrammatical phenomena even where these do not use this term (cf. for example Mair on gerundial and infinitival complements after *begin* and *start* (Mair 2003) and on infinitival complementation in general (Mair 1990), Noël (2003) on infinitives, accusatives and *that*-clauses, and many similar studies).

Finally, it deserves mention that Hoey extends the idea of colligational associations to relationships between words and positions in texts (such as the beginning of a sen-

Tab. 43.7: Grammatical-function colligates of consequence in a random sample from the BNC World edition

Word class	Frequency
Subject	18
Subject of an equative construction	(17)
Subject of a transitive construction	(1)
Object	4
Object of a <i>have</i> -construction	(4)
Adverbial	55
Adverbial with <i>as</i>	(25)
Adverbial with <i>in</i>	(20)
Adverbial with <i>of</i>	(9)
Adverbial with <i>by</i>	(1)
Complement	21
Subject Complement	(21)
Apposition	2
Total	100

tence, the beginning of a paragraph) etc. and between words and particular textual functions (such as disagreeing).

Hoey's view of colligation takes the crucial step towards a systematic corpus-based analysis of grammar and its relation to lexis. Like the work on collocational frameworks and grammar patterns, it shows that grammar and lexis are intertwined in intricate ways. However, also like this work it has so far not been given a strict quantitative underpinning – for example, it lacks a correction for expected baseline frequencies – and thus often remains impressionistic to some degree.

5. Collostructional analysis

The most recent attempt at a comprehensive framework for the corpus-based investigation of lexico-grammar is *collostructional analysis*, a set of methods for investigating the relationship between lexical items and (meaningful) grammatical structures based on their observed and expected co-occurrence in large corpora. Essentially, collostructional analysis is an application of Firth's notions collocation and colligation within a framework that regards grammatical structure as consisting of meaningful signs, so-called 'constructions' (hence its name, a blend of *construction* and *collocational analysis*). Thus, collostructional analysis is rooted in a theoretical tradition that distinguishes it from other current approaches in the field. In addition, it is rooted in a methodological tradition that sets it apart from many implementations of earlier approaches.

As already mentioned, the theoretical tradition is one that assumes that (some or all) grammatical structures are best viewed as meaningful linguistic units (i. e. as signs in the Saussurean sense). There are a broad variety of theories that share this assumption (for example, Hunston/Francis' *Pattern Grammar*, see above); some of these theories differ radically from each other in many other respects, but collostructional analysis can usefully be applied within any of these. In fact, it could even be applied within frameworks

that deny the possibility of meaningful grammatical structures altogether, as long as these theories posit *any* relationship at all between lexical items and grammatical structures (it would then become a version of colligational analysis, albeit a strictly quantified one, see below).

Collostructional analysis usually adopts the terminology and the background assumptions of one specific theory, Construction Grammar (Goldberg 1995). In this theory, a construction is any combination of linguistic entities whose formal or semantic properties are not fully predictable from its component parts and/or more general constructions ('rules') of the language. Crucially in the present context, constructions can have different degrees of specificity. Thus, the notion covers many of the structures referred to as 'collocational frameworks' or 'grammar patterns', but also the whole range of grammatical categories recognized by grammatical theory (including part-of-speech categories, grammatical relations, etc.). Thus, collostructional analysis captures most of the phenomena investigated in grammar-pattern analysis and colligational analysis in a unified methodological and theoretical framework.

Like these approaches, collostructional analysis has so far mainly focused on the relationship between lexis and grammatical structures. According to Construction Grammar, this relationship is determined by semantic compatibility: words occur in (slots provided by) a given construction if their meaning matches that of the construction. Collostructional analysis has confirmed this assumption from several perspectives.

The methodological tradition that collostructional analysis stems from is characterized on the one hand by a detailed, theoretically informed attention to different levels of linguistic structure, and on the other hand by the commitments of quantitative corpus linguistics: (i) the use of large, balanced corpora, (ii) the exhaustive retrieval of all instances of the phenomenon under investigation (even if this requires extensive manual post-editing), and (iii) strict statistical evaluation of the results.

5.1. Overview

If grammatical structures are linguistic signs on a par with lexical items, then the association between grammatical structures and lexical items (or other grammatical structures) can be investigated in the same way as associations between words. Instead of focusing on various types of relationships between two (or more) words, collostructional methods focus on corresponding relationships between a construction and one or more words. There are currently three such methods, each with a different focus on the association between words and grammatical constructions:

- collexeme analysis is used in investigating the association between a construction and the words occurring in a particular slot in this construction (cf., e.g., Stefanowitsch/Gries 2003) – for example, between the verb *give* and the ditransitive construction as opposed to all other constructions;
- distinctive collexeme analysis is used in investigating the association between a word and (one member of) two or more semantically or functionally equivalent constructions (cf., e.g., Gries/Stefanowitsch 2004a) – for example, between the verb *give* and the ditransitive construction as opposed to the prepositional dative;

- covarying collexeme analysis is used in investigating the association between pairs of words occurring in two different slots in the same construction (cf., e.g., Gries/Stefanowitsch 2004b, Stefanowitsch/Gries 2005) – for example, the verb and the direct object in the ditransitive construction.

Like all collocation measures, the three types of collostructional analysis are best described in terms of a two-by-two distribution table like the one shown schematically in Table 43.8.

Tab. 43.8: Distribution table

	B	$\neg B$	
A	O_{11}	O_{12}	R_1
$\neg A$	O_{21}	O_{22}	R_2
	C_1	C_2	N

The three types of collostructional analysis differ only in terms of the values assigned to A, $\neg A$, B, and $\neg B$:

- for collexeme analysis, A corresponds to a given construction, $\neg A$ corresponds to all other constructions in the corpus, B corresponds to a given word (lemma) occurring in a particular slot in A, and $\neg B$ corresponds to all other words in occurring in the corpus;
- for distinctive collexeme analysis, A corresponds to one member of a pair of constructions, $\neg A$ corresponds to the other member of the pair, B corresponds to a given word (lemma) occurring in a particular slot in A and/or $\neg A$, and $\neg B$ corresponds to all other words occurring in A and/or $\neg A$;
- for covarying collexeme analysis, A corresponds to a particular word in Slot 1 of the construction, $\neg A$ corresponds to all other words occurring in Slot 1, B corresponds to a particular word in Slot 2 of the construction, and $\neg B$ corresponds to all other words occurring in Slot 2.

In principle, any distributional statistic can be applied to such a table (see article 36 and article 58) – clearly, nothing hinges theoretically on the choice of association measure. However, given the extremely asymmetric frequency distributions typically found in natural language data, it is highly desirable to use an exact test. In collostructional analysis, the Fisher-Yates exact test is typically used. The association measure is then either the p-value or the negative base-10 logarithm of the p-value. This measure has the advantage of providing information about both the reliability of the obtained association/repulsion and its strength (cf. Stefanowitsch/Gries 2003, 238–239, n. 6 for detailed discussion), but alternative measures such as effect sizes, which would be independent of sample sizes could also be used (cf. Gries (to appear) for an example).

5.2. Collexeme analysis

Collexeme analysis is the most straightforward implementation of collocation analysis in a constructional framework: instead of a node word, the researcher retrieves all instances of a grammatical construction from the corpus, and instead of investigating

collocates (i. e. words occurring in a user-defined span around the node word), one investigates the words occurring in a particular slot provided by that construction (such words are referred to as (potential) collexemes). The latter are typically lemmatized, but looking at word forms is equally possible and tends to yield results that are conceptually similar (cf. Gries (to appear)). Each word's frequencies are entered into a distribution table as described above, and the Fisher-Yates exact test (or some other appropriate statistic) is applied to these tables. As an example, consider the verb *give* and the ditransitive construction. The ICE-GB corpus (see article 20) contains 461 occurrences of *give* used ditransitively, 699 occurrences of other uses, 574 ditransitives with other verbs, and 136,930 uses that do not contain the verb *give*, and are not ditransitive. Table 43.9 shows this information in the appropriate form, together with the expected frequencies for each cell in parentheses.

Tab. 43.9: The distribution of *give* inside and outside of the ditransitive in the ICE-GB

	<i>give</i>	Other verbs	Row totals
Ditransitive	461 (9)	574 (1,026)	1,035
Other constructions	699 (1,151)	136,930 (136,478)	137,629
Column totals	1,160	137,504	138,664

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 0, indicating that the p-value is smaller than the smallest integer that home-issue computers will output (i. e., approx. 4.94e-324). Thus, the association between *give* and the ditransitive is an extremely significant one, but this in itself does not tell the researcher anything about the direction of the association, i. e. whether *give* is significantly more frequent than expected in the ditransitive (in which case it is referred to as a *significantly attracted collexeme*), or whether *give* is significantly less frequent than expected (in which case it is referred to as a *significantly repelled collexeme*). In order to determine this, the observed frequency must be compared to the expected one. In this case, there is a positive association, i. e. *give* is a strongly attracted collexeme of the ditransitive (in fact, the most strongly attracted one). One can now apply the same procedure to all 69 verbs occurring in the ditransitive at least once, and rank the verbs in ascending order by their p-values. Table 43.10 shows the top twenty significantly attracted collexemes of the ditransitive, as well as the two only significantly repelled collexemes.

These results are typical for collexeme analysis in that they show two things. First, there are indeed significant associations between lexical items and grammatical structures. Second, these associations provide clear evidence for semantic coherence: the strongly attracted collexemes all involve a notion of 'transfer', either literally or metaphorically, which is the meaning typically posited for the ditransitive. This kind of result is typical enough to warrant a general claim that collostructional analysis can in fact be used to identify the meaning of a grammatical construction in the first place.

Concerning the repelled collexemes, there is little to say in this case, as there are only two such cases. However, it is worth noting that neither of these involves a notion of 'transfer' and thus they provide further evidence for semantic coherence. In this respect, the results in Table 43.10 are also typical; in many cases the number of repelled collex-

Tab. 43.10: Attracted and repelled collexemes in the ditransitive in the ICE-GB.

ATTRACTED COLLEXEMES		REPELLED COLLEXEMES	
Word	p	Word	p
<i>give</i> (461)	0	<i>make</i> (3)	2.72E-04
<i>tell</i> (128)	1.6E-127	<i>do</i> (10)	2.99E-03
<i>send</i> (64)	7.26E-68		
<i>offer</i> (43)	3.31E-49		
<i>show</i> (49)	2.23E-33		
<i>cost</i> (20)	1.12E-22		
<i>teach</i> (15)	4.32E-16		
<i>award</i> (7)	1.36E-11		
<i>allow</i> (18)	1.12E-10		
<i>lend</i> (7)	2.85E-09		
<i>deny</i> (8)	4.5E-09		
<i>owe</i> (6)	2.67E-08		
<i>promise</i> (7)	3.23E-08		
<i>earn</i> (7)	2.13E-07		
<i>grant</i> (5)	1.33E-06		
<i>Allocate</i> (4)	2.91E-06		
<i>wish</i> (9)	3.11E-06		
<i>accord</i> (3)	8.15E-06		
<i>pay</i> (13)	2.34E-05		
<i>hand</i> (5)	3.01E-05		

emes is much greater, and words displaying a lack of semantic coherence with respect to the construction are often predominant among these.

Note also that this method can easily be extended to words that do not occur at all in a given construction in a given corpus. For such words, collostructional analysis can determine whether they are significantly repelled by the construction or not. If they are significantly repelled, this may indicate that they are categorically barred from occurring in the construction in question – in other words, collostructional analysis allows the researcher to make principled statements about negative evidence (cf. Stefanowitsch 2006).

From a methodological perspective, a comment seems in order concerning the absence of post-hoc corrections in Table 43.10 (and in collostructional analysis in general): from a purely statistical perspective, it could be argued that the procedure described above constitutes a case of multiple testing, and thus the results would have to be corrected accordingly. This is not usually done in collostructional analysis for two reasons: first, there is a tradition in corpus linguistics to view each result as an independent test, and second, the values are mainly used for ranking items, and the ranking would not change due to post-hoc corrections.

5.3. Distinctive collexeme analysis

Distinctive collexeme analysis differs from collexeme analysis in that the association of a verb to a particular slot of a given construction is calculated not against its frequency in the corpus as a whole, but against its frequency in a corresponding slot in another

specific construction or corresponding slots in several other constructions. This strategy is particularly useful for pairs of semantically, pragmatically, or otherwise functionally similar constructions (although it can, in principle, be applied to any pair or set of constructions). As an example, consider the famous pair consisting of the ditransitive construction and the prepositional dative. Many verbs can occur in both of these constructions, and this has led a number of researchers to posit a link between them. However, it is conceivable that some or all of these verbs have significant preferences towards one of the two. Take again the verb *give*, which was shown to be highly significantly associated with the ditransitive, but which also occurs in the prepositional dative. More precisely in the ICE-GB, it occurs in the prepositional dative 146 times, and there are 1,773 occurrences of the latter with other verbs; the frequencies for the ditransitive were already given above. Table 43.11 shows this information in the appropriate form (again with expected frequencies in parentheses).

Tab. 43.11: The distribution of *give* in the ditransitive and the prepositional dative in the ICE-GB

	<i>give</i>	Other verbs	Row totals
Ditransitive	461 (213)	574 (822)	1,035
<i>To</i> -dative	146 (394)	1,773 (1,525)	1,919
Column totals	607	2,347	2,954

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 1.835954E-120, which indicates that *give* highly significantly prefers the ditransitive even when compared to the prepositional dative. Since the comparison is only between these two constructions, this automatically entails that, of the verbs that occur in both constructions, *give* is one that is not associated with the prepositional dative at all. One can now apply the same procedure to all forty verbs that occur at least once in each of the two constructions in the ICE-GB, and rank the results for each construction in descending order of the p-values. Table 43.12 shows the significantly distinctive collexemes for each construction.

Tab. 43.12: Distinctive collexemes in the ditransitive and the prepositional dative in the ICE-GB

Ditransitive (n = 1,035)		<i>To</i> -dative (n = 1,919)	
Word	p	Word	p
<i>give</i> (461:146)	1.84E-120	<i>bring</i> (7:82)	1.47E-09
<i>tell</i> (128:2)	8.77E-58	<i>play</i> (1:37)	1.46E-06
<i>show</i> (49:15)	8.32E-12	<i>take</i> (12:63)	2.00E-04
<i>offer</i> (43:15)	9.95E-10	<i>pass</i> (2:29)	2.00E-04
<i>cost</i> (20:1)	9.71E-09	<i>make</i> (3:23)	6.80E-03
<i>teach</i> (15:1)	1.49E-06	<i>sell</i> (1:14)	1.39E-02
<i>wish</i> (9:1)	5.00E-04	<i>do</i> (10:40)	1.51E-02
<i>ask</i> (12:4)	1.30E-03	<i>supply</i> (1:12)	2.91E-02
<i>promise</i> (7:1)	3.60E-03		
<i>deny</i> (8:3)	1.22E-02		
<i>award</i> (7:3)	2.60E-02		

These results are typical for distinctive-collexeme analysis in that they again provide clear evidence for associations between words and constructions and for semantic compatibility as the main principle governing these associations. Specifically, it has been argued by a number of authors that the ditransitive encodes a direct transfer of a theme from an agent to a recipient in a face-to-face situation, while the prepositional dative encodes a caused movement of a theme by an agent to a different location. The verbs in Table 43.12 reflect this distinction, most clearly in the case of the top collexemes *give* (direct transfer) and *bring* (motion to a different location).

Note that distinctive-collexeme analysis does not produce repelled collexemes, since the method assigns all collexemes to one or the other of the constructions under investigation and thus collexemes that are repelled by one construction are automatically attracted by the other.

The method has so far been applied to several classic cases of ‘alternations’ such as the verb-particle constructions or active vs. passive, but also pedagogically relevant alternations such as the *will* future vs. the *going-to* future or the *s*-genitive vs. the *of* construction. The extension to more than two alternative constructions referred to as multiple distinctive collexeme analysis can be used to investigate these cases and others in even more detail; straightforward extensions would be to investigate active vs. *be* passive vs. *get* passive or *will* future vs. *going-to* future vs. *shall* future etc.

5.4. Covarying-collexeme analysis

Covarying-collexeme analysis differs from the previous two methods in that it is not primarily concerned with the association between a word and a grammatical construction, but with the association between two words occupying specific slots in a given construction. Among the collostructional methods, it is thus most similar to traditional collocate-based or colligate-based studies in that it focuses on the relationship between words, but it differs from these methods in that, unlike collocate-based studies, it takes grammatical structure into account, and unlike colligate-based studies, it defines the words via constructions rather than via word classes in a given span. The method is useful for investigating the kinds of issues that the more traditional methods address. Take again the ditransitive construction. This construction provides three slots in addition to the verb: an agent slot (the subject in an active-voice sentence), a recipient slot (the first object in an active sentence), and a theme slot (the second object in an active sentence). A strong collocational link is expected between the verb and the theme slot (since the theme is the thing undergoing the action denoted by the verb, selectional restrictions should hold). As an example consider the word *ask* and the theme *question*. *Ask* occurs 23 times in the ditransitive, 9 times with *question* and 14 times with other themes. *Question* occurs 9 times in the ditransitive, always with the verb *ask*. Given the total number of dit transitives with object NPs in the corpus, all other figures can be derived automatically. They are shown in Table 43.13, together with the expected frequencies for each cell in parentheses.

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 5.7E-17, indicating a very strong association between these words in the ditransitive (they are referred to as a *significantly attracted collexeme pair*). One can now apply the same

Tab. 43.13: The distribution of *ask* and *question* in the ditransitive in the ICE-GB

	<i>question</i>	Other Object NPs	Row totals
<i>ask</i>	9 (0)	14 (23)	23
Other verbs	0 (9)	1,182 (1,173)	1,182
Column totals	9	1,196	1,205

Tab. 43.14: Attracted and repelled pairs of co-varying V-Object collexemes in the ditransitive in the ICE-GB

ATTRACTED COLLEXEMES		REPULLED COLLEXEMES	
Word pair	p	Word pair	p
ask–question	5.70E-17	give–what	5.26E-08
tell–what	7.04E-15	give–that	0.0005
tell–that	1.51E-13	give–pound	0.0089
do–good	5.66E-09	give–one	0.0367
offer–job	7.21E-09	give–letter	0.0833
take–minute	2.82E-08	give–job	0.1385
write–letter	4.07E-08	give–money	0.2057
guarantee–place	4.99E-08	give–card	0.245
send–copy	2.12E-07	give–minute	0.245
wish–best	2.89E-07	give–it	0.3019
wish–success	2.89E-07	give–account	0.3789
tell–story	5.93E-07	give–love	0.3789
send–cheque	6.66E-07	give–quid	0.3789
set–deadline	4.14E-06	give–room	0.3789
take–hour	7.48E-06	give–freedom	0.4266
lend–money	1.61E-05	give–sth.	0.429
drop–line	2.48E-05	offer–what	0.5086
drop–note	2.48E-05	give–cash	0.564
tell–all about NP	3.67E-05	give–detail	0.564
tell–truth	3.67E-05	give–position	0.564

procedure to all verbs and all nouns in the theme slot and sort the results as before. Table 43.14 shows the top twenty significantly attracted collexeme pairs, as well as the only four significantly repelled ones.

These results are typical for item-based co-varying-collexeme analysis: first, like the other two methods, they provide clear evidence for associations between words and constructions; and second, they represent typical collocations based on semantic coherence between the words in question. In this case, this semantic coherence is anchored in frame-based knowledge about what people typically do with which object. This is obvious not only in the case of *ask a question*, but also in *offer s. o. a job*, *write (s. o.) a letter*, *send (s. o.) a cheque*, etc. In many cases the verb-theme combinations represent fixed or semi-fixed expressions (like *tell you what*, *take (s. o.) a minute*, *wish s. o. all the*

best, or drop (s. o.) a line. Work on item-based co-varying collexemes has uncovered different types of semantic coherence between words in addition to the very concrete, frame-based one shown here, e. g. image-schematic coherence, coherence based on semantic prototypes, and coherence based on metaphors.

6. Outlook and desiderata

All of the methods discussed here – from collocational framework analysis and pattern grammar over colligation analysis to collostructional analysis – have produced a wealth of evidence concerning the association between lexical items and grammatical structures. However, there are several areas in which these methods can and should be improved.

6.1. Clustering collocates and collexemes

Collocation-based studies of words and/or grammatical categories and constructions are always faced with the problem that their result is simply a list of items ranked according to frequency or some statistical association measure. Such a list is not, in itself, an analysis of the phenomenon in question; typically, it must be grouped into semantic and/or syntactic classes before its relevance becomes clear. This grouping is usually done on the basis of intuitive common-sense criteria; clearly, a more objective, bottom-up approach would be highly desirable.

One statistical technique that lends itself well to this task is cluster analysis (see article 40), which has been used, for example, for the identification of syntactic categories (e. g. Brill et al. 1990), co-occurrence classes (e. g. Hindle 1990; Pereira/Tishby/Lee 1993; Li/ Abe 1996), and semantic classes (e. g. Waterman 1996; Schütze/Pedersen 1997; Schulte im Walde 2000).

In the context of collostructional analysis, for example, Gries/Stefanowitsch (to appear) use hierarchical agglomerative clustering to identify semantic classes among the collexemes in one slot by clustering them according to the collexemes in a different slot, for example, in the *way*-construction (as in *He made his way to the station*). They cluster the verb collexemes by the prepositional collexemes introducing the locative PP, and find clear and robust clusters of verbs of physical force (e. g., *force, push*), verbs of non-linear movement (e. g., *weave, wind*), and verbs of body-part related movement (e. g., *shoulder, elbow*). Such results provide strong evidence for the assumption that constructions are regularly associated with different related senses that are reflected by closely related groups of collexemes and, more generally, that the interpretation of collocates and collexemes can benefit greatly from further multivariate analysis.

6.2. The inclusion of additional variables

Collocation-based studies of lexis and/or grammar do not typically include additional variables, such as channel (spoken/written), register (formal/informal), dialect, gender, etc. However, a number of studies have shown that such variables have an influence, especially in the domain of lexico-grammar (cf., e. g., Biber/Conrad/Reppen 1998).

In order to allow for an easy integration of additional variables into collocation-based studies, Stefanowitsch/Gries (2005) propose an extension of the collocation-based method on the basis of configural frequency analysis using binomial tests (cf. von Eye 1990). This procedure allows the identification of positive and negative associations between variables in multi-dimensional contingency tables (as opposed to the two-by-two tables underlying traditional collocational methods). It may thus be used, among other things, to investigate triples of linguistic elements (for example, a construction and its collexemes in two different slots, as in Stefanowitsch/Gries' (2005) extension of co-varying-collexeme analysis, or two linguistic elements and an external variable such as channel, as in Gries (to appear), and Stefanowitsch/Gries (to appear)).

6.3. Word-sense sensitive analysis

For the most part, collocation-based studies of lexis and/or grammar ignore the fact that words are generally polysemous. This is mostly a matter of necessity, as there are currently no large corpora annotated for word senses (see article 26). However, it has been shown that the association of a given word to a construction may be contingent on specific senses of the word in question (cf. Roland/Jurafsky 2002). Clearly, thus, the inclusion of word senses into collocation-based approaches to grammar remains a highly desirable goal.

6.4. Dispersion

All approaches discussed here use co-occurrence frequencies to analyze the relationship between lexis and grammar, sometimes (but not always) subjected to statistical evaluation. However, even where statistical procedures are used, these have so far failed to take into account the fact that high co-occurrence frequencies can be deceptive if they are due to the influence of a small number of corpus files or the output produced by a small number of speakers (cf. Gries 2006). Future work should therefore devise ways to weigh the frequency of co-occurrence of lexical and grammatical elements on the basis of their dispersion in the corpus as a whole.

7. Final remarks: Association measures vs. raw frequencies

Finally, it is still unclear whether association measures based on statistical tests are in fact superior to raw frequencies. On a priori grounds, statistical association measures may be argued to be superior due to their higher degree of sophistication, but this assumption has been called into question, for example, by Stubbs (1995) and Kilgarriff (2005). Ultimately, this is largely a matter of empirical research, which is still largely lacking. The experimental evidence that does exist, however, provides empirical support for the superiority of statistical association measures (more precisely, the Fisher-Yates exact test outlined above). Gries/Hampe/Schönefeld (2005, to appear) compare the pre-

dictive power of collostruction strength, frequency, and subcategorization probability by means of sentence-completion tasks and a self-paced reading-time experiment and find that collostruction strength clearly outperforms the other variables. However, much more research is needed to confirm these results and provide solid evidence for more reliable generalizations in this exciting research area.

8. Literature

- Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Brill, Eric/Magerman, David/Marcus, Mitch/Santorini, Beatrice (1990), Deducing Linguistic Structure from the Statistics of Large Corpora. In: *Proceedings of the DARPA Speech and Language Workshop*. Hidden Valley: PA, 275–282.
- Evert, Stefan (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Unpublished PhD dissertation, University of Stuttgart.
- Evert, Stefan/Kermes, Hannah (2003), Experiments on Candidate Data for Collocation Extraction. In: *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Morristown, NJ: Association for Computational Linguistics, 83–86.
- Eye, Alexander von (1990), *Introduction to Configural Frequency Analysis: The Search for Types and Antitypes in Crossclassifications*. Cambridge: Cambridge University Press.
- Firth, John R. (1968), *Selected Papers of J.R. Firth 1952–59*. Edited by F. R. Palmer. London: Longman.
- Goldberg, Adele E. (1995), *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Gries, Stefan Th. (2003), Testing the Sub-test: A Collocational-overlap Analysis of English *-ic* and *-ical* Adjectives. In: *International Journal of Corpus Linguistics* 8(1), 31–61.
- Gries, Stefan Th. (2006), Some Proposals towards More Rigorous Corpus Linguistics. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 191–202.
- Gries, Stefan Th. (to appear), Corpus Data in Usage-based Linguistics: What's the Right Degree of Granularity for the Analysis of Argument Structure Constructions? In: Brda, Mario/Žic Fuchs, Milena (eds.), *Expanding Cognitive Linguistic Horizons*. Amsterdam/Philadelphia: John Benjamins.
- Gries, Stefan Th./Hampe, Beate/Schönenfeld, Doris (2005), Converging Evidence: Bringing Together Experimental and Corpus Data on the Association of Verbs and Constructions. In: *Cognitive Linguistics* 16(4), 635–676. To appear in: Newman, John/Rice, Sally (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford: CSLI.
- Gries, Stefan Th./Hampe, Beate/Schönenfeld, Doris (to appear), Converging Evidence II: More on the Association of Verbs and Constructions.
- Gries, Stefan Th./Stefanowitsch, Anatol (2004a), Extending Collostructional Analysis: A Corpus-based Perspective on ‘Alternations’. In: *International Journal of Corpus Linguistics* 9(1), 97–129.
- Gries, Stefan Th./Stefanowitsch, Anatol (2004b), Co-varying Collexemes in the Into-causative. In: Achard, Michel/Kemmer, Suzanne (eds.), *Language, Culture, and Mind*. Stanford, CA: CSLI, 225–236.
- Gries, Stefan Th./Stefanowitsch, Anatol (to appear), Cluster Analysis and the Identification of Collexeme Classes. To appear in: Newman, John/Rice, Sally (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford: CSLI.
- Hindle, Donald (1990), Noun Classification from Predicate Argument Structures. In: *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA, 268–275.

- Hoey, Michael (1997), From Concordance to Text Structure: New Uses for Computer Corpora. In: Lewandowska-Tomaszczyk, Barbara/Melia, James (eds.), *PALC'97: Practical Applications in Language Corpora*. Łódź: Łódź University Press, 2–23.
- Hoey, Michael (2000), A World beyond Collocation: New Perspectives on Vocabulary Teaching. In: Lewis, Michael (ed.), *Teaching Collocations*. Hove, UK: Language Teaching Publications, 224–243.
- Hoey, Michael (2004), Textual Colligation: A Special Kind of Lexical Priming. In: *Language and Computers* 49(1), 171–194.
- Hunston, Susan/Francis, Gill (1999), *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.
- Justeson, John S./Katz, Slava M. (1991), Co-occurrences of Antonymous Adjectives and their Contexts. In: *Computational Linguistics* 17(1), 1–19.
- Justeson, John S./Katz, Slava M. (1995a), Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. In: *Natural Language Engineering* 1, 9–27.
- Justeson, John S./Katz, Slava M. (1995b), Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. In: *Computational Linguistics* 21(1), 1–27.
- Kilgarriff, Adam (2005), Language is Never, Ever, Ever Random. In: *Corpus Linguistics and Linguistic Theory* 1(2), 263–276.
- Krenn, Brigitte (2000), *The Usual Suspects: Data-oriented Models for the Identification and Representation of Lexical Collocations*. Saarbrücken: DFKI and Universität des Saarlandes.
- Krenn, Brigitte/Evert, Stefan (2001), Can we Do Better than Frequency? A Case Study on Extracting PP-verb Collocations. In: *Proceedings of the ACL Workshop on Collocations*. Toulouse, France, 39–46.
- Li, Hang/Abe, Naoki (1996), Learning Dependencies between Case Frame Slots. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, 10–15.
- Mair, Christian (1990), *Infinitival Complement Clauses in English: A Study of Syntax in Discourse*. Cambridge: Cambridge University Press.
- Mair, Christian (2003), Gerundial Complements after *begin* and *start*: Grammatical and Sociolinguistic Factors, and How they Work against Each Other. In: Rohdenburg, Günter/Mondorf, Britta (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 330–345.
- Noël, Dirk (2003), Is There Semantics in All Syntax? The Case of Accusative and Infinitive Constructions vs. *That*-clauses. In: Rohdenburg, Günter/Mondorf, Britta (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 348–377.
- Pereira, Fernando/Tishby, Naftali/Lee, Lillian (1993), Distributional Clustering of English Words. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, 183–190.
- Renouf, Antoinette/Sinclair, John M. (1991), Collocational Frameworks in English. In: Aijmer, Karin/Altenberg, Bengt (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 128–144.
- Roland, Douglas/Jurafsky, Daniel (2002), Verb Sense and Verb Subcategorization Probabilities. In: Stevenson, Suzanne/Merlo, Paola (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. Amsterdam/Philadelphia: John Benjamins, 325–346.
- Schulte im Walde, Sabine (2000), Clustering Verbs Semantically According to their Alternation Behaviour. In: *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, 747–753.
- Schütze, Hinrich/Pedersen, Jan O. (1997), A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. In: *Information Processing and Management* 33, 307–318.
- Sinclair, John M. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stefanowitsch, Anatol (2006), Negative Evidence and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 2(1), 61–77.

- Stefanowitsch, Anatol/Gries, Stefan Th. (2003), Collostructions: Investigating the Interaction between Words and Constructions. In: *International Journal of Corpus Linguistics* 8(2), 209–243.
- Stefanowitsch, Anatol/Gries, Stefan Th. (2005), Covarying Collexemes. In: *Corpus Linguistics and Linguistic Theory* 1(1), 1–43.
- Stefanowitsch, Anatol/Gries, Stefan Th. (to appear), Register and Constructional Semantics: A Collostructional Case Study. In: Kristiansen, Gitte/Dirven, René (eds.), *Cognitive Sociolinguistics*. Berlin/Heidelberg/New York: Mouton de Gruyter.
- Stubbs, Michael (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.
- Waterman, Scott A. (1996), Distinguished Usage. In: Boguraev, Branimir/Pustejovsky, James (eds.), *Corpus Processing for Lexical Acquisition*. Cambridge, MA: The MIT Press, 143–172.
- Wulff, Stefanie (2003), A Multifactorial Corpus Analysis of Adjective Order in English. In: *International Journal of Corpus Linguistics* 8(2), 245–282.

*Anatol Stefanowitsch, Bremen (Germany)
and Stefan Th. Gries, Santa Barbara, CA (USA)*

44. The induction of verb frames and verb classes from corpora

1. Introduction
2. Induction of verb frames from corpora
3. Induction of verb classes from corpora
4. Acknowledgments
5. Literature

1. Introduction

Creating lexical information resources manually is an expensive effort: it takes a long time to define detailed lexical knowledge, then the information needs to be updated regularly because of neologisms, sublanguages and language change, and the lexicon will rarely if ever be complete. For these reasons and also given the increasing availability of computing power and corpus resources, one line of research at the interface of corpus and computational linguistics aims at an automatic acquisition of lexical information, utilising existing corpora and applying computational algorithms. The retrieved lexical information is stored in machine-readable lexicons, and can be updated dynamically and quickly. Also, the resulting lexical resources can be integrated into computational tasks and applications in Natural Language Processing (NLP), such as parsing, machine translation, question answering, and many more.

Within the area of automatic lexical acquisition, the induction of lexical verb information has been a major focus, because verbs play a central role for the structure and the meaning of sentences and discourse. The levels of information that are relevant for a verb lexicon concern all lexical aspects of verbs, ranging from phonological and morpho-

- Stefanowitsch, Anatol/Gries, Stefan Th. (2003), Collostructions: Investigating the Interaction between Words and Constructions. In: *International Journal of Corpus Linguistics* 8(2), 209–243.
- Stefanowitsch, Anatol/Gries, Stefan Th. (2005), Covarying Collexemes. In: *Corpus Linguistics and Linguistic Theory* 1(1), 1–43.
- Stefanowitsch, Anatol/Gries, Stefan Th. (to appear), Register and Constructional Semantics: A Collostructional Case Study. In: Kristiansen, Gitte/Dirven, René (eds.), *Cognitive Sociolinguistics*. Berlin/Heidelberg/New York: Mouton de Gruyter.
- Stubbs, Michael (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.
- Waterman, Scott A. (1996), Distinguished Usage. In: Boguraev, Branimir/Pustejovsky, James (eds.), *Corpus Processing for Lexical Acquisition*. Cambridge, MA: The MIT Press, 143–172.
- Wulff, Stefanie (2003), A Multifactorial Corpus Analysis of Adjective Order in English. In: *International Journal of Corpus Linguistics* 8(2), 245–282.

*Anatol Stefanowitsch, Bremen (Germany)
and Stefan Th. Gries, Santa Barbara, CA (USA)*

44. The induction of verb frames and verb classes from corpora

1. Introduction
2. Induction of verb frames from corpora
3. Induction of verb classes from corpora
4. Acknowledgments
5. Literature

1. Introduction

Creating lexical information resources manually is an expensive effort: it takes a long time to define detailed lexical knowledge, then the information needs to be updated regularly because of neologisms, sublanguages and language change, and the lexicon will rarely if ever be complete. For these reasons and also given the increasing availability of computing power and corpus resources, one line of research at the interface of corpus and computational linguistics aims at an automatic acquisition of lexical information, utilising existing corpora and applying computational algorithms. The retrieved lexical information is stored in machine-readable lexicons, and can be updated dynamically and quickly. Also, the resulting lexical resources can be integrated into computational tasks and applications in Natural Language Processing (NLP), such as parsing, machine translation, question answering, and many more.

Within the area of automatic lexical acquisition, the induction of lexical verb information has been a major focus, because verbs play a central role for the structure and the meaning of sentences and discourse. The levels of information that are relevant for a verb lexicon concern all lexical aspects of verbs, ranging from phonological and morpho-

logical to syntactic, semantic, and pragmatic criteria. This article introduces work that focuses on the acquisition of lexical verb properties at the syntax-semantics interface, and addresses the automatic induction of verb frames from corpora (section 1), and the acquisition of verb classes (section 2). As is true for automatic lexical acquisition in general, the combination of corpus data and computational algorithms is not always straightforward, and we find a variety of solutions: corpus data can be used on various levels of annotation, i. e., as raw text, with part-of-speech tags, as parsed text with structural information, etc. (cf. the relevant articles in section IV of this book on preprocessing corpora); algorithms might fit a certain acquisition task better or worse, depending on the mathematical properties of the algorithm and how they relate to the linguistic task (cf. article 36); the acquisition results are often difficult to compare because they rely on different theoretical assumptions and produce different types of output, and there is not always an evaluation method available. In the course of this article, each section provides an overview of existing approaches to the induction of verb frames and verb classes, and describes their assumptions, procedures and evaluation.

2. Induction of verb frames from corpora

The potential of verbs to choose their complements (in this article, the term ‘complement’ is used to subsume the terms ‘argument’ and ‘adjunct’) is referred to as ‘verb subcategorisation’ or ‘verb valency’, and a combination of functional complements that are evoked by a verb is often called a ‘verb subcategorisation frame’, or simply a ‘verb frame’. For example, the verb ‘bake’ can subcategorise for a direct object (in addition to the obligatory subject), as in (1).

- (1) Elsa bakes a chocolate cake.

Alternatively, ‘bake’ might subcategorise for a direct object plus an indirect object, or a temporal prepositional phrase, as illustrated by (2) and (3), but cannot be used with e.g. a *that*-clause as in (4). With a different verb such as ‘say’ the frame in (4) would have been acceptable: ‘Elsa says that she likes cakes’.

- (2) Elsa bakes Tim a chocolate cake.
- (3) The chocolate cake baked for 1 hour.
- (4) *Elsa bakes that she likes cakes.

Verb frames might distinguish between obligatory and optional verb complements; however, this distinction is not always clear-cut, cf. the prepositional phrase (PP) in example (3): is the PP obligatory or optional? (See Meyers/Macleod/Grishman 1994 for one definition of the criteria for distinguishing between obligatory and optional verb complements.)

Typically, verb frames are illustrated as a set of the complements they include, such as $\{\text{Subj}, \text{Obj-dir}, \text{Obj-indir}\}$, or $\{\text{Subj}, \text{PP-for}\}$. Depending on the framework, the details of the frame description vary. For example, languages where subjects are obligatory need not explicitly include subjects in the verb frame; languages with case marking tend to

refer to the case of noun phrases instead of using direct vs. indirect objects; some approaches might distinguish PP arguments and adjuncts, others not; PPs can be referred to by a very general label (such as ‘PP’ only), by semantic category labels (such as ‘PP-tmp’, ‘PP-loc’), or by the specific preposition (such as ‘PP-for’, ‘PP-at’); etc.

Subcategorisation/valency is not restricted to the syntactic options of verb complements, but also refers to the semantic and the pragmatic level, cf. Helbig (1992) and Fischer (1999), among others. Example (5) presents a clause where ‘bake’ subcategorises for a direct object as in (1), but appears strange, because a stone is typically not baked. The degree of acceptability with respect to the semantic realisation of a complement varies, so a verb is said to define ‘selectional preferences’ for its complements.

- (5) ?Elsa bakes a stone.

Selectional preferences do not only refer to the syntactic function of a complement within a verb frame, but take into account the ‘semantic role’ of the respective complement, cf. the thematic proto-roles in Dowty (1991), and the argument structure in Grimshaw (1992). For example, the direct object in the causative transitive clause ‘Elsa melts the chocolate’ and the subject in the inchoative intransitive variant ‘The chocolate melted’ is the nominal phrase ‘the chocolate’, and this NP represents the patient role of the verb ‘bake’ in both variants. As the example illustrates, selectional preferences are required by the semantic roles of complements, which are in turn determined by the verb and the syntactic function of the complement within a certain verb frame. The phenomenon that verb frames and semantic roles can be used in alternative constructions is referred to as ‘diathesis alternation’; cf. Levin (1993) for a prominent collection of English verb alternations.

The induction of verb frames from corpora was one of the first issues when empirical lexical acquisition from corpora started out. The reason for this specific interest is that subcategorisation frames of verbs provide useful information for the structural analysis of sentences, which is necessary in e.g. parsing, cf. article 28. For example, Briscoe/Carroll (1993) found that half of all parse failures on unseen data are caused by inaccurate subcategorisation information. Later, the syntactic frame information in automatic lexical acquisition was gradually expanded to include semantic information referring to selectional preferences or semantic roles, and also definitions of diathesis alternations. What follows in this section is organised accordingly: section 1.1. describes approaches to acquiring syntactic verb frame types, and section 1.2. introduces extensions to the syntactic frame definitions.

2.1. Approaches to inducing subcategorisation frames

Approaches to automatically acquiring lexical verb information on subcategorisation frames can be defined and distinguished with respect to several dimensions, which we list here as (a) to (e):

- (a) *Corpus selection and preparation:* Which corpus is selected as the data resource, and what kind of annotation is provided?
- (b) *Frame types:* How many and which types of verb frames are distinguished?

- (c) *Acquisition method*: Which computational methods are used, in order to induce the subcategorisation frames?
- (d) *Filtering*: Are the subcategorisation frames as obtained under (c) filtered for noise, and what kind of method is used for filtering?
- (e) *Evaluation*: How is the resulting frame information evaluated?

In what follows, this section elaborates on the above criteria and exemplifies them by approaches to subcategorisation acquisition. The examples refer to representative (but not exhaustive) work for English; additional approaches for languages other than English follow.

Corpus selection

The first step (a) in frame acquisition is to select a corpus (cf. article 20 for an overview of corpora). As is true for empirical acquisition in general, researchers try to use as much data as are available (with respect to the language they are concerned with) and which can be processed with applicable computing resources. For example, in the early stages of subcategorisation acquisition, Brent (1993) used 2.6 million words of the Wall Street Journal corpus (WSJ), Ushioda et al. (1993) used 600,000 words of the same corpus, Manning (1993) used 4 million words of the New York Times newswire, and Briscoe/Carroll (1993) used 1.2 million words from the Susanne corpus, the Corpus of Spoken English (SEC), and the Lancaster-Oslo/Bergen Corpus (LOB). Comparing the early work with more recent approaches illustrates the increasing amount of available data and the decreasing restrictions on data processing. For example, Carroll/Rooth (1998) used the whole British National Corpus (BNC) with 100 million words.

In addition to the quantitative influence of the corpus size, the acquisition result is determined by the qualitative properties of the corpus, i. e., the genre of the corpus, the speech type (written vs. spoken corpora), the corpus age, etc. As a prominent example highlighting the influence of corpus choice on acquisition results, Roland/Jurafsky (2002) compared the frequencies of verb subcategorisation frames as obtained from five different corpora; two corpora were derived from psychological experiments in which participants were asked to produce single isolated sentences; two corpora were written texts, extracted from the Brown corpus and the Wall Street Journal corpus (Marcus/Marcinkiewicz/Santorini 1993), and one corpus contained telephone conversations. Roland/Jurafsky reported differences between the frame types and the frame type frequencies, and that the two major sources of the acquisition differences were (a) the discourse type, and (b) the semantic choices, i. e., the word senses represented in the corpora.

Corpus annotation

Methods for frame induction differ in the level of annotation they presuppose. Some approaches use raw corpus data, others either preprocess the corpus (i. e., lemmatiser/tagger/parser are applied to annotate the raw data), or use existing annotations provided by the corpus, such as the WSJ which is manually annotated on several levels (i. e., lemmas, part-of-speech (POS) tags, parse trees). For example, Brent (1993) performed frame acquisition from raw corpus data; Ushioda et al. (1993) used corpus data annotated with part-of-speech tags, and Manning (1993), Briscoe/Carroll (1997), as well as most subsequent work assumed partially or fully parsed corpus data. Recent work such as Kinyon/Prolo (2002) and O'Donovan et al. (2005) relied on the annotation provided by

a treebank (here: the Penn Treebank). Approaches which work on unannotated data are *a priori* more restricted in the linguistic details of their frame variants than approaches which take deeper morpho-syntactic information into account.

Frame types

Existing approaches differ strongly with respect to the desired number and types of subcategorisation frames they induce: are all available subcategorisation frames relevant, or are the frame types restricted to a subset? How fine-grained are the frame types, e.g. do they distinguish between different kinds of clauses or prepositional phrases? And do the approaches address the distinction between arguments and adjuncts, or generalise over the functions? For example, Brent's approach (1993) detected six frame types which only addressed direct objects and subcategorised clauses and infinitives. Ushioda et al. (1993) used a larger variety of complement types, but also restricted the experiments to six frame types, not distinguishing arguments and adjuncts. Manning (1993) defined 19 frame types with limited information on prepositions, still not distinguishing between arguments and adjuncts. Briscoe/Carroll (1997) acquired lexical information on 163 frame types, including the distinction between arguments and adjuncts, and a fine-grained reference to prepositional phrase types. Carroll/Rooth (1998) allowed all combinations of verb-adjacent constituents as frame types, as based on their context-free grammar parses.

Approaches which start out with no or few restrictions on frame types are more challenging (because they do not rely on existing frame information) but also more flexible than those which induce the syntax-semantics structure from a treebank. However, both classes of approaches are important; the more flexible approaches enable the induction of non-pre-existing, domain-independent subcategorisation lexicons and allow for unforeseen categories, and the treebank-based approaches are strong for theory-related, domain-specific subcategorisation knowledge.

As the overview of selected approaches illustrates, the degree of details within the frame types and – related to this – the number of types, depend on the underlying corpus information and also on individual decisions of which complements to include in the frames. It is important to note that there is no optimum with respect to size and details of the frame types. The ‘optimal’ subcategorisation lexicon depends on the NLP task/application which uses such a lexicon. For example, the argument/adjunct distinction is more relevant in machine translation (where it is important to relate the constituent functions between two languages in some detail) than in question answering (where the question/answer functions are rather generalised to enhance the query results).

Acquisition method

Another criterion in frame induction refers to the method for the lexical acquisition. Here we need to distinguish two steps, which may be interrelated: (i) the identification of the verbs in the corpus, and (ii) the identification and quantification of the frame types. The identification of the verbs is more or less difficult with respect to the level of corpus annotation that is accessed: raw corpus data provides less cues than part-of-speech tagged or even parsed corpus data. Also, languages with richer morphology (such as German) facilitate the detection of verbs, as compared to languages like English with poor morphology. Below, a series of methods is introduced, whose chronological order illustrates an increase in both the amount of corpus annotation and also the complexity of the acquisition approach.

For example, Brent (1993) identified English verbs in a raw corpus as all lexical items that appeared both with and without the suffix ‘-ing’. The result was filtered by heuristics on the lexical context, e. g. potential verbs which directly follow a determiner were not considered. The verb complements were identified by a finite-state grammar which defined linear patterns, such as ‘to V’ referring to a subcategorised infinitival clause. For its simplicity, Brent’s approach is surprisingly successful, but cannot be extended to sufficiently cover additional frame types, as no reliable cues exist for many frames.

Ushioda et al. (1993) used part-of-speech tags to identify verbs; to identify frame types, they defined a finite-state grammar for chunking, and regular expressions for linear chunk patterns. Like Brent’s approach above, the procedure is sufficient for a small number of frame types, but difficult to extend.

Manning (1993) used a finite-state parser to parse only clauses in the corpus with auxiliaries; relying on the restricted sentence structure, he identified the constituents following the verb as the verb complements. His approach is more reliable for a larger set of frame types, but restricts itself to a certain surface pattern, i. e., clauses with auxiliaries.

Later approaches made use of more complex corpus annotation: Briscoe/Carroll (1997) used the ranked output analyses of a probabilistic parser trained on a treebank, and extracted the verb and the subcategorised complements (plus their lexical heads) from the parses. Carroll/Rooth (1998) used a head-lexicalised probabilistic context-free grammar (HL-PCFG), trained the grammar with an unsupervised algorithm, and induced lexicalised subcategorisation information from the trained grammar model. The approaches of Briscoe/Carroll as well as Carroll/Rooth allowed all patterns which occurred according to their grammar, and derived the frame types according to these patterns.

Kinyon/Prolo (2002) defined a mapping from the Penn Treebank annotation to obtain verbs and their subcategorisation frame types. O’Donovan et al. (2005) derived their verb and frame information after they had performed an automatic annotation of the Penn Treebank with LFG (Lexical Functional Grammar) structures. The latter two approaches therefore relied on the definitions in a treebank to induce the verb-frame lexicon.

Filtering

Once the verb frame information is acquired, most approaches perform an additional step: they filter the empirical outcome. Brent (1993) suggested a hypothesis test to determine the reliable association between a verb and a frame type, referring to a binomial test. His filter was adapted in subsequent work, e. g., by Manning (1993) and Briscoe/Carroll (1997). However, Korhonen/Gorrell/McCarthy (2000) showed that a much simpler filter, which defines a cut-off on the relative frequency of a verb-frame pair, is sufficient and performs better than the hypothesis test. Following a different intuition, Korhonen (2002) – whose work built on Briscoe/Carroll (1997) – suggested a filter based on verb semantics: she demonstrated the usefulness of semantic verb classes (cf. section 2) by smoothing the empirical subcategorisation frames with back-off estimates on the verbs’ semantic classes by Levin (1993), and subsequently applied a simple threshold to the estimates. The range of approaches to filtering illustrates that, on the one hand filtering is seen as an important step after the acquisition procedure, but on the other hand the type and the complexity of the filtering methods differ greatly.

Evaluation

There are multiple possibilities for evaluating the empirical frame information. Existing approaches performed either (a) manual judgement, or (b) an evaluation against frame types listed in existing manually built lexicons, or (c) an evaluation by integrating the frame information into an NLP task or application. For example, Brent (1993) evaluated his English subcategorisation frames by hand judgement, reporting an f-score of 73.85 %. Wauschkuhn (1999) did the same for German on a choice of seven verbs. He reported an f-score of 61.86 %. Manning (1993) and Carroll/Rooth (1998) evaluated their frames against the *Oxford Advanced Learner's Dictionary* (Hornby 1985) and reported an f-score of 58.20 % and 76.95 %, respectively. Briscoe/Carroll (1997) used the *Alvey NL Tools Dictionary* (Boguraev et al. 1987) and the *COMLEX Syntax Dictionary* (Grishman/Macleod/Meyers 1994) and achieved a 46.09 % f-score. O'Donovan et al. (2005) also used COMLEX, achieving 27.3 %/64.3 % with/without PP specifications. Schulte im Walde (2002a, 2002b) used the German dictionary *Duden – Das Stilwörterbuch* (Bibliographisches Institut & F. A. Brockhaus AG), reporting an f-score of 57.24 %/62.30 % with/without PP specifications.

Even though the f-scores above suggest qualitative differences between the various approaches, it is important to note that a comparison between the results is difficult. Generally speaking, a manual inspection of the frame results has the advantage of being more flexible than comparing the frame information against pre-defined dictionary entries. However, in the case of manual judgement being performed by only one person, it runs the risk of being subjective, so it is necessary to rely on multiple annotators for evaluation. On the other hand, an evaluation against dictionaries is bound to the assumptions and definitions in a dictionary, which might differ from those in the lexical acquisition approach: the dictionary provides more details in some cases but less details in others, with respect to the kinds of complements included in the frame types (such as the subcategorisation of clauses, or arguments vs. adjuncts), and the granularity of the complement properties (such as the kinds of prepositional phrases, or case information in morphologically rich languages). Last but not least, the automatic induction of verb frames is more or less difficult with regard to how many different frame types the approaches call for, and how fine-grained the frame type information is. In conclusion, one should only compare the outcome of approaches whose target frame types are sufficiently similar, and whose evaluation methods are comparable to a large extent. The f-score results reported above should therefore be interpreted with caution.

Integrating a subcategorisation lexicon within an application is one way to compare the outcomes of various approaches, because the improved performance of the application when using subcategorisation information can be measured. This method has rarely been employed for comparison reasons, because the induced frame lexicons usually differ with respect to their original purpose, and also a comparison across languages is difficult. However, a few research groups did apply their result to NLP tasks or applications. Work based on Briscoe/Carroll (1997, 2002) applied subcategorisation information to improve the coverage and accuracy of parsing systems: Carroll/Minnen/Briscoe (1998) used the 1997-system and showed that subcategorisation information significantly improved the accuracy of their wide-coverage parser in inducing grammatical relations. Carroll/Fang (2004) extended the 2002-system with a subcategorisation lexicon and showed that the extension helped a deep HPSG parser to improve its coverage and the parsing success rate.

Languages

The above criteria and approaches illustrate the variety of frame acquisition ideas, and also the development over time, utilising an increasing amount of data and defining more complex algorithms and filters. So far, we have mainly referred to approaches for English. But we also find approaches to the automatic induction of syntactic frames in languages other than English: for German, Eckle-Kohler (1999) performed a semi-automatic acquisition of subcategorisation information for 6,305 verbs. She worked on POS-tagged corpus data and defined linguistic heuristics by regular expression queries over the usage of 244 frame types including PP definitions. Wauschkuhn (1999) constructed a valency dictionary for 1,044 German verbs. He extracted a maximum of 2,000 example sentences for each verb from annotated corpus data, and constructed a context-free grammar for partial parsing. The syntactic analyses provided valency patterns, which were grouped in order to extract the most frequent pattern combinations, resulting in a verb-frame lexicon with 42 frame types. Schulte im Walde (2002a) developed a German context-free grammar containing frame-predicting grammar rules, and used the unsupervised training environment of HL-PCFGs (Carroll/Rooth 1998) to train the grammar on 18.7 million words of German newspaper corpora. She induced subcategorisation frames for more than 14,000 German verbs, for 38 purely syntactic frame types and a refinement of 178 frame types including prepositional phrase distinctions. For Portuguese and Greek, de Lima (2002) and Georgala (2003), respectively, also utilised the same HL-PCFG framework to learn verb-frame combinations for the respective languages. For Czech, Sarkar/Zeman (2000) used the syntactic dependency definitions in the Prague Dependency Treebank (PDT) to induce subcategorisation frames. A frame was defined as a subset of the annotated dependents of a verb in the treebank. As a major task, they learned the argument-adjunct distinction in the frame types. For Dutch, Spranger/Heid (2003) developed a chunker to extract verb subcategorisation. For French, Chesley/Salmon-Alt (2006) created a multi-genre corpus with random occurrences of 104 frequent verbs from the Frantext online literary database. They applied a dependency-parser to the corpus and obtained 27 different subcategorisation frames in the form of any combination of a restricted set of constituents (direct objects, pre-specified PPs, clauses, adjectival phrases, and reflexive clitic NPs).

In addition to the ‘usual’ differences between the approaches to subcategorisation acquisition (cf. the discussions of (a) to (e) above), the approaches across languages naturally differ as a consequence of the properties of the respective languages, such as morphological marking, word order, etc. For example, Carroll/Rooth (1998) defined a HL-PCFG with flat grammar rules for English, whereas Schulte im Walde (2002a) – who used the same framework for German – defined mostly binary context-free rules because the freer word order in German would have required an enormous amount of flat rules (covering all possible constellations of complement orderings, combined with adverbial modifiers, etc.), creating training problems for a lexicalised grammar, and possibly causing a sparse-data problem. Another example of language differences concerns the relevance of complement types within the verb frames. For example, in some languages subjects are obligatory (e. g., English, German), whereas in others (e. g., Italian, Spanish) they are not, so the relevance of including subject information in the frames differs; also, certain complements such as adjectival phrases are of minor (e. g., in German) vs. major (e. g., in French) importance with respect to their productivity.

2.2. Approaches to empirical extensions of verb frames

So far, we have discussed the induction of purely syntactic verb frames, plus refinement by prepositional phrase types in some approaches. But syntactic frames are only one part of verb subcategorisation, as mentioned above. In this section, we address the acquisition of verb frames with additional semantic subcategorisation, i.e., we introduce approaches which empirically define selectional preferences or semantic roles for verb frames. In addition, we refer to approaches which build on the induction of syntactic and semantic subcategorisation, and address the diathesis alternation of verbs, namely the alternative usage of frames and roles.

Selectional preferences

As demonstrated in section 1 of this article, the degree of acceptability with respect to the semantic realisation of a complement varies, so a verb is said to define ‘selectional preferences’ for its complements. From a practical point of view, selectional preferences for complements are useful because they refer to a generalisation of specific complement heads and therefore improve a sparse-data situation. For example, in lexicalised parsing, the lexical heads that are incorporated into the parser cause a sparse-data problem. Referring to selectional preferences instead of specific lexical heads might help the parser because if it is confronted with e.g. ‘drink a beverage’, then it can abstract from seen instances such as ‘drink tea’, ‘drink coffee’, etc. to unseen instances such as ‘drink cocoa’.

In order to define selectional preferences for frame complements, it is necessary to refer to an inventory of semantic categories, such as ‘animate’ vs. ‘inanimate’, or ‘banana’ vs. ‘teacher’, etc. The choice of semantic categories and the level of granularity depend on the theoretical assumptions of the researchers, and in practice the categories are often restricted to the definitions in an existing resource. The reason is that, on the one hand we demand a generalisation over nominal complements in order to talk about abstract preferences, but on the other hand we do not *a priori* find generalisations in corpus data. So it is helpful to refer to an external categorisation. The following example approaches utilise ‘WordNet’ (Fellbaum 1998), a lexical semantic taxonomy originally developed for English at the University of Princeton, and since then transferred to additional languages, cf. the Global WordNet Association (www.globalwordnet.org) for more details. The lexical database was inspired by psycholinguistic research on human lexical memory. It organises nouns, verbs, adjectives and adverbs into classes of synonyms (‘synsets’). Words with several senses are assigned to multiple classes. The synsets are connected by lexical and conceptual relations such as hypernymy, hyponymy, meronymy, etc. The hypernym-hyponym relation imposes a multi-level hierarchical structure on the taxonomy. The noun synsets in WordNet provide a choice of semantic categories on different levels of generalisation, which can be used to define selectional preferences for verbs. For example, the verb ‘drink’ would specify a strong selectional preference for the WordNet synset ‘beverage’ with respect to its direct object, intuitively because on the one hand the synset generalises over its hyponyms such as ‘coffee’, ‘tea’, ‘milk’, etc., and on the other hand it is more specific to the verb’s complements than its hypernyms such as ‘food’, or even ‘substance’.

In the following, a choice of approaches which utilise WordNet is presented. As described above, WordNet provides a framework that is suitable for defining selectional preferences, and has therefore been used extensively for this task. The selection of ap-

proaches is by far not exhaustive, but provides an overview and pointers to more information on selectional preference acquisition, with and without WordNet.

Resnik (1997) defined selectional preference as the association strength between a predicate and the semantic categories of its complements. The starting point of his approach was co-occurrence counts of predicates and complements within a specific syntactic relationship (such as ‘direct object’), as based on a corpus. The co-occurrence counts were assigned to those WordNet synsets which contain the respective heads of the complements, and propagated upwards in the WordNet hierarchy. For ambiguous complements, the count was split over all WordNet synsets containing that complement. This procedure was repeated for all complements, and the counts were accumulated for each synset. Furthermore, the procedure was applied twice: (a) for each specific predicate of interest, e. g., for specific verbs, and (b) without relation to a specific predicate, i. e., accumulating over a class of predicates such as all verbs in the corpus. The association strength was then calculated by applying the information-theoretic measure of relative entropy to the two probability distributions based on the complement counts over WordNet synsets: The *prior* probability of a complement class (i. e., a WordNet synset such as ‘beverage’) *regardless* of the identity of the predicate is compared with the *posterior* probability of a complement class *with regard* to a specific predicate. Relative entropy calculates the distance between the respective probability distributions; the more similar the two probability distributions are, the weaker the association between predicate and complement class, and therefore the weaker the selectional preference of the predicate for that class.

Li/Abe (1998) also based their approach on co-occurrence counts of predicates and complements within a specific syntactic relationship. The selectional preferences for a predicate-complement structure were described by a cut in the WordNet hierarchy, i. e., a set of WordNet nodes. The cut was determined by the Minimum Description Length (MDL), a principle from information theory for data compression and statistical estimation. A selectional preference model where the chosen set of WordNet nodes is nearer the WordNet root is simpler to describe (by means of the number of bits for encoding the model) but with a poorer fit to the data, i. e., the specific WordNet leaves; a model nearer the WordNet leaves is more complex but with a better fit to the data. The MDL principle finds the cut in the hierarchy which minimises the sum of encoding both the model and the data.

Abney/Light (1999) provided a stochastic generation model to determine the selectional preferences of a predicate-complement relationship. The co-occurrence probabilities were estimated by a Hidden Markov Model (HMM) for each predicate structure. The HMM was defined and trained on the WordNet hierarchy, with the initial state being the (artificial) root node of WordNet. Each HMM run was a path through the hierarchy from the root to a word sense, plus the word generated from the word sense. The most likely path indicated the verbs’ selectional preferences.

Clark/Weir (2002) estimated the joint frequencies for a predicate-complement relationship and a specific WordNet class in the same way as Resnik (1997). Their generalisation procedure then used the statistical chi square test to find the most suitable class: a bottom-up check of each node in the WordNet hierarchy determined whether the probability of the parent class was significantly different from that of the children classes. In that case, the search was stopped at the respective child node as the most suitable selectional preference representation.

Even though the above approaches used different algorithms to calculate selectional preferences, they all rely on similar data (verb-complement co-occurrence counts from a chunker or a parser) and attempt to characterise the selectional preferences of a verb by WordNet noun synsets. *A priori*, it is difficult to tell whether any of the approaches is optimal. Brockmann/Lapata (2003) therefore compared the approaches of Resnik, Li/Abe, and Clark/Weir with respect to German verbs and their NP and PP complements, using a common corpus. The models, as well as a combination of the models, were evaluated against human ratings, demonstrating that there was no method which performs best overall. They added a model combination, using multiple linear regression, and the combined method actually obtained a better fit to the experimental data than the single methods. The comparison demonstrates that it is not necessarily the case that one approach outperforms all other approaches. Rather, it is important to compare the variety of approaches with respect to a certain task, or even try to find combinations that complement each other.

Semantic roles

A second current of adding semantic information to subcategorisation frames is concerned with the definition of semantic roles for complements. Differently to selectional preferences, semantic roles are not generalisations of lexical heads, but represent the semantic relationship between a predicate and a complement within a certain frame type. To refer back to our example in section 1, the NP ‘the chocolate’ represents the patient role of the direct object in the transitive clause ‘Elsa melts the chocolate’ and also of the subject in the inchoative intransitive variant ‘The chocolate melted’. In practical terms, semantic roles are useful in applications such as question answering, where e. g. a question word such as ‘who’ in ‘who killed ...’ needs to be matched to an agent role for the verb ‘kill’, abstracting over syntactic functions and lexical heads.

As for selectional preference acquisition, we do not *a priori* find semantic roles in corpus data; thus, the approaches to semantic role labeling attempt to induce regularities from unlabeled data, or rely on manually annotated data. In the following, two prominent projects concerned with semantic role labeling are introduced, ‘FrameNet’ and ‘PropBank’. Within the projects, corpora are annotated with semantic information; the annotation is partly manual and partly semi-automatic; the semi-automatic labeling explores unsupervised methods for role labeling. The annotated data can be used for supervised approaches to learning semantic subcategorisation.

FrameNet (Baker/Fillmore/Lowe 1998) is based on Fillmore’s frame semantics (Fillmore 1982) and thus describes ‘frames’, i. e., the background and situational knowledge needed to understand a word or expression. Each FrameNet frame provides its set of semantic roles, the participants and properties of the prototypical situation. The Berkeley FrameNet project is building a dictionary which links their frames to the words and expressions that introduce them, illustrating them with example sentences from the British National Corpus. FrameNet started out for English, but there is already cross-lingual transfer of the framework to German (Erk et al. 2003), Spanish (Subirats/Sato 2004), and Japanese (Ohara et al. 2004).

The PropBank project (Palmer/Gildea/Kingsbury 2005) is creating a corpus of text annotated with information about semantic roles by adding a layer of predicate-complement relations to the syntactic structures of the Penn Treebank. In contrast to FrameNet, PropBank defines semantic roles on a per-verb basis, but not across verbs. The

PropBank is designed as a broad-coverage resource, covering every instance of a verb in the corpus, to facilitate the development of more general NLP systems.

Whole lines of research on semantic roles (partly working on the above databases) have been advanced via the framework of recent and ongoing shared tasks, i. e., competitions where the organisers define a task (and provide the necessary data) in order to compare different approaches to that specific task. In *Senseval* (www.senseval.org) the task is word sense disambiguation. Rich data sets with deep syntactic information are provided for this task, which started out in 1998 (cf. article 26 section 5). Also, within the Conference on Natural Language Learning, the shared task was devoted to semantic role labeling in some of the events.

Diathesis alternation

Diathesis alternations concern the (systematic) alternative use of frames and semantic roles. Thus, after having assigned semantic information to verb frames, the next natural step is a study of diathesis alternations. Sentences (6) and (7) illustrate an example of diathesis alternation, namely the *benefactive alternation* in English, cf. Levin (1993).

- (6) Martha carved a toy for the baby.
- (7) Martha carved the baby a toy.

The benefactive alternation is characterised by an alternation between (i) a transitive frame plus a benefactive *for-PP*, and (ii) a double object frame; in addition, the semantic categories of the direct objects in (6) and (7) overlap, as well as the semantic categories of the *for-PP* in (6) and the indirect object in (7). The alternation is called *systematic*, since it applies to a range of semantically similar verbs, cf. Apresjan (1973). For example, the benefactive alternation transfers to other *build verbs* such as ‘bake’ and ‘cook’, and *preparation verbs* such as ‘pour’ and ‘prepare’. This property of regularity makes diathesis alternations an important issue for the creation of verb classes, cf. section 2.

Even though a large number of approaches have been concerned with the automatic acquisition of syntactic subcategorisation and there is a substantial amount of work devoted to semantic labeling, few approaches exist for inducing diathesis alternations, and most of these are concerned with case studies. This is probably because an *explicit* definition of diathesis alternations is rarely necessary, while an *implicit* definition (acquiring and applying syntactic subcategorisation combined with semantic subcategorisation) is usually sufficient in the relevant NLP tasks. In the following, three example approaches to the explicit learning of diathesis alternation are presented.

McCarthy (2001) introduced a method to identify which English verbs participate in a diathesis alternation. In a first step, she used the subcategorisation frame acquisition system of Briscoe/Carroll (1997) to extract frequency information on subcategorisation frame types for verbs from the BNC. The subcategorisation frame types were manually linked with the Levin alternations, and thereby defined the verbal alternation candidates. Following the acquisition of the syntactic information, the nominal fillers of the NP and PP complements in the verb-frame tuples were used to define selectional preferences for the respective complement slots. For this step, McCarthy utilised the selectional preference acquisition approach of Minimum Description Length of Li/Abe (1998). In the final step, McCarthy defined two methods to identify the participation of verbs in diathesis alternations: (i) The MDL principle compared the costs of encoding the tree cut models

of selectional preferences for the relevant complement slots in the alternation frames. If the cost of combining the models was cheaper than the cost of the separate models, the verb was classified as undergoing the respective alternation. (ii) A similarity-based method calculated the similarity of the two tree cut models with reference to the alternating complement slots for verbs that participated in diathesis alternations. A threshold decided the participation.

Lapata/Brew (2004) performed a case study on the induction of diathesis alternations, studying the dative and benefactive alternation for English verbs. They used a shallow parser to identify verb frames and their frequencies in the BNC, and defined a simple probabilistic model to generate preferences for the Levin classes.

Tsang/Stevenson (2004) based their model of diathesis alternation on distributional similarity between WordNet trees, rather than WordNet classes. The WordNet nominal trees were activated by probability distributions over verb-frame-noun pairs, and standard similarity measures determined the similarity of verb-frame alternations. A threshold defined the participation in an alternation; the work was a case study on the causative alternation.

The three above approaches are difficult to compare because they focus on different alternations and are not evaluated on common data. Tsang/Stevenson introduced their approach as an enhancement of McCarthy's and showed that their results outperformed the previous approach in the general case (i.e., when applied to random rather than hand-selected data). For empirical linguistics it might be interesting to see further developments of explicit approaches to automatically detect diathesis alternations, especially for languages other than English.

3. Induction of verb classes from corpora

Verb classes categorise verbs into classes such that verbs in the same class are as similar as possible, and verbs in different classes are as dissimilar as possible; the kind of similarity is defined by the creators of the verb classes. For example, syntactic verb classes categorise verbs according to syntactic properties of interest, semantic verb classes categorise verbs according to semantic properties of interest, etc. From a practical point of view, verb classes reduce redundancy in verb descriptions, since they refer to the common properties of the verbs in the classes; in addition, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, by referring to verbs in the same class: under this criterion, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is available for rare events.

This section is concerned with the *automatic* creation of verb classes, which is supposed to avoid tedious *manual* definitions of the verbs and the classes. The outcome of the creation process depends on several factors, which are summarised as follows:

- the purpose of the classification,
- the choice of the verbs of interest,
- the definition of features that describe the verb properties of interest and can be obtained from corpora,

- the choice of an algorithm for class formation and verb assignment to classes, and
- the evaluation of the resulting classification.

In the remainder of this section, we address these parameters. Section 2.1. provides an overview of different types of verb classes, section 2.2. presents approaches to the automatic creation of verb classes, and section 2.3. addresses the evaluation of classifications. As mentioned above, we focus on the empirical acquisition of verb classes and only occasionally refer to manual classifications.

3.1. Types of verb classes

Even though one could think of various linguistic properties for the classification of verbs, much work on the automatic induction of verb classes has concentrated on verb classes at the syntax-semantics interface. An important reason for this is that few corpora are semantically annotated and provide semantic annotation off-the-shelf (such as FrameNet (Baker/Fillmore/Lowe 1998) and PropBank (Palmer/Gildea/Kingsbury 2005), cf. section 1.2.). Instead, the automatic construction of syntax-semantics verb classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its behaviour: to a certain extent, the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its complements, cf. Pinker (1989) and Levin (1993), among others. Even though the meaning-behaviour relationship is not perfect, the following prediction is used: if a verb classification is induced on the basis of features describing verb behaviour, then the resulting behaviour-classification should agree with a semantic classification to a certain extent. From a practical point of view, such verb classes have successfully been applied in NLP. For example, the English verb classification by Levin (1993) was used in NLP applications such as word sense disambiguation (Dorr/Jones 1996), machine translation (Dorr 1997), document classification (Klavans/Kan 1998), and subcategorisation acquisition (Korhonen 2002). In the following, individual approaches to acquire verb classes at the syntax-semantics interface are introduced with respect to their target classification and the choice of features used to empirically model the verb properties of interest.

Brent (1991) and Siegel (1998) described approaches to aspectual verb classes, distinguishing between states and events. Both approaches chose features that were indicators of verbal aspect: Brent used syntactic cues such as occurrences of the progressive and adverbial constructions in the verb context; Siegel used a more extensive set of 14 linguistic indicators including Brent's cues, and adding e. g. tense distinctions and prepositional phrases indicating a duration.

A major line of approaches to verb classes at the syntax-semantics interface induced empirical information on verb behaviour from corpora, focusing on subcategorisation frames, prepositional phrases, semantic categories of complements, and alternation behaviour, in line with section 1. For example, Dorr/Jones (1996) extracted the syntactic patterns from Levin's class descriptions (distinguishing positive and negative instances), and showed that these patterns correspond closely to the affiliation of the verbs with their semantic classes. Merlo/Stevenson (2001) approached three verb classes – unergative, unaccusative, and object-drop verbs – and defined verb features that rely on linguistic heuristics to describe the thematic roles of subjects and objects in transitive and

intransitive verb usage. The features included heuristics for transitivity, causativity, animacy, and syntactic features. For example, the degree of animacy of the subject roles was estimated as the ratio of occurrences of personal pronouns to all subjects for each verb, based on the assumption that unaccusatives occur less frequently with an animate subject when compared to unergative and object-drop verbs. Joanis (2002) and Joanis/Stevenson (2003) presented an extension of their work that approached 14 Levin classes. They defined an extensive feature space including part-of-speech, auxiliary frequency, syntactic categories and animacy, plus selectional preference features taken from WordNet. Stevenson/Joinis (2003) then applied various approaches to automatic feature selection in order to reduce the feature set to the relevant features, addressing the problem of too many irrelevant features. They reported a semi-supervised chosen set of features based on seed verbs (i. e., representative verbs for the verb classes) as the most reliable choice.

Schulte im Walde (2000, 2006) described English/German verbs according to the probabilities of subcategorisation frames including prepositional phrase types, plus selectional preferences referring to the WordNet/GermaNet top-level synsets. The classification target was semantic verb classes such as ‘manner of motion’, ‘desire’, and ‘observation’.

Esteve Ferrer (2004) acquired verb properties referring to syntactic subcategorisation frames; the target classification referred to the manual Spanish verb classes developed by Vázquez et al. (2000), with the three semantic classes ‘trajectory’, ‘change’, and ‘attitude’, subdivided into 31 subclasses. The Spanish verb classes were similar to Levin’s English classes, but grouped together different subclasses.

Merlo et al. (2002) and Tsang/Stevenson/Merlo (2002) introduced a multi-lingual aspect to the work by Merlo/Stevenson (2001). Merlo et al. (2002) showed that the classification paradigm was applicable to other languages than English by using the same features as defined by Merlo/Stevenson (2001) for the respective classification of Italian verbs. Tsang/Stevenson/Merlo (2002) used the content of Chinese verb features to refine the English verb classification: the English verbs were manually translated into Chinese and given part-of-speech tags, passive particles, causative particles, and sublexical morphemic properties. Verb tags and particles in Chinese are overt expressions of semantic information that is not expressed as clearly in English. The multi-lingual set of features outperformed either set of monolingual features. The multi-lingual work demonstrates that a) there are features that are useful for the task of verb class acquisition cross-linguistically, and b) an existing feature set in this framework can be extended and improved by exploiting features from a different language.

The overview of the selected approaches illustrates that there are various types of syntax-semantics target classifications, and that the chosen verb features vary accordingly across the targets. On the one hand, a core of features (such as subcategorisation frames) has established itself within the syntax-semantics descriptions; on the other hand, the choice and extraction of empirical features from corpora for verb class creation is still developing.

3.2. Approaches to acquiring verb classes

Based on the verb descriptions introduced in the previous section, approaches to acquiring verb classes have used various supervised or unsupervised methods (cf. article 40) to decide on class membership. For example, Brent (1991) simply defined a confidence

interval for his cue frequencies, and a threshold to decide between a stative and an event verb. Siegel (1998), by comparison, applied three supervised machine learning algorithms (logistic regression, decision trees, genetic programming) to his aspectual classification, plus an unsupervised partitioning algorithm which was based on a random assignment and improved by a greedy search.

Most work in the tradition of Merlo and Stevenson (Merlo/Stevenson 2001; Joanis 2002; Merlo et al. 2002; Tsang/Stevenson/Merlo 2002) used decision trees to establish the verb classes. Schulte im Walde (2000), Stevenson/Joanis (2003) as well as Esteve Ferrer (2004) performed unsupervised clustering, applying agglomerative hierarchical approaches. Schulte im Walde (2006) partitioned verbs into classes by using the unsupervised iterative k-Means algorithm. Even though different classification and clustering approaches were applied to a similar task, it is difficult to compare the above approaches, since none of them were evaluated on common data sets.

So far, few approaches have addressed the polysemy of verbs by using soft-clustering algorithms and multiple assignment of verbs to classes. For example, Rooth et al. (1999) produced soft semantic clusters for English which at the same time represented a classification of verbs as well as of nouns. The conditioning of the verbs and the nouns on each other was made through hidden classes and the joint probabilities of classes. Verbs and nouns were trained by the Expectation-Maximisation (EM) algorithm. The resulting model defined conditional membership probabilities for each verb and noun in each class.

Korhonen/Krymolowski/Marx (2003) used the Information Bottleneck, an iterative soft clustering method based on information-theoretic foundations, to cluster verbs with possible multiple senses. They reported that polysemic verbs with a clear predominant sense or regular polysemy were frequently clustered together. Homonymic verbs or verbs with strong irregular polysemy tended to resist any classification.

Last but not least, we find whole projects devoted to the creation of (verb) classes. Prominent examples are – as introduced in section 1.2. – WordNet (Fellbaum 1998), which organises English nouns, verbs, adjectives and adverbs into classes of synonyms, and FrameNet (Baker/Fillmore/Lowe 1998), which assigns English verbs, nouns and adjectives to FrameNet frames, referring to common situational knowledge. Even though much of the work in these and other projects is performed manually, selected issues are supported by (semi-)automatic methods.

3.3. Evaluation of verb classes

There is no absolute scheme for automatically evaluating the induced verb classifications. A variety of evaluation measures from diverse areas such as theoretical statistics, machine vision, web-page clustering and coreference resolution do exist, but so far, no generally accepted method has been established. We can distinguish two currents of evaluation methods: (i) methods which address how well the data underlying the verb descriptions are modelled by the resulting classification, and (ii) methods which compare the resulting classification against a gold standard.

The silhouette value (Kaufman/Rousseeuw 1990) represents an example of type (i), evaluating the modelling of the data. It measures which verbs lie well within a class and which verbs are marginal to a class by comparing the verbs' distances to verbs in the

same class with the distances to verbs in the neighbour class. The distances between verbs in the same class should be smaller than between verbs in different classes; the data are thus well separated by the clustering result. Stevenson/Joanis (2003) and Esteve Ferrer (2004) applied this evaluation as one measure of their classification quality.

Evaluation methods of type (i) do not assess whether the clustering result resembles a desired verb classification. By contrast, when applying an evaluation method of type (ii), one needs a gold standard resource of verb classes to compare the clustering result with. Most approaches so far have referred to hand-crafted small-scale verb classes which they developed for the purpose of evaluation. Large-scale resources are rare; two instances for English are the Levin classes and WordNet. But even with a gold standard at hand the evaluation task is still difficult, because there are various ways to compare the two sets of classes. Questions such as the following are difficult to answer: how to map the classes within the two sets onto each other, especially when the number of classes is different; whether an evaluation of classes can be reduced to an evaluation of the verb pairs within the classes; how to deal with ambiguity; etc. Schulte im Walde (2003, chapter 4) performed an extensive comparison of various evaluation methods against a gold standard, referring not only to general classification criteria, but also to the task-specific linguistic demands. She determined three evaluation measures to be the most appropriate ones to apply: (a) the f-score of a pair-wise precision and recall measure, (b) an adjusted pair-wise precision measure, and (c) the adjusted Rand index. (a) The pair-wise precision and recall measure goes back to a suggestion by Hatzivassiloglou/McKeown (1993), who performed an automatic classification of English adjectives and calculated precision and recall based on common class membership of adjective pairs in the automatic and the gold standard classification. (b) Since the recall value shows strong class size biases, Schulte im Walde/Brew (2002) focused on the precision value and adjusted it by a scaling factor based on the size of the respective verb class. This adjusted pair-wise precision measure (APP) was applied to evaluating verb classes by Schulte im Walde/Brew themselves, and Korhonen/Krymolowski/Marx (2003). (c) The adjusted Rand index (Hubert/Arabie 1985) also measures the agreement between verb pairs in the classes, but is corrected for the chance compared to the null model that the classes were constituted at random, given the original number of classes and verbs. This measure was applied by Schulte im Walde (2003, 2006), Stevenson/Joanis (2003), and Esteve Ferrer (2004).

The example evaluations illustrate that there is still a need for a generally accepted evaluation method. However, it is also clear that the different approaches to verb class induction have started to agree on a selection of measures.

4. Acknowledgments

Many thanks to Pia Knöferle, Anna Korhonen, Anke Lüdeling, Alissa Melinger, Sebastian Padó, Nils Reiter, Kristina Spranger, Suzanne Stevenson, and two anonymous reviewers for their feedback on earlier versions of this article.

5. Literature

- Abney, Steven/Light, Marc (1999), Hiding a Semantic Class Hierarchy in a Markov Model. In: *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*. College Park, MD, 1–8.
- Apresjan, Jurij D. (1973), Regular Polysemy. In: *Linguistics* 142, 5–32.
- Baker, Collin/Fillmore, Charles/Lowe, John (1998), The Berkeley FrameNet Project. In: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*. Montreal, Canada, 86–90.
- Boguraev, Branimir/Briscoe, Ted/Carroll, John/Carter, David/Grover, Claire (1987), The Derivation of a Grammatically-indexed Lexicon from the *Longman Dictionary of Contemporary English*. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Stanford, CA, 193–200.
- Brent, Michael R. (1991), Automatic Semantic Classification of Verbs from their Syntactic Contexts: An Implemented Classifier for Stativity. In: *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*. Berlin, Germany, 222–226.
- Brent, Michael R. (1993), From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. In: *Computational Linguistics* 19(2), 243–262.
- Briscoe, Ted/Carroll, John (1993), Generalized Probabilistic LR Parsing for Unification-based Grammars. In: *Computational Linguistics* 19(1), 25–60.
- Briscoe, Ted/Carroll, John (1997), Automatic Extraction of Subcategorization from Corpora. In: *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*. Washington, DC, 356–363.
- Briscoe, Ted/Carroll, John (2002), Robust Accurate Statistical Annotation of General Text. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain, 1499–1504.
- Brockmann, Carsten/Lapata, Mirella (2003), Evaluating and Combining Approaches to Selectional Preference Acquisition. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, 27–34.
- Carroll, Glenn/Rooth, Mats (1998), Valence Induction with a Head-lexicalized PCFG. In: *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*. Granada, Spain, 36–45.
- Carroll, John/Fang, Alex (2004), The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In: *Proceedings of the 1st International Joint Conference on Natural Language Processing*. Sanya City, China, 107–114.
- Carroll, John/Minnen, Guido/Briscoe, Ted (1998), Can Subcategorisation Probabilities Help a Statistical Parser? In: *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*. Montreal, Canada, 118–126.
- Chesley, Paula/Salmon-Alt, Susanne (2006), Automatic Extraction of Subcategorization Frames for French. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, 253–258.
- Clark, Steven/Weir, David (2002), Class-based Probability Estimation Using a Semantic Hierarchy. In: *Computational Linguistics* 28(2), 187–206.
- Dorr, Bonnie J. (1997), Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. In: *Machine Translation* 12(4), 271–322.
- Dorr, Bonnie J./Jones, Doug (1996), Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, 322–327.
- Dowty, David (1991), Thematic Proto-roles and Argument Selection. In: *Language* 67, 547–619.
- Eckle-Kohler, Judith (1999), *Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora*. Berlin: Logos Verlag.
- Erk, Katrin/Kowalski, Andrea/Padó, Sebastian/Pinkal, Manfred (2003), Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In: *Proceed-*

- ings of the 41st Annual Meeting of the Association for Computational Linguistics.* Sapporo, Japan, 537–544.
- Esteve Ferrer, Eva (2004), Towards a Semantic Classification of Spanish Verbs Based on Subcategorisation Information. In: *Proceedings of the Student Research Workshop at the Annual Meeting of the Association for Computational Linguistics.* Barcelona, Spain, 37–42.
- Fellbaum, Christiane (ed.) (1998), *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.
- Fillmore, Charles (1982), Frame Semantics. In: *Linguistics in the Morning Calm*, 111–137.
- Fischer, Klaus (1999), Verb Valency – an Attempt at Conceptual Clarification. In: *The Web Journal of Modern Language Linguistics* 4–5. Published by the School of Modern Languages, University of Newcastle upon Tyne. Available at: <http://wjml.ncl.ac.uk/issue04-05/fischer.htm>
- Georgala, Effi (2003), A Statistical Grammar Model for Modern Greek: The Context-free Grammar. In: *Proceedings of the 24th Annual Meeting of the Linguistics Department of the Aristotle University of Thessaloniki.* Thessaloniki, Greece, 183–193.
- Grimshaw, Jane B. (1992), *Argument Structure.* Cambridge: The MIT Press.
- Grishman, Ralph/Macleod, Catherine/Meyers, Adam (1994), COMLEX Syntax: Building a Computational Lexicon. In: *Proceedings of the 15th International Conference on Computational Linguistics.* Kyoto, Japan, 268–272.
- Hatzivassiloglou, Vasileios/McKeown, Kathleen R. (1993), Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.* Columbus, OH, 172–182.
- Helbig, Gerhard (1992), *Probleme der Valenz- und Kasustheorie.* (Konzepte der Sprach- und Literaturwissenschaft 51.) Tübingen: Max Niemeyer Verlag.
- Hornby, Albert S. (1985), *Oxford Advanced Learner's Dictionary of Current English.* Oxford: Oxford University Press.
- Hubert, Lawrence/Arabie, Phipps (1985), Comparing Partitions. In: *Journal of Classification* 2, 193–218.
- Joanis, Eric (2002), Automatic Verb Classification Using a General Feature Space. MSc thesis, Department of Computer Science, University of Toronto.
- Joanis, Eric/Stevenson, Suzanne (2003), A General Feature Space for Automatic Verb Classification. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics.* Budapest, Hungary, 163–170.
- Kaufman, Leonard/Rousseeuw, Peter J. (1990), *Finding Groups in Data – an Introduction to Cluster Analysis.* New York: John Wiley & Sons, Inc.
- Kinyon, Alexandra/Prolo, Carlos A. (2002), Identifying Verb Arguments and their Syntactic Function in the Penn Treebank. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation.* Las Palmas de Gran Canaria, Spain, 1982–1987.
- Klavans, Judith L./Kan, Min-Yen (1998), The Role of Verbs in Document Analysis. In: *Proceedings of the 17th International Conference on Computational Linguistics.* Montreal, Canada, 680–686.
- Korhonen, Anna (2002), *Subcategorization Acquisition.* PhD thesis, Computer Laboratory, University of Cambridge. Published as Technical Report UCAM-CL-TR-530.
- Korhonen, Anna/Gorrell, Genevieve/McCarthy, Diana (2000), Statistical Filtering and Subcategorization Frame Acquisition. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* Hong Kong, China, 199–205.
- Korhonen, Anna/Krymolowski, Yuval/Marx, Zvika (2003), Clustering Polysemic Subcategorization Frame Distributions Semantically. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.* Sapporo, Japan, 64–71.
- Lapata, Mirella/Brew, Chris (2004), Verb Class Disambiguation Using Informative Priors. In: *Computational Linguistics* 30(1), 45–73.
- Levin, Beth (1993), *English Verb Classes and Alternations.* Chicago: The University of Chicago Press.
- Li, Hang/Abe, Naoki (1998), Generalizing Case Frames Using a Thesaurus and the MDL Principle. In: *Computational Linguistics* 24(2), 217–244.

- de Lima, Erika (2002), The Automatic Acquisition of Lexical Information from Portuguese Text Corpora with a Probabilistic Context-free Grammar. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Manning, Christopher D. (1993), Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, 235–242.
- Marcus, Mitchell P./Marcinkiewicz, Mary Ann/Santorini, Beatrice (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19(2), 313–330.
- McCarthy, Diana (2001), Lexical Acquisition at the Syntax-semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. PhD thesis, Department of Informatics, University of Sussex.
- Merlo, Paola/Stevenson, Suzanne (2001), Automatic Verb Classification Based on Statistical Distributions of Argument Structure. In: *Computational Linguistics* 27(3), 373–408.
- Merlo, Paola/Stevenson, Suzanne/Tsang, Vivian/Allaria, Gianluca (2002), A Multilingual Paradigm for Automatic Verb Classification. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 207–214.
- Meyers, Adam/Macleod, Catherine/Grishman, Ralph (1994), Standardization of the Complement Adjunct Distinction. In: *Proceedings of the 7th EURALEX International Congress*. Göteborg, Sweden. Available at: <http://nlp.cs.nyu.edu/comlex/adj-compl.ps>.
- O'Donovan, Ruth/Burke, Michael/Cahill, Aoife/van Genabith, Josef/Way, Andy (2005), Large-scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. In: *Computational Linguistics* 31(3), 329–365.
- Ohara, Kyoko Hirose/Fujii, Seiko/Ohori, Toshio/Suzuki, Ryoko/Saito, Hiroaki/Ishizaki, Shun (2004), The Japanese FrameNet Project: An Introduction. In: *Proceedings of the LREC Workshop on 'Building Lexical Resources from Semantically Annotated Corpora'*. Lisbon, Portugal, 249–254.
- Palmer, Martha/Gildea, Daniel/Kingsbury, Paul (2005), The Proposition Bank: An Annotated Corpus of Semantic Roles. In: *Computational Linguistics* 31(1), 71–106.
- Pinker, Steven (1989), *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge: MIT Press.
- Resnik, Philip (1997), Selectional Preference and Sense Disambiguation. In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*. Washington, DC, 52–57.
- Roland, Douglas/Jurafsky, Daniel (2002), Verb Sense and Verb Subcategorization Probabilities. In: Stevenson, Suzanne/Merlo, Paola (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam: John Benjamins, 325–346.
- Rooth, Mats/Riezler, Stefan/Prescher, Detlef/Carroll, Glenn/Beil, Franz (1999), Inducing a Semantically Annotated Lexicon via EM-based Clustering. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD, 104–111.
- Sarkar, Anoop/Zeman, Daniel (2000), Automatic Extraction of Subcategorization Frames for Czech. In: *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, 691–697.
- Schulte im Walde, Sabine (2000), Clustering Verbs Semantically According to their Alternation Behaviour. In: *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, 747–753.
- Schulte im Walde, Sabine (2002a), A Subcategorisation Lexicon for German Verbs Induced from a Lexicalised PCFG. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain, 1351–1357.
- Schulte im Walde, Sabine (2002b), Evaluating Verb Subcategorisation Frames Learned by a German Statistical Grammar against Manual Definitions in the Duden Dictionary. In: *Proceedings of the 10th EURALEX International Congress*. Copenhagen, Denmark, 187–197.
- Schulte im Walde, Sabine (2003), Experiments on the Automatic Induction of German Semantic Verb Classes. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

- Schulte im Walde, Sabine (2006), Experiments on the Automatic Induction of German Semantic Verb Classes. In: *Computational Linguistics* 32(2), 159–194.
- Schulte im Walde, Sabine/Brew, Chris (2002), Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 223–230.
- Siegel, Eric V. (1998). Linguistic Indicators for Language Understanding: Using Machine Learning Methods to Combine Corpus-based Indicators for Aspectual Classification of Clauses. PhD thesis, Department of Computer Science, Columbia University.
- Spranger, Kristina/Heid, Ulrich (2003). A Dutch Chunker as a Basis for the Extraction of Linguistic Knowledge. In: Gaustad, Tanja (ed.), *Computational Linguistics in the Netherlands 2002. Selected Papers from the 13th CLIN Meeting*. Groningen, The Netherlands, 93–109.
- Stevenson, Suzanne/Joanis, Eric (2003), Semi-supervised Verb Class Discovery Using Noisy Features. In: *Proceedings of the Conference on Computational Natural Language Learning*. Edmonton, Alberta, 71–78.
- Subirats, Carlos/Sato, Hiroaki (2004), Spanish FrameNet and FrameSQL. In: *Proceedings of the LREC Workshop on ‘Building Lexical Resources from Semantically Annotated Corpora’*. Lisbon, Portugal, 13–16.
- Tsang, Vivian/Stevenson, Suzanne (2004), Using Selectional Profile Distance to Detect Verb Alternations. In: *Proceedings of the NAACL Workshop on Computational Lexical Semantics*. Boston, MA, 30–37.
- Tsang, Vivian/Stevenson, Suzanne/Merlo, Paola (2002), Cross-linguistic Transfer in Automatic Verb Classification. In: *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan, 1023–1029.
- Ushioda, Akira/Evans, David A./Gibson, Ted/Waibel, Alex (1993), The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora. In: *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text*. Columbus, OH, 95–106.
- Vázquez, Gloria/Fernández, Ana/Castellón, Irene/Martí, María Antonia (2000), *Clasificación Verbal: Alternancias de Diátesis*. (Quaderns de Sintagma 3.) Lleida: Universitat de Lleida.
- Wauschkuhn, Oliver (1999), Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. Doctoral thesis, Institut für Informatik, Universität Stuttgart.

Sabine Schulte im Walde, Stuttgart (Germany)

45. Corpus linguistics and word meaning

1. Word meaning
2. Corpus linguistic studies of word meaning
3. Summary and integration: Lexical priming
4. Acknowledgements
5. Literature

1. Word meaning

Most studies of word meaning (or word sense) begin by questioning the common-sense notion of the ‘word’ and with good reason. However, in this article I am going to begin by treating as unproblematic the term ‘word’, since that will enable us better to explore

- Schulte im Walde, Sabine (2006), Experiments on the Automatic Induction of German Semantic Verb Classes. In: *Computational Linguistics* 32(2), 159–194.
- Schulte im Walde, Sabine/Brew, Chris (2002), Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, 223–230.
- Siegel, Eric V. (1998). Linguistic Indicators for Language Understanding: Using Machine Learning Methods to Combine Corpus-based Indicators for Aspectual Classification of Clauses. PhD thesis, Department of Computer Science, Columbia University.
- Spranger, Kristina/Heid, Ulrich (2003). A Dutch Chunker as a Basis for the Extraction of Linguistic Knowledge. In: Gaustad, Tanja (ed.), *Computational Linguistics in the Netherlands 2002. Selected Papers from the 13th CLIN Meeting*. Groningen, The Netherlands, 93–109.
- Stevenson, Suzanne/Joanis, Eric (2003), Semi-supervised Verb Class Discovery Using Noisy Features. In: *Proceedings of the Conference on Computational Natural Language Learning*. Edmonton, Alberta, 71–78.
- Subirats, Carlos/Sato, Hiroaki (2004), Spanish FrameNet and FrameSQL. In: *Proceedings of the LREC Workshop on ‘Building Lexical Resources from Semantically Annotated Corpora’*. Lisbon, Portugal, 13–16.
- Tsang, Vivian/Stevenson, Suzanne (2004), Using Selectional Profile Distance to Detect Verb Alternations. In: *Proceedings of the NAACL Workshop on Computational Lexical Semantics*. Boston, MA, 30–37.
- Tsang, Vivian/Stevenson, Suzanne/Merlo, Paola (2002), Cross-linguistic Transfer in Automatic Verb Classification. In: *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan, 1023–1029.
- Ushioda, Akira/Evans, David A./Gibson, Ted/Waibel, Alex (1993), The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora. In: *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text*. Columbus, OH, 95–106.
- Vázquez, Gloria/Fernández, Ana/Castellón, Irene/Martí, María Antonia (2000), *Clasificación Verbal: Alternancias de Diátesis*. (Quaderns de Sintagma 3.) Lleida: Universitat de Lleida.
- Wauschkuhn, Oliver (1999), Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. Doctoral thesis, Institut für Informatik, Universität Stuttgart.

Sabine Schulte im Walde, Stuttgart (Germany)

45. Corpus linguistics and word meaning

1. Word meaning
2. Corpus linguistic studies of word meaning
3. Summary and integration: Lexical priming
4. Acknowledgements
5. Literature

1. Word meaning

Most studies of word meaning (or word sense) begin by questioning the common-sense notion of the ‘word’ and with good reason. However, in this article I am going to begin by treating as unproblematic the term ‘word’, since that will enable us better to explore

the relationship between traditional categories used to determine word meaning and the findings of corpus linguistics that are pertinent to this area of study. By the end, however, the term will be seen to have a similar status to the atom in chemistry and micro-physics, neither small enough to account for the nature of matter nor large enough to discuss important chemical properties, and yet continuing to have some usefulness.

Before considering the impact of corpus linguistics on the study of word meaning, it will be necessary to give a brief account of some of the key categories adopted in traditional structural linguistics to account for the meanings that are assigned to words. To begin with, traditional accounts of the word largely agree that most words are **Polysemous**, i. e. that they have multiple related meanings (e. g. Carter 1998). There is, though, little agreement as to how to determine how many senses a word may have, and a glance at any group of modern dictionaries will often turn up significant differences in the number of senses recognised for any particular word.

Polysemy has been handled in a variety of ways. Papers by Copestake/Briscoe (1995) and by Pustejovsky/Bouillon (1995) in a special issue of *Journal of Semantics* argue that the logical composition and semantics of lexical items produce polysemy. Boguraev (1979) and Hirst (1987) seek to account for the disambiguation of polysemous items in terms of the priming effects of the contexts in which they occur, an approach that influenced the second of the corpus-driven approaches we shall be considering later.

Asher (1993), on the other hand, tackles polysemy and ambiguity from a discourse perspective, and an awareness of the communicational needs served by lexis might be argued to underpin a more general approach to its description known as event structure. Vendler (1967) classified events into states, processes, achievements and accomplishments and used this classification to provide a more rigorous account of verbs. This inspired a range of works all seeking a description on the same principles. Dowty (1979), for example, works with four primitives BE, BECOME, CAUSE and DO, while Parsons (1990) makes use of event variables and Pustejovsky (1992) argues for more complex event structures. All of these approaches, however they differ, have as their implicit metaphor the one we began with – that there are sub-atomic particles below the level of the verb/event, a metaphor indeed that is enshrined in the subtitle of Parsons (1990).

Other linguists, e. g. Bierwisch (1982), have sought to provide generalisable conceptual schemata to account for some aspects of polysemy. Thus, for example, Bierwisch notes that the word *letter* may refer to the original piece of paper on which an epistle has been inscribed, multiple copies of such a piece of paper (as in *the letter was sent to all Heads of Department*), the content of the epistle and the whole epistolary genre. Such a range of senses, or conceptual shifts as he terms them, have their parallels in words such as *book*, *symphony* and *picture*, and can be handled in terms of a set of generalisable principles applicable to the conceptual schemata that map onto the separate words. Such an approach not only accounts for some polysemy (though very far from all) but for some co-hyponymy. Its abstractness and doubtful general applicability, however, prevent its further discussion here.

Bierwisch's position is part of a more general position which posits two levels of semantics – one is conceptual and exists independently of individual languages; the other is semantic and language-specific. As Prandi (2004, 155) rather neatly puts it, “just as a sculptor can carve his figures out of shaped as well as out of rough stone, so linguistic structures can carve lexical values out of a shaped as well as out of a shapeless purport.” At the conclusion of this article, we shall have arrived at a position not incompatible with this, though the method of argumentation will be quite different.

A key principle in the description of word meaning is that articulated by Lyons (1977), namely that a word (Lyons actually talks of lexical items, a distinction to which we shall return later) derives its meaning from the relationships it forms with other words in its semantic field. The first and most obvious relationship that may be used to locate the meaning of a word is that of **synonymy**. The statement that *beneath* means the same as *below* is a claim about their synonymy. Of course, as is generally recognised, true synonyms are rare and possibly non-existent. For this reason, Bawcom (2003) favours the term ‘similonym’. Cruse (2000) notes that if true synonymy existed it would not be of interest to linguists. Synonymy, he says, occurs when the similarity in the meaning of two words is more salient than the differences, and what is interesting, he argues, is the question of identifying under what conditions we attend to the similarities between the words rather than the differences.

Another category of word relationship that has been used to pinpoint a word’s meaning is that of **antonymy**. Antonymy is recognised when two words are selected from the same paradigm but with radically contrastive meaning. Thus the statement that *below* is one end of a vertically organized spectrum and that *above* is at the other end is a claim that they are antonyms. Lyons separates antonyms from opposites (Lyons 1977) and both he and Cruse distinguish a number of sub-classes of antonymy, based on criteria such as the gradability of the items or the markedness of one of the members of the pair (Lyons 1963, 1977; Cruse 1976, 1986, 2000). In this article, I shall however lump the different sub-classes together, since my argument is not affected by such distinctions.

Although markedness may be marked morphologically (e. g. by the addition of *un* to *happy*, *fortunate*, *prepared* and so on), it is often implicit, which means that the presence or absence of markedness is part of our stored knowledge about an antonymic pair. Jones (2002) argues that antonymy has a psychological reality (as markedness suggests) and that because of this a number of uses are available to antonymic pairs that would otherwise be inconceivable.

A third category of word relationship that plays an important part in determining word meaning is that of **hyponymy**. This is the relationship central to many definitions, particularly those of nouns. When Humpty Dumpty in Lewis Carroll’s *Alice’s Adventures Through the Looking Glass* defines the borogove as “a thin shabby-looking bird with its feathers sticking out all round – something like a live mop”, he is making use of the superordinate-hyponym relation between *bird* and *borogove*, and in so doing is asserting that a relationship exists between *borogove* and its **co-hyponyms** of *bird* such as *robin* and *owl*. The idea that a word’s meaning can be arrived at by determining what it is not, by examining all the other paradigmatic choices in the system, can be traced back to Saussure (1972 [1916]). However, the semantic field within which the paradigmatic system operates is itself bounded. We compare *red* with *white*, not with *argumentative* (Trier 1973 [1931], cited in Prandi 2004). Co-hyponyms may sometimes be identified by reference to shared appearance in a particular syntactic pattern. Considering the Location-Subject pattern (as in *The garden was swarming with bees*), Dowty (2001) notes that it is associated with a small number of semantic sets such as small local movements (e. g. *crawl*, *drip*, *bubble*), and animal-type sounds, often repetitive (e. g. *buzz*, *cackle*, *chatter*). These sets might each be seen as co-hyponymous.

More generally, and by way of summing up, word meaning can be identified by recognising the word’s relationships with other words in the same semantic field (or the concepts that these words articulate) (Cruse 2000). Some of these words will be syn-

onyms, antonyms, superordinates, hyponyms, co-hyponyms and so on but some will form relationships not covered by the categories so far considered. Thus a *rabbit* is a type of *animal*, it lives in a *burrow* and is used as a *pet* (or as *dog food*).

2. Corpus linguistic studies of word meaning

The categories so far considered are part of the battery of linguistic strategies that have developed over the years to describe word meaning. I have not attempted to be comprehensive but the terminology I have briefly characterised is representative and widely accepted within structural semantics. It should not be inferred from the way I have organised my article that the linguists already cited are uninterested in corpora nor that they have not tested their ideas against the evidence that corpora provide. Both Carter (1998) and Jones (2002) in fact do so, and Jones in particular advances our understanding of the way antonyms are used together very considerably. But in general, the terminology we have inherited for the description of word meaning was developed prior to the availability of electronic corpora and it makes sense to test the insights that such a terminology brings against corpus data. To evaluate its usefulness in accounting for authentic instances of polysemous use, I want now to consider a set of data, derived from a corpus of the *Independent* (a British ‘quality’ newspaper).

2.1. Case study: *dry*

The concordance selection that follows represents 21 uses of the word *dry*:

1. He also had a **dry** wit which was occasionally able to surprise.
2. The Loire is represented by ten wines, among which I particularly liked the smoky, austere **dry** Touraine Sauvignon Blanc from Dutertre Pere et Fils, and an unusual honeyed, more classically gooseberryish Cheverny Sauvignon Blanc from Philippe Tessier.
3. Here it was a mystery trip, and as they waited for it to unfold, they huddled into parkas around the bookies who were keeping their blackboards **dry** under sensible brown umbrellas.
4. Now there, you might think, is one hell of a used myth, squeezed **dry** nearly ten years ago.
5. And because technology has left this early nineteenth-century landscape high and **dry** there was no belching smoke, no infernal clanking to mar the illusion.
6. Inquiries were being made at **dry** cleaners throughout Liverpool.
7. Britain's tourist industry will suffer more in the mid-90s when the supply of school-leavers starts to **dry** up.
8. A single sycamore leaf falls with **dry** scrapes and clicks; the noise of a squirrel shell-ing acorns is quite different.
9. In any case, everywhere looks pretty much the same when filled with clouds of **dry** ice and illuminated with strobes.

10. Each of the farmers in the valley will be given a ‘shopping list’ of incentives covering the restoration of hay meadows, heather moorland and hedgerows to the rebuilding of **dry** stone walls, rights of way and the development of woodlands.
11. The result was a cold, **dry**, emotionless dissertation which in itself looked harmless enough.
12. The duty would extend to all defects, including **dry** rot and adverse planning decisions, ‘of which the vendor knows or ought to know.’
13. The novice faces daily drudgery, toiling in the kitchens, cleaning, helping the seniors wash and **dry** themselves.
14. Racing: High hopes for **dry** run to give Oliver the Edge.
15. ‘I have conceded absolutely nothing,’ she said at the press conference, happy to be back on **dry** land and in charge again.
16. The cow that gave milk to everybody is now **dry**, at least for the next few years, and that will change everything,’ said one Arab analyst.
17. Instead, a little muesli, a slice of **dry** toast and a black coffee to set me up for the weights.
18. Not a **dry** eye in the house, we wager.
19. We whisk past the **dry** dock and the transport cafe with peeling window frames.
20. ‘We have a **dry** law here,’ he said, though his friend a young man in a red headband, took a different tack.
21. George Bush is already home and **dry**, according to NBC, in a swathe of mainly rural states from Idaho in the northern Rockies to Florida in the South and needs only another ten electoral votes to win.

Arguably, every instance of use of the word *dry* in this concordance represents a different meaning of the word and if we ask ourselves how we know the meanings are different, we are likely to find ourselves using the set of categories already mentioned. We noted earlier Lyons’ claim (1977) that word meaning can be isolated with respect to the set of sense relations the word forms with other items in the same semantic field. It follows that we should be able to provide reasons for each polysemous use of the word *dry* from its antonymous, synonymous, hyponymous and other relations with other words (or, rather, with particular uses of other words). Thus, both to justify a separate sense for use 2 and to isolate the meaning it actually has, we are likely to refer to the system of two members – *dry* and *sweet* (antonyms) – that operates with the semantic set of wines. Notice that in such a system, while *dry* has a radically different meaning from that we would most readily associate with the adjective, *sweet* is closer to its prototypical sense (though a dry wine with sugar stirred into it would not be sweet in this sense.) A different antonymous set – *dry* and *buttered* – would in the same way probably be referenced to account for the sense used in use 17, and yet another antonymous set – *dry* and *moist* – for use 18 (use 14 is apparently in the same semantic domain, with *soft* a possible antonym). There is an expression *milch cow* (and for some reason it is always *milch*, not *milk*), and though it is rarely used, the latent antonymy it offers could be thought to help pinpoint the meaning of *dry* in use 16. Although *dry* in use 15 has no obvious antonym in this sense, the whole phrase *dry land* has a quasi-antonymous relationship with water. In terms of the whole family of senses of *dry*, we might want to argue, however, that the central antonymous relationship is that between *wet* and *dry*, which motivates the choice in use 3.

To identify many of the separate senses, we would use not antonymy but synonymy. In the case of use 1, it is more likely that we might want to call upon the category of approximate synonymy – *dry* and *sardonic*, perhaps – and similarly in the case of use 11, we might note the near-synonymous relation of *dry* to *boring* or *dull*. While it is true, as Cruse says, that we are interested in synonyms as “words whose semantic similarities are more salient than their differences”, it remains the case that if the meaning of a word is defined in terms of its relationship to a synonym, it is still the difference that will be in focus. So *result* and *consequence* are synonyms but *consequence* has a strong tendency to occur with negative adjectives (*a grim consequence*) while *result* has a tendency to occur with positive adjectives (*a great result*) (Hoey 2005). In the case of use 11, *dry* seems restricted in its *boring* or *dull* meaning to verbal outputs. So a lecture, sermon or article can be *dry*, but a concert, film or journey can only be *boring* or *dull*.

It is when we turn to the ways senses are constructed out of hyponymous relationships that we first encounter limitations to the claim that the sense of a word is describable in terms of its relationships to other words. *Dry* in *dry stone walls* (use 10) does not participate in any relation that seems to occur naturally (*cemented stone walls* is an antonymous possibility but not one I am aware of ever having encountered). As a phrase, however, it functions as a hyponym of *stone walls* or simply *walls*. The same point may be made of *dry rot* (use 12) and *dry law* (use 20). In other cases, even as a phrase the hyponymous relation is not constitutive of the meaning. *Dry cleaners* are arguably not a hyponym of *cleaners*. Indeed the expressions are close to synonymous in one of the meanings of *cleaners* as the following example (drawn from the Macmillans corpus) illustrates:

22. You can't, you took it to the cleaners on Saturday

Likewise *dry ice* is definitely not a hyponym of *ice*, related though the two expressions are.

Despite the problems we have encountered with the application of the superordinate-hyponym relationship, the categories of antonymy, synonymy and, to some extent, hyponymy have allowed us to isolate particular senses of *dry* as exemplified in uses 1, 2, 3, 6 (in part), 10, 11, 12, 16, 17, 18, 19 and 20. This leaves us with nine senses for which the traditional accounts of word relationships are inadequate. At this juncture we must turn to the insights of corpus linguists. In most cases reference to collocation and idiom is required – *squeezed dry*, *high and dry*, *dry cleaners*, *dry up*, *wash and dry*, *home and dry*, in particular, would be hard to describe without such reference. Collocation can be defined as the association between two words in a language that is made consciously or subconsciously by users of the language; it can be identified in corpora as the recurrent occurrence of two words together at a level of frequency not accounted for by the separate statistical frequency of the words in the language taken as a whole (for a discussion of definitions of collocation, see Partington 1998; Hoey 2005; Evert 2005; and article 58).

On idiom, Sinclair (1991) has this to say:

“The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.”
(Sinclair 1991, 110)

The combinations *squeezed dry*, *high and dry*, *dry cleaners*, *dry up*, *wash and dry*, and *home and dry* all would appear to be accountable for as semi-preconstructed choices of the kind Sinclair describes (or ‘chunks’ in Biber et al. 1999). But if we were to leave it at that, we would be understating the inadequacy of antonymy, synonymy, hyponymy and the other terms for any account of word meaning. To illustrate problems with antonymy, it is worth looking again at use 2 in our selective concordance of *dry*. In the first place, if we are to talk of antonymy in such a case, the basic antonymy is not between *dry* and *sweet* but between *dry wine* and *sweet wine*, both of which are strong collocations. The reason for asserting that this is the case is that the *dry/sweet* antonymic pair is established by, and with an important caveat which we shall come to shortly, recognised through the collocation with *wine*. At first sight, indeed, there appears to be a case for claiming that the antonymy is between two semi-preconstructed phrases. However, searches on Google produce data that suggest a more complex picture:

Tab. 45.1: Google counts for *wine* expressions (searches undertaken in April 2006).

PHRASE	GOOGLE HITS	PHRASE	GOOGLE HITS
red wine	16,900,000	dry red wine	338,000
white wine	10,600,000	dry white wine	1,140,000
sweet wine	587,000	red dry wine	952
dry wine	266,000	white dry wine	935
sweet red wine	66,200	red sweet wine	8,500
sweet white wine	99,100	white sweet wine	225

The first thing to note about these data is that the antonymic pair *dry/sweet* appears to interlock in complex ways with the antonymic pair *red/white*, both governed by the collocation with *wine*. (As hinted above, we shall have occasion to modify the wording of this claim below.) Where the *red/white* opposition is not evoked, it is worth noting that there are twice as many hits for *sweet wine* as for *dry wine*.

The next thing to note about the statistics given above is that references to *red wine* outnumber references to *white wine* at a ratio of 17:11 and yet when it comes to specifications of the dryness/sweetness of the wines, *white wine* outnumbers *red wine* at a ratio of roughly 3:1. It may also be noted that *dry white wine* and *sweet white wine* are clear examples of semi-preconstructed phrases, with low frequencies for the alternatives *white dry wine* and *white sweet wine*. The same can also broadly be said for the phrases *dry red wine* and *sweet red wine*, which outnumber considerably the frequencies for *red sweet wine* and *red dry wine*. However, the number of hits for *red sweet wine* is considerably larger than those for *red dry wine*, *white sweet wine* or *white dry wine*.

The best explanations of these data raise questions about the relationship of antonymy that we assumed underpinned our identification of the flavour sense of *dry*. Dryness is a default flavour of both red and white wine. This would explain the greater frequency of *sweet wine* over *dry wine*, since there is normally no need to specify dryness. Furthermore, for many speakers, there seems to be no available *dry/sweet* antonymy for talking about *red wine*. Informal checks amongst wine-drinking friends revealed that they did not know of the existence of *sweet red wines*, though, had they visited Austria or Hungary, their response might well have been very different. So when such speakers refer to a *dry red wine*, they cannot be accessing the *dry/sweet* antonymy to make sense

of *dry*. The relative instability of the sequence of *sweet* and *red*, as evidenced by the relatively high number of references to *red sweet wine*, further suggests that, at least for those who used this sequence of words, sweetness was likewise understood without reference to the notion of dryness.

On the other hand, for my wine drinking informants, the option of *sweet white wine* is a recognised reality, and the figures quoted above seems to support the view that there is an established antonymic relation between the semi-preconstructed pairs *dry white wine* and *sweet white wine*. Even here, though, there are grounds for doubt. Use 2 refers not to *dry white wine* but to *dry Sauvignon Blanc*, a hyponym of wine. Although a white wine, *Sauvignon Blanc* is not recognised as existing in a sweet version for the great majority of wine drinkers. Although a Google search threw up 26 references to *sweet Sauvignon Blanc* (and a further six to *slightly sweet Sauvignon Blanc*), this possibility is not, I would argue, accessed in references to *dry Sauvignon Blanc*. Instead, as with *dry red wine*, the dryness is understood non-antonymically. Interestingly, German has the adjective + noun phrase (*roter Wein* ‘red wine’) but also the adjective + noun compound (*Rotwein*). These can be described as sweet or dry, i.e. *trockener* ‘dry’ *Rotwein* etc. One does not have #*Trockenwein*, though there is *Süßwein* which is however not the same as *süßer Wein* ‘sweet wine’, *Süßwein* being confined to dessert wines like port or madeira. The antonym to *trocken* would be *lieblich*.

In short, it makes more sense to say that the meanings of *dry* and *sweet* are arrived at independently of any antonymy that might be perceived between them, and that this antonymy is a post-hoc explanation of a cluster of largely non-contrastive expressions with shared lexical context. Teubert/Čermáková (2004) demonstrate succinctly the difficulties attached to providing antonymous descriptions, and their work seems compatible with the argument just presented. A similar argument based on similar data could be presented for hyponymy and co-hyponymy. In each case it is not the relationship that gives us access to word meaning; it is the occurrence of word meanings in particular contexts that allows us to generalise in terms of hyponymy and co-hyponymy. Bierwisch (1982, 4), commenting on Putnam, notes that “someone who cannot tell an elm from a beech tree is not able to determine the extension of these terms, even though he might know that their extensions are different.” In other words (which, however, Bierwisch might not acknowledge as a paraphrase), the co-relation of *elm* and *beech tree* with *tree* allows a person to posit that they are co-hyponyms but this will not help such a person identify the trees (and one tree could be identifiable for this person without reference to the other.) The case concerning synonymy is rather different in that here the perception of shared meaning is not a product of the co-text, it being normal for synonyms to differ in their collocations (Partington 1998; Hoey 2005).

There are a number of ways out of these apparent circularities that have been proposed by corpus linguists. They are not mutually exclusive, though it does not follow that each presupposes the others. Unsurprisingly, all focus on the syntagmatic dimension, given that this is the dimension that corpus linguistics can describe most directly, and all directly address the issue of polysemy as well as raising questions about the status of the word as a unit.

2.2. Collocational environments

A key early corpus-based approach to word meaning is associated with John Sinclair, who noted in the preparation of the *Collins COBUILD Dictionary* (Sinclair et al. 1987)

that the existence of distinct collocational environments for a word was a reliable guide for the identification of polysemous uses (see Sinclair 1987, for an account of the lexicographic practices used in the preparation of that dictionary). The principles underlying this are fleshed out in Sinclair (1991). The same principles are employed by Partington (1998) in his study of synonymy. After a detailed account of the apparently synonymous words *sheer*, *pure*, *absolute* and *complete* in their hyperbolic uses, he notes that “each of these synonyms was ... found in its own phraseological patterns, and this fact supports the argument ... that every lexical item in the language has its own individual and unique pattern of behaviour.” In a similar vein, but focussing on grammatical rather than collocational patterns, Mindt (1991) notes that semantic distinctions correlate with characteristic syntactic and morphological contexts.

Sinclair’s focus on the collocational profiles of polysemous uses of a word led him both to recognise contextual constraints other than collocation and to question whether the lexicon is word-centred. In Sinclair (1996, revised 2004), he argues against the use of the concept of ‘word’, wishing instead to talk of ‘lexical items’. In this, of course, he is far from alone; we have already noted in passing Lyons’ preference for the latter term. What distinguishes Sinclair’s position is that his concept is much more extended and fuzzy-bounded and is described in terms of a range of corpus-driven criteria. He makes use in particular of collocation (already discussed), colligation (the co-occurrence of a word or phrase with a grammatical function or a grammatical category, see also article 43), semantic preference (the co-occurrence of a word or phrase with members of a semantic set) and semantic prosody (crudely, the pragmatic interpretation selected by the writer/speaker for the whole utterance, reflected in the choices s/he makes, and accessed by the reader/listener). In a brilliant interpretation of a set of 151 concordance lines for the preconstructed phrase *naked eye*, drawn from The Bank of English, he shows how layers of selections interlock to create something much less fixed than a traditional idiom but nevertheless much more intimately co-selected than can be accounted for in terms of word choices in grammatical slots. His observations can be simplified as follows:

- In 95% of the lines, the phrase *naked eye* is preceded by the word *the* (collocation)
- 90% of the lines have a preposition immediately preceding the position held by *the*, with *to* and *with* dominating (colligation)
- Preceding both the preposition and *the*, there is almost always some reference to (in)visibility; exact figures are less easily come by here because some cases are more central than others (semantic preference)
- Where the visibility choice is realised as an adjective, the preposition referred to above is *to* and in most cases where the visibility choice is a verb, the preposition is *with* (colligation)
- To the left of all these choices, there is characteristically some indication of difficulty of seeing (semantic prosody).

The following line illustrates all these features:

23. agents too small to see with the naked eye

It will be seen that this account of lexical item meaning treats the words as atoms which combine to make molecules that function as unities in combination with other such

molecules. It is a dominantly syntagmatic account and draws on none of the categories we have been considering in the first part of this paper.

We can apply Sinclair's approach to word meaning to use 5. The preconstructed phrase *high and dry* occurs in 23 concordance lines generated from a corpus of the *Guardian* newspaper of 1990 and 1991 (renumbered from 1 for convenience):

1. am and with whom he had added 91, **high and dry** on 99. Weston had been
2. transferred across the Atlantic and **high and dry** between two cultures. Tim
3. pose problems (best remember it by '**high and dry**') It is part of the Far East p
4. priest, and leaves the rest of the cast **high and dry** as they work gingerly and
5. et: Gower dismissal leaves Hampshire **high and dry** – Nottinghamshire v Hamp
6. p breaks up, the so-called stars are left **high and dry** just playing out the old sta
7. retrain for new jobs. We might be left **high and dry**, pensioned off with inadeq
8. said the parents felt they had been left **high and dry** by social services. No soci
9. il route from east to west has been left **high and dry** as a result. Effjohn, the Fi
10. back in the sea. Far from being left **high and dry**, they are simply trying to r
11. Walter Hammond was fumingly left **high and dry** with the tail. The tyro tot w
12. tales about how we're going to get left **high and dry** if we devote too much time
13. miscast and Laura San Giacomo is left **high and dry** as the unlikely provider of
14. German Grand Prix while Senna is left **high and dry** /hl> NIGEL M
15. pened. The Pride of Baltimore was left **high and dry**, all too visibly in breach of
16. Mr Rushdie is concerned, he was left **high and dry**, clinging only to some tent
17. r wild-card game, the Dolphins looked **high and dry** trailing the Kansas City Ch
18. verhampton: Looters leave newcomers **high and dry** – Michael White looks at th
19. to leave Smith, not best pleased, **high and dry** on 99. This was something
20. leaving both customer and salesman **high and dry** with a newly clinched orde
21. ptiness. The boat, marking time in the **high and dry**, serves a function as
22. cer: Well take the Cup to leave United **high and dry** again – Scottish Cup, at
23. they? Once you've exchanged, you're **high and dry**, tight? Crack out the cham

In 17 of the lines (74%), the phrase is preceded by one of the morphological variants of LEAVE (collocation). Of these lines, eleven are in the passive voice (colligation). In all but three of the 17 lines with LEAVE, what is left is either one or more human beings or a team (semantic preference). Intriguingly, two of the exceptions are vehicles (lines 9 and 15, a train – the Venice Simplon-Orient Express – and a boat – The Pride of Baltimore), a category of things which are sometimes referred to as if they were human (e.g. with the pronoun *she*). The other exception is an animal (line 10) and I shall have occasion to refer to this example immediately below. Finally, we may note that in all but one case – that of the animal – the lines with LEAVE are all associated with insoluble problems (semantic prosody). Line 8 from the above concordance, expanded to include its immediate context, illustrates all the features just mentioned:

8. Ms Amphlett criticised the absence of home and social assessments, and said the parents felt they had been left high and dry by social services. No social worker had visited any of the parents after the removal of the children, and the parents themselves felt abused.

The exception in line 10 is the only literal use of the phrase and describes an animal for whom being *high and dry* is no problem. Interestingly, though, the animal is described as suffering from human assumptions that being high and dry is an insuperable problem, which strongly supports the analysis just given.

10. THE recent gales have swept a large number of baby seals onto Welsh beaches and well-meaning but dim-witted folk have been going to enormous trouble to chuck them back in the sea. Far from being left high and dry, they are simply trying to recover from their prolonged battering, says Arthur Bryant who runs a hospital for marine animals at New Quay, Dyfed.

Sinclair's approach to the meaning and structure of lexical items is part of a larger theory in which the language user constructs his or her utterances by moving between the idiom principle and the open choice principle. Another corpus-based approach to word meaning that seeks to resolve the difficulties encountered above also has its roots in a larger theory, that of lexical priming (Hoey 2004, 2005). This approach, like Sinclair's, makes use of collocation, colligation, and semantic association (exactly equivalent to Sinclair's semantic preference). In addition it makes use of pragmatic association (which overlaps with but is not identical to semantic prosody) and a range of textual features including textual semantic association (association with particular textual relations or patterns of organisation).

3. Summary and integration: Lexical priming

The underlying assumption of lexical priming is that each time we encounter a word, we note and keep a record of the contexts in which it has occurred. As patterns of recurrence of context emerge, we become primed to associate the word with these recurrent contexts, the contexts here including all the features listed above. Importantly for a description of word meaning, primings nest. So if we are primed to associate *wine* with *white*, the combination *white wine* is then available to be in turn primed for *dry*. So combinations larger than the word have the same properties as the word on its own. Likewise, we may be primed for the contexts of pieces of language smaller than the word. So *some* is primed to occur for many speakers with problematic states in *burden-some*, *wearisome*, *troublesome*, *tiresome*, *gruesome* and *cumbersome*; such a priming is in principle indistinguishable from the primings for words and combinations of words. This is the same idea underlying many accounts of morphological productivity. See for example article 41 or Hay (2003), who shows that frequency of simplex and complex expressions determines how likely it is that they will be recognized as complex (but see also Baroni 2001, who could not find a similar effect for Italian derivations with the prefix *ri-*).

So, returning to the atom analogy we have used before, we are finding that elements below the level of the atom are fully describable.

Also important for a view of word meaning is that primings are psychological phenomena and each person's experience of any particular word is inevitably unique. It follows then that the primings for a word may vary from person to person, being based

on different encounters with the word in different contexts. Therefore for most wine drinkers, *red wine* is primed to occur only with *dry* but for a few well-informed or adventurous drinkers, the phrase has become more weakly primed to occur with *sweet*. Many of this minority are primed for the pattern FLAVOUR + COLOUR + wine, and so accept, and produce, the combination *sweet red wine*. Others are only primed for *dry red wine* and therefore can say *red sweet wine* without overriding their primings. Since primings vary from person to person, so do word meanings, albeit in largely trivial and unthreatening ways. One implication, though, of the claim that each person's primings are potentially unique is that a corpus can only indicate the patterns for which a word may be primed for an individual; it cannot demonstrate that such priming actually has occurred for any user of the language.

The implications of a theory of lexical priming for an account of word meaning can be gleaned from consideration of the way in which it might account for use 7 from the concordance selection for *dry* above. Whereas Sinclair was drawing conclusions in general about the combination *naked eye*, here the focus is on explaining how *dry up* gets its meaning in the line

7. Britain's tourist industry will suffer more in the mid-90s when the supply of school-leavers starts to **dry up**.

An examination was made of a concordance of 41 lines of the words *dry up*, again generated from the Guardian corpus of 1990 and 1991 and here renumbered from 1:

1. Then he shrugs. 'They are going to all **dry up** sometime, I suppose.' Gooch
2. of smoke and bury their houses and **dry up** their water courses. Their landsc
3. lem. If purchasers for corporate assets **dry up** then prices will slide into a down
4. aid. 'If inflows into the secondary banks **dry up**, then they're in trouble. This has
5. ystem of bribes, rewards, and brutality. **Dry up** the cash which keeps all that in
6. r all, be getting his teeth into the case? '**Dry up**, Lewis,' rasped Morse, handing
7. into adults. Ponds and ditches **dry up**; sudden storms can wash them
8. some aid is siphoned off, now watch it **dry up** ... Watch the march of the
9. of Mark Hughes and Brian McClair **dry up** as 1989, and United, faded
10. of organic rhubarb. Tears mysteriously **dry up** and heart uncurdles. Brain, goin
11. severe effect on the rivers. They often **dry up** in summer, river channels are no
12. on Gibraltar: Employment opportunities **dry up** as British troops prepare to
13. in. The company has seen orders **dry up** as its major customers – the local
14. while chalk bedded rivers. The rivers **dry up**, the appearance deteriorates an
15. international funding would slowly **dry up** to the point where refugees
16. th lack of work as raw material supplies **dry up** because of unpaid bills. Internal
17. out of the sums they realise. If these **dry up**, who will liquidate the
18. later, when the adrenalin began to **dry up**, did Connors the realist speak. 'I
19. funds for the contras began to **dry up**, North happily drew up a list of
20. 4–5 million. As finance began to **dry up**, the world began to take off.
21. war and the arms and money began to **dry up** / The rebels' desperate efforts to
22. line from the schools began to **dry up** as multi-option, non-team game
23. 's Tsing-Tao. When supplies began to **dry up**, a Russian friend furnished a bot
24. sure which, the radio work began to **dry up** as the Cold War years crept on,

25. rewarding allies and punishing foes, to **dry up** Iraq's sources of cash and credit
 26. ears of Thatcherism. They will have to **dry up** liquidity in some other way,' he s
 27. been left, like a cut-off meander, to **dry up** because the powerful current of
 28. invoking the 'great goddess Nature' to '**dry up** in her the organs of increase', au
 29. g a series of absorbent 'fuzz panels' to **dry up** as much of the sound as require
 30. this year and council building set to **dry up** before long. The Government's h
 31. many small wells and springs to **dry up** in summer. Israel counters with t
 32. ms. Consequently abstraction tends to **dry up** the river sources. Halcrow
 33. friends initially, but that work tends to **dry up** very quickly,' he said. Jack
 34. of new money which is unlikely to **dry up** when the markets throw a
 35. regional phone companies will **dry up** before the infrastructure is
 36. he is serious, or that the roles ever will **dry up**, but 'in a certain way, it is a relief
 37. inquiries from potential investors will **dry up** and the only outside agents for
 38. r. She is no longer afraid her muse will **dry up**, she is confident and feels better
 39. finance for experimental projects will **dry up** as the economic crisis deepens.
 40. rapid climate changes. Many rivers will **dry up** or shrink. The level of the Great
 41. rolling stock to Network SouthEast will **dry up** by April 1993. That year will be

These data suggest the following observations on *dry*.

Firstly, and self-evidently, *dry* is primed to collocate with *up*.

In this combination, it is primed for semantic association with LIQUIDS in Subject or Object function, occurring in the data with this semantic association eight times (lines 2, 7, 10, 11, 14, 31, 32 and 40). This semantic association contributes to distinguishing one sense of *dry up* from its other sense(s).

In the remaining 33 lines, the combination *dry up* occurs with *began to* seven times (lines 18–24) (21% of the relevant data set), indicating a fairly strong collocational priming. This draws attention to the process meaning of the combination for most users.

Of the 33 lines of *dry up* without semantic association with LIQUIDS, 28 (85%) show *dry up* being used ergatively (e. g. lines 33–39), a strong colligational priming. This points to a lack of explicit agency in the meaning of *dry up* for most users.

Of the 28 instances of *dry up* + ERGATIVE (excluding LIQUIDS), 23 (82%) have abstract Subjects (e. g. lines 36–39) (colligation).

Of the 23 instances of ABSTRACT SUBJECT + *dry up* + ERGATIVE (– LIQUIDS), 20 have as Subject something that would normally be regarded as good (e. g. investment (line 35 – see the expanded version below), orders [in the company sense] (line 13), muse (line 38)); none have something bad. So most users are primed to see *dry up* in the combination of primings ABSTRACT SUBJECT + *dry up* + ERGATIVE (– LIQUIDS) as something negative that happens (pragmatic association).

16 of the 20 instances (80%) of GOOD ABSTRACT SUBJECT + *dry up* + ERGATIVE (– LIQUIDS) describe a problem (in Sinclair's terms, we have a semantic prosody), and 10 of the 20 (50%) are followed by a Response in a Problem-Solution pattern (textual semantic association). An example is the following expansion of line 35:

35. ... the nascent cable television shudders at this and claims that investment in cable largely from the North American regional phone companies will dry up before the infrastructure is complete if BT gets the go ahead. The cable industry wants a 15-year breathing space before BT can compete ...

So most speakers are primed to see *dry up* in this nested combination of primings as a marker of a Problem to be solved.

For simplicity's sake, I have acted as if these individual primings only operate with particular combinations. In fact, there is considerable cross-fertilisation, so that, for example, ergativity is associated with the LIQUIDS use (e.g. line 40) and Problem is associated with concrete Subjects (e.g. line 3). A fuller analysis would deal with each of these factors separately.

If we look again at the line from the original data set for *dry*, we see that it largely conforms to the description arrived at on the basis of the concordance:

7. Britain's tourist industry will suffer more in the mid-90s when the supply of school-leavers starts to **dry up**.

In this line *dry* (of course) collocates with *up*, the combination *dry up* is used ergatively, we have an abstract Subject which would normally be regarded as good, and the context makes it clear that the drying up will be a problem, though the lack of a larger context for this line makes it impossible to verify whether it is part of a Problem-Solution pattern. The only deviation from the description provided above is that we have *starts to* in place of *began to*, which suggests that we are typically primed for a semantic association with COMMENCE, with *began* as the most normal manifestation of this association.

What the approach just described shares with Sinclair's, apart from a heavy overlap of categories of analysis, and despite the differences of theoretical starting point, is an approach to word meaning that does not make use directly of traditional terminologies. It is not being argued that the categories of synonymy, antonymy and hyponymy cannot fruitfully continue to be used, but they will need to be used within a syntagmatic framework and recognised as outcomes of analyses that take account of the idiosyncrasies of individual words and of individual (groups of) users. What is now urgently needed is corpus-driven work on ways of integrating word-concept relationships (as in *dry* as an experience of a flavour offered by Sauvignon Blanc) to the semantic and pragmatic associations and collocational and colligational idiosyncrasies of words such as *dry* identified in the latter part of this article. Out of attempts at such integration might grow a new and more adequate terminology for the description of word meaning.

4. Acknowledgements

The data in the first set of concordance lines were generated for me by Dr Steven Jones of the University of Manchester from a corpus of *Independent* newspapers created by Professor Antoinette Renouf.

5. Literature

- Asher, N. (1993), *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers.
- Baroni, M. (2001), The Representation of Prefixed Forms in the Italian Lexicon: Evidence from the Distribution of Intervocalic [s] and [z] in Northern Italian. In: Booij, G./van Marle, J. (eds.), *Yearbook of Morphology 1999*. Dordrecht: Springer, 121–152.

- Bawcom, L. (2003), Bawcom's Blossom. Unpublished presentation given at the Tuscan Word Centre, May 2003.
- Biber, D./Johansson, S./Leech, G./Conrad S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Bierwisch, M. (1982), Formal and Lexical Semantics. In: *Linguistische Berichte* 80, 3–17.
- Boguraev, B. (1979), Automatic Resolution of Linguistic Ambiguities. PhD Thesis, Computer Laboratory, University of Cambridge.
- Carter, R. (1998), *Vocabulary: Applied Linguistic Perspectives*. 2nd edition. London: Routledge.
- Copestake, A./Briscoe, T. (1995), Semi-productive Polysemy and Sense Extension. In: *Journal of Semantics* 12, 15–67.
- Cruse, D. A. (1976), Three Classes of Antonyms in English. In: *Lingua* 38, 281–292.
- Cruse, D. A. (1986), *Lexical Semantics*. Cambridge: Cambridge University Press.
- Cruse, D. A. (2000), *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.
- Dowty, D. (1979), *Word Meaning in Montague Grammar*. Dordrecht: Reidel.
- Dowty, D. (2001), *The Semantic Asymmetry of 'Argument Alternations' (and Why it Matters)*. Available at <http://www.ling.ohio-state.edu/~dowty> (accessed February 2007).
- Evert, S. (2005), *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Doctoral dissertation, University of Stuttgart. Available at <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/> (accessed May 2007).
- Hay, J. (2003), *Causes and Consequences of Word Structure*. New York and London: Routledge.
- Hirst, G. (1987), *Semantic Interpretation and the Resolution of Ambiguity*. (Studies in Natural Language Processing.) Cambridge: Cambridge University Press.
- Hoey, M. (2004), The Textual Priming of Lexis. In: Bernardini, S./Aston, G./Stewart, D. (eds.), *Corpora and Language Learners*. Amsterdam: John Benjamins, 21–41.
- Hoey, M. (2005), *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Jones, S. (2002), *Antonymy: A Corpus-based Perspective*. London: Routledge.
- Lyons, J. (1963), *Structural Semantics*. Oxford: Blackwell.
- Lyons, J. (1977), *Semantics*. Cambridge: Cambridge University Press.
- Mindt, D. (1991), Syntactic Evidence for Semantic Distinctions in English. In: Ajmer, K./Altenberg, B. (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 182–196.
- Parsons, T. (1990), *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, Massachusetts: MIT Press.
- Partington, A. (1998), *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Prandi, M. (2004), *The Building Blocks of Meaning: Ideas for a Philosophical Grammar*. Amsterdam/Philadelphia: John Benjamins.
- Pustejovsky, J. (1992), The Syntax of Event Structure. In: Levin, B./Pinker, S. (eds.), *Lexical and Conceptual Semantics*. Cambridge MA and Oxford, UK: Blackwell, 47–81.
- Pustejovsky, J./Bouillon, P. (1995), Aspectual Coercion and Logical Polysemy. In: *Journal of Semantics* 12(2), 133–162.
- Putnam, H. (1975), The Meaning of 'Meaning'. In: Gunderson, K. (ed.), *Language, Mind and Knowledge*. Minneapolis: University of Minnesota Press, 131–193.
- Saussure, F. de (1972 [1916]), *Cours de linguistique générale*. Critical edition by T. de Mauro. Paris: Payot.
- Sinclair, J. (ed.) (1987), *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.
- Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1996), The Search for Units of Meaning. In: *Textus* 9(1), 75–106.
- Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

- Sinclair, J./Hanks, P./Fox, G./Moon, R./Stock, P. (eds.) (1987), *Collins COBUILD English Language Dictionary*. London: Collins.
- Teubert, W./Čermáková, A. (2004), Directions in Corpus Linguistics. In: Halliday, M. A. K./Teubert, W./Yallop, C./Čermáková, A. (eds.), *Lexicology and Corpus Linguistics*. London: Continuum, 113–166.
- Trier, J. (1931), *Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Bd. 1: Von den Anfängen bis zum Beginn des 13. Jahrhunderts*. Heidelberg: Winter. Reprinted in Lee, A. van der/Reichmann, O. (eds.) (1973), *Aussätze und Vorträge zur Wortfeldtheorie*. The Hague: Mouton, 40–65.
- Vendler, Z. (1967), Verbs and Times. In: Vendler, Z. (ed.), *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press, 97–121.

Michael Hoey, Liverpool (UK)

46. Theory-driven corpus research: Using corpora to inform aspect theory

1. Introduction
2. Can corpora contribute to linguistic theory?
3. Using corpora to inform aspect theory
4. Conclusions
5. Literature

1. Introduction

The theory-driven versus data-driven distinction in linguistics is a manifestation of the conflict between rationalism and empiricism in philosophy. The extremist views of these two approaches to linguistics are vividly illustrated by Fillmore's (1992) cartoon figures of the armchair linguist and the corpus linguist. The armchair linguist thinks that what the corpus linguist is doing is uninteresting while the corpus linguist believes that what the armchair linguist is doing is untrue. It is hardly surprising that the divorce of theory and empirical data results in either untrue or uninteresting theories because any theory that cannot account for authentic data is a false theory while data without a theory is just a meaningless pile of data. As such, with exceptions of a few extremists from either camp who argue that “Corpus linguistics doesn't mean anything” (see Andor 2004, 97), or that nothing meaningful can be done without a corpus (see Murison-Bowie 1996, 182), the majority of linguists (e.g. Leech 1992; Meyer 2002) are aware that the two approaches are complementary to each other. In Fillmore's (1992, 35) words, “the two kinds of linguists need each other. Or better, [...] the two kinds of linguists, wherever possible, should exist in the same body”.

- Sinclair, J./Hanks, P./Fox, G./Moon, R./Stock, P. (eds.) (1987), *Collins COBUILD English Language Dictionary*. London: Collins.
- Teubert, W./Čermáková, A. (2004), Directions in Corpus Linguistics. In: Halliday, M. A. K./Teubert, W./Yallop, C./Čermáková, A. (eds.), *Lexicology and Corpus Linguistics*. London: Continuum, 113–166.
- Trier, J. (1931), *Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Bd. 1: Von den Anfängen bis zum Beginn des 13. Jahrhunderts*. Heidelberg: Winter. Reprinted in Lee, A. van der/Reichmann, O. (eds.) (1973), *Aussätze und Vorträge zur Wortfeldtheorie*. The Hague: Mouton, 40–65.
- Vendler, Z. (1967), Verbs and Times. In: Vendler, Z. (ed.), *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press, 97–121.

Michael Hoey, Liverpool (UK)

46. Theory-driven corpus research: Using corpora to inform aspect theory

1. Introduction
2. Can corpora contribute to linguistic theory?
3. Using corpora to inform aspect theory
4. Conclusions
5. Literature

1. Introduction

The theory-driven versus data-driven distinction in linguistics is a manifestation of the conflict between rationalism and empiricism in philosophy. The extremist views of these two approaches to linguistics are vividly illustrated by Fillmore's (1992) cartoon figures of the armchair linguist and the corpus linguist. The armchair linguist thinks that what the corpus linguist is doing is uninteresting while the corpus linguist believes that what the armchair linguist is doing is untrue. It is hardly surprising that the divorce of theory and empirical data results in either untrue or uninteresting theories because any theory that cannot account for authentic data is a false theory while data without a theory is just a meaningless pile of data. As such, with exceptions of a few extremists from either camp who argue that “Corpus linguistics doesn't mean anything” (see Andor 2004, 97), or that nothing meaningful can be done without a corpus (see Murison-Bowie 1996, 182), the majority of linguists (e.g. Leech 1992; Meyer 2002) are aware that the two approaches are complementary to each other. In Fillmore's (1992, 35) words, “the two kinds of linguists need each other. Or better, [...] the two kinds of linguists, wherever possible, should exist in the same body”.

This article discusses the use of corpus data in developing linguistic theory (section 2) and presents an effort to achieve a marriage between theory-driven and corpus-based approaches to linguistics via a series of case studies of aspect (section 3), which has long been studied, but rarely with recourse to corpus data.

2. Can corpora contribute to linguistic theory?

To answer this question, we must first of all find out what linguistics is about. We will then discuss the use of intuitions and corpora as evidence in linguistic theorizing and explore how corpus data can contribute to linguistic theory.

2.1. What linguistics is about

It has been argued that linguistics is “the study of abstract systems of knowledge idealized out of language as actually experienced”, i.e. “idealized internalized I-language” (Widdowson 2000, 6). If linguistics is defined as such, we must admit that any linguistic analysis involving performance data (i.e. “E-language”) has nothing to do with “linguistics” and should claim no place in “linguistics” at all (cf. Leech 2000, 685). The assumption underlying Widdowson’s definition is Chomsky’s (1965; 1986) claim that competence can be separated from performance to be studied alone. But can the two be separated?

The competence vs. performance divide is rooted in the hypothesis that grammar is autonomous within the human mind. Generative grammarians argue that our use of language (performance, E-language) cannot reflect our internal knowledge of language (competence, I-language), because of the constraints in naturally occurring language. Performance errors have been likened to abnormal conditions like tiredness and drunkenness in human communication (e.g. Radford 1997, 2). Only the internal grammar, which is based on native intuitions and not polluted by performance constraints, is said to be part of competence. The corollary of this argument is the sharp distinction between langue and parole (Saussure 1916 [1966]), between performance and competence (Chomsky 1965), and between grammar and usage (Newmeyer 2003). Nevertheless, this dichotomy is arguably over-stated. Evidence from recent research in psycholinguistics, neurolinguistics, and biology shows that the hypothesis of an autonomous grammar, which underlies the sharp distinction between competence and performance, is unsustainable (see Shei 2004 for a review). Rather, grammar is constantly shaped by culture (or environmental factors) and interpersonal interactions. In de Beaugrande’s (1997, 302) words, “performance can crucially determine the development and quality of competence”. On the other hand, performance does not spring from nowhere – it is a natural and actual product of competence. Clearly, as Leech (1992, 108) observes, “the putative gulf between competence and performance has been overemphasized”.

Given the nature of this interdependence, the Chomskyan linguists’ practice of forcing a sharp distinction between competence and performance is simply misleading in that it is in essence merely an “idealization of language for the sake of simplicity” (Abney 1996, 11). In doing so, real language is replaced by idealized language which does not

exist but which purports to sustain an explanation of language (cf. de Beaugrande 1997) while attested language data is denied a place in theory building. In the dialectic view of the relationship between competence and performance, therefore, the assertion is simply unsustainable that performance “cannot constitute the subject-matter of linguistics” (Chomsky 1965, 20), because competence is not directly accessible and our only gateway to it is through performance (cf. Meyer/Nelson 2006). Linguistics is in fact concerned with what language really is – as reflected by our knowledge, as well as use, of language. Just as Kennedy (1998, 270) argues:

“Furthermore, description of the system we use is not the only legitimate goal of the study of language. The linguistic system is both derived from and instantiated by specific instances of use. It is thus perfectly legitimate to describe language both in terms of the system we use and our use of this system and for the description thus to encompass language as possibility as well as probability of use”.

2.2. Intuitions and corpus data in theory building

Intuitions and corpus data are two important types of evidence in linguistic theory. Linguistic intuitions can be used in introspection to invent (grammatical, ungrammatical, or questionable) example sentences for linguistic analysis, or make judgments about the acceptability/grammaticality or meaning of an expression. They are always useful in linguistics as the linguist can invent purer examples instantly for analysis. This is so because intuitions are readily available and invented examples are free from language-external influences existing in naturally occurring language. Intuitions are even in a sense indispensable in linguistic theorizing because categorization, which usually involves intuitive judgments, is unavoidable in theory building (see section 3). Nevertheless, intuitions should be applied with caution (cf. Seuren 1998, 260–262). Firstly, it is possible to be influenced by one’s dialect or sociolect (cf. also Krishnamurthy 2000b, 172). Consequently, what appears grammatically unacceptable to one speaker may be perfectly felicitous to another (cf. also Wasow/Arnold 2005, 1482; see Schütze 1996 for further discussion of grammaticality judgments). Secondly, when one invents an example to support or disprove an argument, one is consciously monitoring one’s language production. Therefore, even if one’s intuitions are correct, the example may not represent typical language use. Thirdly, introspective data is decontextualized because it exists in the analyst’s mind rather than in any real linguistic context. Context is particularly relevant to acceptability and grammaticality judgments. With proper contexts, what might appear ungrammatical or unacceptable out of context can become grammatical and acceptable while “our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances)” (Krishnamurthy 2000a, 32–33). Fourthly, results based on intuitions alone are difficult to verify as introspection is not observable. Fifthly, excessive reliance on intuitions blinds the analyst to the realities of language usage (cf. Meyer/Nelson 2006). For example, we tend to notice the unusual but overlook the commonplace because of the psychological salience of rare words and usages (Sinclair 1997, 33; Krishnamurthy 2000b, 170–171). Finally, there are areas in linguistics where intuitions cannot be used reliably but must rely upon corpus data, e.g. language variation, historical linguistics, register and style, first and second language acquisition (Meyer 2002; Léon 2005, 36).

With that said, we must hasten to add that we do not mean that linguistic intuitions are useless and should be abandoned in theory building as corpus-driven linguists have advocated (see section 2.3.). Absolutely not. We use our intuitions in the first place to decide what to examine and then to interpret what we find (cf. Krishnamurthy 2000b, 172). Without intuitions, it is also impossible to make judgments of the acceptability or meaning of an expression, or to categorize data – even though contextual clues are sometimes available. Our point is simply that intuition as a type of linguistic evidence should be used in conjunction with empirical evidence collected from other sources rather than being relied upon solely as the basis of linguistic theory. However, there has been an overreliance upon introspective evidence in theoretical linguistics (cf. also Gast 2006). One of the symptoms is that, as Adger (2002, no page) notes, “[s]ome of the data of core linguistics is actually generated by the theory, in that experimental and intuition-based data are collected to test theory, and the collection techniques are designed with this aim in mind”. That explains why such data can be “biased towards the point that is to be proved, i. e. linguists may see what they want to see” (Johansson 1995; cited in Krishnamurthy 2000b, 170). It is simply a vicious circle to develop a linguistic hypothesis on the basis of an analyst’s introspective data, which is used in turn to verify the same hypothesis. As such, Labov (1972, 199) argued that “linguists cannot continue to produce theory and data at the same time”.

Another symptom of the overreliance is that intuitions are given such a privileged status in generative linguistics that “evidence other than intuitions is brought in only as supporting evidence” (Wasow/Arnold 2005, 1484) whilst “contradictory evidence from other sources [...] is simply ignored” (Wasow/Arnold 2005, 1486). Given that it “would simply be a waste of time and energy” to “devise operational and experimental procedures” (Chomsky 1969, 81) while usage data cannot seriously “constitute the actual subject matter of linguistics” (Chomsky 1965, 4), the overreliance in the generative tradition upon intuitions is quite unsurprising. Nevertheless, as “All of us, even native speakers or ‘expert speakers’, have only a partial knowledge of that language” (Krishnamurthy 2000b, 172), a linguistic theory will become more reliable and convincing if linguists care to check whether their intuitions are “in accord with what people actually say and write” (Wasow/Arnold 2005, 1486).

In contrast with introspective data which relies solely on one’s own intuitions, a corpus pools together the linguistic intuitions of a range of speakers and/or writers. Corpora comprise samples of spoken/written language which has already occurred naturally in real linguistic context. As people speak and write on the basis of their intuitions in real contexts, corpus data is also intuition-based; but it is more natural than introspective data because it is not created specifically for linguistic analysis. In relation to data collected through introspection of an individual, corpus data typically reflects the intuitions of a much greater number of language users. Corpora can also provide frequencies readily, which cannot be predicated by intuition reliably (cf. McEnery/Wilson 2001, 15). As such, corpus data allows a linguist to avoid any potential bias in his/her own intuitions and distinguish what is statistically central and typical from what is statistically marginal in theory development. In short, a corpus typically provides data that is attested, contextualized and quantitative. It can also find differences that intuitions alone cannot perceive (cf. Francis/Hunston/Manning 1996; Kennedy 1998, 272). In addition, corpora have opened up or foregrounded a number of new areas of linguistic research that would not have been possible on the basis of intuitions alone, most notably register and variation studies (see section 3.3. for an example; see also article 38 in this volume).

While the corpus-based approach has won widespread popularity and has been used in nearly all branches of linguistics (see McEnery/Xiao/Tono 2006 for an overview), corpora have also become the target of a number of criticisms. For example, Chomsky (1957) has argued that a corpus only contains a finite number of sentences while language is “an infinite set of sentences” (though see de Beaugrande (2002, 105) for a counter-argument for this definition of language; and see article 2 in this volume for discussion of early generative linguistics). Since a corpus does not include each and every possible sentence of language, corpus data is by nature “skewed” – “[s]ome sentences won’t occur because they are obvious, others because they are false, still others because they are impolite” (Chomsky 1962, 159, cited in Leech 1991, 8). These criticisms are certainly valid, especially when they were made in the 1950s. Corpora, especially those used in what McEnery/Wilson (2001) call “early corpus linguistics”, were ready targets of such criticisms because of their small sizes and inadequate sampling. Chomsky can indeed be considered as the person who has helped to “shape the approach taken by the corpus today” (McEnery/Wilson 2001, 19), because these criticisms have led to such key concepts as balance, representativeness and sampling in corpus linguistics which, coupled with developments in technology, and especially the development of ever more powerful computers offering ever increasing processing power and massive storage at relatively low cost, have made corpora of today as large and balanced as practically possible to be maximally representative of the language or language variety under consideration (see article 9 in this volume). While it might be true that a 100-million word balanced corpus is still skewed to some extent, it is certainly less skewed than a dataset obtained through introspection on the basis of one analyst’s intuitions. Intuitions can be skewed because “the process of introspection may not be systematic” (McEnery/Wilson 2001, 15) and because intuitions are discriminating – “[m]atters of wit, curiosity and love of the unusual, the absurd, etc., have a further impact on the intuition” (Sinclair 1997, 33). Corpora have been criticized for being skewed simply because they are observable and open to scrutiny whereas intuitions are not.

Chomsky (1965) argued against corpora also because corpus data is likely to contain performance errors which have nothing to do with one’s knowledge of language. This criticism is true, but it is reasonable to assume that a corpus is generally composed of sentences which are grammatical (cf. also McEnery/Wilson 2001, 16). Corpus data at least provides evidence of what speakers believe to be grammatically acceptable in their language. Intuitions are not error-free either, though, not to mention the bias as noted earlier. Labov (1975) has shown that one’s intuitions of grammaticality may not necessarily be a true reflection of one’s internal grammar. Furthermore, as a corpus presents data in context, it allows for research into what types of performance errors occur under what conditions and are typically associated with what contexts. Theories of this type cannot be developed on the basis of decontextualized introspective data but they are of practical importance in linguistics. In our view, therefore, a “performance grammar” (Chomsky 1962, 537–538) that copes with regular and irregular language phenomena (including performance errors) is of greater importance than a “competence grammar” that has little bearing on “everyday production or comprehension of language” (Schütze 1996, xi).

However, while the corpus-based approach has some advantages over the intuition-based approach, it also has some known weaknesses. Firstly, as a corpus cannot possibly include all sentences in a language, sampling is unavoidable and the representativeness

of the corpus becomes an issue. Nowadays, representativeness is still regarded as an “act of faith” (Leech 1991, 127) for lack of a reliable scientific measure of corpus balance, though the confidence about a corpus can be increased when the corpus increases to a respectable size and achieves a wide coverage (cf. article 9 in this volume). Secondly, statistical methods that are more sophisticated and rigorous are required to interpret corpus data. Quantitative analysis is equally important as qualitative analysis in corpus research. Many statistical measures which are commonly used in corpus linguistics assume that linguistic features are evenly distributed in language – in different corpora or in different samples in a corpus – which may not be the case. Hence, we support Gries’s (2006) argument for “more rigorous corpus linguistics” (see article 36 for further discussion of statistical methods in corpus exploration). Thirdly, a corpus does not provide negative evidence (but see Stefanowitsch 2006 for a counter-argument). A corpus, however large and balanced it is, cannot be exhaustive except for highly specialized cases (e.g. the corpus of the Bible mentioned in article 9), because natural language is infinite. As such, corpora cannot tell us what is possible or not possible in language. If a construction does not exist in a corpus, you cannot say that it does not exist in the language (but according to Stefanowitsch 2006, it is possible to tell what is “significantly absent” from what is “accidentally absent” on the basis of a properly annotated corpus); neither can you say that a construction found in a corpus is necessarily grammatically acceptable because a corpus may contain “performance errors”. Nevertheless, everything included in a corpus is what language users have actually produced – for good or ill. The emphasis of corpus research is on “the repetitive and routine nature of language use” (Stubbs 2001a, 152), though hapax legomena can also be of interest in some studies (e.g. article 41). Corpus data is useful in showing what is statistically central and typical in language. If a “performance error” is repeated sufficiently often by a sufficiently large group of native language users, the “error” might as well be approached from the perspective of language variation or language change, while in the case of learner corpora it is precisely repetitive patterns of such performance errors that make data of this kind useful in interlanguage studies. Finally, while the corpus-based approach is good at yielding interesting findings, it cannot explain what we find in corpora. The explanations must be developed using other methodologies and evidence from other sources, including intuitions.

In spite of the philosophical tension between theoretical linguists and corpus linguists, the intuition-based approach and corpus-based approach are not necessarily antagonistic. But rather the two approaches corroborate each other and can be “gainfully viewed as being complementary” (McEnery/Wilson 2001, 19). Given that both introspective data and corpus data have their own weaknesses as noted earlier, it is our view that the theory-driven and data-driven approaches to linguistics should be combined to take advantage of their strengths while circumventing their weaknesses. Broadly speaking, compared with the more traditional intuition-based approach, which rejected or ignored corpus data, the corpus-based approach can achieve improved reliability because it does not go to the extreme of rejecting intuitions while attaching importance to empirical data. The key to using corpus data is to find the balance between the use of corpus data and the use of one’s intuitions. As Leech (1991, 14) observes:

“Neither the corpus linguist of the 1950s, who rejected intuitions, nor the general linguist of the 1960s, who rejected corpus data, was able to achieve the interaction of data coverage and the insight that characterize the many successful corpus analyses of recent years”.

Unsurprisingly, a number of areas in modern linguistics have relied upon a fusion of corpus evidence and intuitions, ranging from the more practical aspects such as sociolinguistic studies (see article 6), language teaching (see article 7) and lexicography (see article 8), to more theory-driven research including syntax (see article 42) and grammar (see article 43).

Theory-driven corpus research has so far been confined largely to the distribution of forms rather than semantic aspects of language (cf. Kennedy 1998, 272). While meanings related to forms (e.g. semantic prosody, semantic preference, and pattern meaning) have also become a focus of recent corpus research (e.g. Louw 1993, 2000; Stubbs 1995; Partington 1998; 2004; Xiao/McEnery 2006a), core semantic notions such as aspect have rarely been approached from a corpus-based perspective. This is probably because the study of aspectual meaning involves much greater use of intuitions than lexical and grammatical studies and thus has been approached traditionally without recourse to corpus data. However, as we will see in the case studies presented in section 3, the theory-driven and data-driven approaches can be fruitfully combined in the development of aspect theory. But before we present the case studies, it is appropriate to discuss how corpora can contribute to linguistics.

2.3. Corpus-based versus corpus-driven linguistics

Whether corpora should be used at all in linguistics is one issue, and how corpora should be used is another. Having established that corpus data can indeed contribute to linguistic theory, this section discusses how corpora are used to achieve this goal. Even among those who advocate the use of corpus data, there are different opinions and different approaches. One further area where differences diverge in corpus linguistics is with regard to the question of corpus-based and corpus-driven approaches. While, as we will see shortly, the distinction between the two is overstated, what underlies the proposed distinction is highly relevant to the discussion of the present article – how pre-corpus theoretical premises and intuitions should be incorporated in corpus research. In a nutshell, corpus-driven linguists aim to build theory “from scratch” – claiming that they are completely free from pre-corpus theoretical premises – and base their theories exclusively on corpus data, assuming that “all the relevant information is contained in the corpus itself, and the linguist’s task is to *extract* that information and make it visible” (Gast 2006, 114), whilst corpus-based linguists tend to approach corpus data “from the perspective of moderate ‘corpus-external’ premises” (*ibid.*) with the aim of testing and improving such theories.

In the corpus-based approach, it is said that corpora are used mainly to “expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study” (Tognini Bonelli 2001, 65). Corpus-based linguists are accused of not being fully and strictly committed to corpus data as a whole as they have been said to discard inconvenient evidence (i.e. data not fitting the pre-corpus theory) by “insulation”, “standardization” and “instantiation”, typically by means of annotating a corpus. In contrast, corpus-driven linguists are said to be strictly committed to “the integrity of the data as a whole” (*ibid.*, 84) and therefore, in this latter approach, it is claimed that “[t]he theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus” (*ibid.*, 85). Upon interrogating

the available evidence, nevertheless, it is found that the proposed sharp distinction between the corpus-based vs. corpus-driven approaches is overstated and that this “radically empiricist” way of doing corpus research” (Gast 2006, 114) is an idealized extreme. There are four basic differences between the corpus-based vs. corpus-driven approaches: types of corpora used, attitudes towards existing theories and intuitions, focuses of research, and paradigmatic claims. Let us discuss each in turn.

Regarding the type of corpus data used, there are three issues – representativeness, corpus size and annotation. Let us consider these one by one. According to corpus-driven linguists, there is no need to make any serious effort to achieve corpus balance and representativeness because the corpus is said to balance itself when it grows to be big enough, as the corpus achieves so-called cumulative representativeness. This initial assumption of self-balancing via cumulative representativeness, nonetheless, is arguably unwarranted. For example, one such cumulatively representative corpus is a corpus of Zimbabwean English that Louw (1991) used in his contrastive study of collocations of in British English and Zimbabwean English. This study shows that the collocates of *wash* and *washing*, etc. in British English are *machine*, *powder* and *spin* whereas in Zimbabwean English the more likely collocates are *women*, *river*, *earth* and *stone*. The different collocational behavior was attributed to the fact that the Zimbabwean corpus has a prominent element of literary texts such as Charles Mungoshi’s novel *Waiting for the Rain*, “where women washing in the river are a recurrent theme across the novel” (Tognini Bonelli 2001, 88). One could therefore reasonably argue that this so-called cumulatively balanced corpus was skewed. Especially where whole texts are included, a practice corpus-driven linguists advocate, it is nearly unavoidable that a small number of texts may seriously affect, either by theme or in style, the balance of a corpus. Findings on the basis of such cumulatively representative corpora may not be generalizable beyond the corpora themselves as their balance is easily affected by the availability of electronic text of different types.

The corpus-driven approach also argues for very large corpora. While it is true that the corpora used by corpus-driven linguists are very large (for example, the Bank of English has grown to 524 million words), size is not all-important, as Leech (1991, 8–29) and McCarthy/Carter (2001) note (but see Krishnamurthy 2000b for a counter-argument). Another problem for the corpus-driven approach relates to frequency. While it has been claimed that in the corpus-driven approach corpus evidence is exploited fully, in reality frequency may be used as a filter to allow the analyst to exclude some data from their analysis. For example, a researcher may set the minimum frequency of occurrence for a pattern which it must reach before it merits attention, e.g. it must occur at least twice – in separate documents (Tognini Bonelli 2001, 89). Even with such a filter, a corpus-driven grammar would consist of thousands of patterns which would bewilder the learner. It is presumably to avoid such bewilderment that the patterns reported in the *Grammar Patterns* series (Francis/Hunston/Manning 1996, 1998), which are considered as the first results of the corpus-driven approach, are not even that exhaustive. Indeed, faced with the great number of concordances, corpus-driven linguists are often found to analyze only the n^{th} occurrence from a total of X instances. This is in reality currently the most practical way of exploring a very large corpus which is unannotated. Yet if a large corpus is reduced to a small dataset in this way, there is little advantage in using very large corpora and it can hardly be claimed that corpus data is exploited fully and the integrity of the data is respected. It appears, then, that the corpus-driven approach is not so different from the corpus-based approach – while the latter allegedly

insulates theory from data or standardizes data to fit theory, the former filters the data via apparently scientific random sampling, though there is no guarantee that the corpus is not explored selectively to avoid inconvenient evidence.

The corpus-driven linguists have strong objections to corpus annotation. This is closely associated with the second difference between the two approaches – different attitudes towards existing theories and intuitions. It is claimed that the corpus-driven linguists come to a corpus with no preconceived theory, with the aim of postulating linguistic categories entirely on the basis of corpus data, though corpus-driven linguists do concede that pre-corpus theories are insights cumulated over centuries which should not be discarded readily and that intuitions are essential in analyzing data. This claim is a little surprising, as traditional categories such as nouns, verbs, prepositions, subjects, objects, clauses, and passives are not uncommon in so-called corpus-driven studies. When these terms occur they are used without a definition and are accepted as given. Also, linguistic intuitions typically come as a result of accumulated education in preconceived theory. So applying intuitions when classifying concordances may simply be an implicit annotation process, which unconsciously makes use of preconceived theory. As implicit annotation is not open to scrutiny, it is to all intents and purposes unrecoverable and thus more unreliable than explicit annotation. Like the purely rationalist approach to linguistics that rejects corpus data, corpus-driven linguists take a radically empiricist approach that aims to reject everything outside a corpus in spite of the known weaknesses of corpus data (see section 2.2.). In contrast, corpus-based linguists do not have such a hostile attitude towards existing theory. The corpus-based approach typically has existing theory as a starting point and corrects and revises such theory in the light of corpus evidence. As part of this process, corpus annotation is common. Annotating a corpus, most notably part-of-speech tagging, inevitably involves developing a tagset on the basis of an existing theory, which is then tested and revised constantly to mirror the attested language use. In spite of the usefulness of corpus annotation as a result, which greatly facilitates corpus exploration, annotation as a process is also important. As Aarts (2002, 122) observes, as part of the annotation process the task of the linguist becomes “to examine where the annotation fits the data and where it does not, and to make changes in the description and annotation scheme where it does not”. The claimed independence of preconception on the part of corpus-driven linguists is clearly an overstatement. A truly corpus-driven approach, if defined in this way, would require something such as someone who has never received any education related to language use and therefore is free from preconceived theory, for as Sampson (2001, 135) observes, schooling plays an important role in forming one’s intuitions. Given that preconceived theory is difficult to totally reject and dismiss, and intuitions are indeed called upon in corpus-driven linguistics, we cannot see any real difference between the corpus-driven demand to re-examine pre-corpus theories in the new framework and corpus-based linguists’ practice of testing and revising such theories. Furthermore, if the so-called proven corpus-driven categories in corpus-driven linguistics, which are supposed to be already fully consistent with and directly reflect corpus evidence, also need refinement in the light of different corpus data, the original corpus data is arguably not representative enough. The endless refinement will result in inconsistent language descriptions which will place an unwelcome burden on the learner. In this sense, the corpus-driven approach is no better than the corpus-based approach.

The third important difference between the corpus-driven and corpus-based approaches is their different research foci. As the corpus-driven approach makes no distinc-

tion between lexis, syntax, pragmatics, semantics and discourse (because all of these are pre-corpus concepts and they combine to create meaning), the holistic approach provides, unsurprisingly, only one level of language description, namely, that of functionally complete units of meaning or language patterning. In studying patterning, corpus-driven linguists concede that while collocation can be easily identified in KWIC concordances of unannotated data, colligation is less obvious unless a corpus is grammatically tagged. Yet a tagged corpus is the last thing the corpus-driven linguists should turn to, as grammatical tagging is based on preconceived theory, and consequently results in a loss of information, in their view. To overcome this problem, Firth's definition of colligation is often applied in a loose sense – in spite of the claim that corpus-driven linguistics is deeply rooted in Firth's work – because studying colligation in Firth's original sense necessitates a tagged or even a parsed corpus. According to Firth (1968, 181), colligation refers to the relations between words at the grammatical level, i. e. the relations of “word and sentence classes or of similar categories” instead of “between words as such”. But nowadays the term *colligation* has been used to refer not only to significant co-occurrence of a word with grammatical classes or categories (e. g. Hoey 1997, 2000; Stubbs 2001b, 112) but also to significant co-occurrence of a word with grammatical words (e. g. Krishnamurthy 2000a). The patterning with grammatical words, of course, can be observed and computed even using a raw corpus.

A final contrast one can note between corpus-based and corpus-driven approaches is that the corpus-based approach is not as ambitious as the corpus-driven approach. The corpus-driven approach claims to be a paradigm within which a whole language can be described. No such claim is entailed in the corpus-based approach. Yet the corpus-based approach, as a methodology which makes use of corpus data and intuitions, has been applied in nearly all branches of linguistics.

The discussion in this section shows that the sharp distinction forced between the corpus-based vs. corpus-driven approaches to linguistics is in reality fuzzy. If the purely intuition-based rationalist approach discussed in the previous section and the radically empiricist corpus-driven approach characterized in Tognini Bonelli (2001) are viewed as the two ends of a scalar rationalism-empiricism continuum (i. e. “the armchair linguist” and “the corpus linguist” in Fillmore's (1992) humourous account), it can be said that just as the former goes to one extreme by rejecting corpus data, the latter goes to the other extreme by rejecting everything outside a corpus. The corpus-based approach lies in between the extremes, seeking to strike a balance between the use of corpora and the use of intuitions. As both intuitions and corpus data have known weaknesses which can be avoided when the two types of complementary and corroborating evidence are taken into account (see section 2.2.), the corpus-based approach is arguably more reliable than the extremist methods that reject either corpora or intuitions.

Having established that corpus data can contribute to linguistic theory and that in so doing the corpus-based approach is more appropriate, we will present a case study of aspect in the remainder of the article, which seeks to achieve a marriage between theory-driven and corpus-based approaches to linguistics.

3. Using corpora to inform aspect theory

Aspect is a linguistic phenomenon that is related to the temporal properties of linguistically described situations in the world (situation aspect) and how these situations are presented (viewpoint aspect). Situation aspect is composed of inherent features whereas

viewpoint aspect is composed of non-inherent features of aspect. The two components of aspect interplay to determine the aspectual meaning of an utterance (cf. Smith 1997). As the temporal notion denoted by aspect is essential to human languages, aspect has long been the subject of intensive studies by both semanticists and grammarians. However, while corpora have been used extensively in a wide range of areas in linguistics, research on aspect has rarely used corpus data. Yet corpora have a role to play both in developing and testing such theories.

With a few exceptions, most studies on aspect published to date have been based on a handful of examples invented through introspection (e.g. Verkuyl 1993; Smith 1997; Klein/Li/Hendriks 2000), some of which are, if not intuitively unacceptable, unnatural and atypical of attested language use (see Xiao/McEnery 2004a; 2004b for further discussion). Furthermore, those proposals have not been tested with corpus data, which can serve as a test-bed for the linguistic theory proposed as well as for the intuitions on which the theory is based (cf. section 2.2.). This section reports on the corpus-based research in aspect which we have recently undertaken.

3.1. Situation aspect: A corpus-based two-level model

Situation aspect is concerned with the aspectual classification of verbs and situations according to their temporal features such as dynamicity, durativity and telicity. While the earliest literature on aspectual classification dates as far back as Aristotle, modern approaches to aspect are normally considered to start with Vendler (1967), who classified verbs into four classes: state, activity, accomplishment, and achievement, as shown in Table 46.1.

Tab. 46.1: Vendler's four verb classes

Classes	[±dynamic]	[±durative]	[±telic]	Examples
State	—	+	—	know, love, believe, possess
Activity	+	+	—	run, walk, swim, push a cart
Accomplishment	+	+	+	run a mile, walk to school, paint a picture
Achievement	+	—	+	recognize, spot, find, lose, reach, win

As can be seen in the table, Vendler's analysis basically works at the lexical level (cf. Verkuyl 1993, 33), though it also involves predicates rather than simply verbs alone. As such, Vendler has to put *run* and *walk* under the category of activity and put *run a mile* and *walk to school* under the category of accomplishment. Ever since Vendler (1967), a number of theories have been proposed to account for the compositional nature of situation aspect. The most important models include Verkuyl (1993) and Smith (1997). However, all of the models are deeply flawed. For example, Verkuyl incorrectly argues that durativity is linguistically irrelevant and that external arguments also contribute to situation aspect, while Smith's model only works at the sentential level.

Xiao/McEnery (2004a) developed a two-level model of situation aspect on the basis of an investigation of the English and Chinese languages using a fusion of native speaker intuitions and evidence from corpora. The new model of situation aspect consists of three components: a lexicon, a layered clause structure and a set of rules mapping verb classes onto situation types.

In this new theory, situation aspect is modeled as verb classes at the lexical level and as situation types at the sentential level. Using a newly established five-way classification system, situation aspect is classified into six verb classes at the lexical level, namely, individual-level state (ILS), stage-level state (SLS), activity, semelfactive, accomplishment, and achievement (see Table 46.2). The verb classes at the lexical level constitute the lexicon of our model.

Tab. 46.2: Feature matrix system of verb classes

Classes	[±dynamic]	[±durative]	[±bounded]	[±telic]	[±result]
Activity	+	+	-	-	-
Semelfactive	+	-	±	-	-
Accomplishment	+	+	+	+	-
Achievement	+	-	+	+	+
ILS	-	+	-	-	-
SLS	±	+	-	-	-

At the sentential level, situation aspect is classified into the same six basic situation types and five derived situation types. Situation types are the composite result of the rule-based interaction between verb classes and complements, arguments, peripheral adjuncts and viewpoint aspect at three layers of the clause structure: the nucleus, core, and clause levels.

Our two-level approach to modeling situation aspect was motivated by the deficiencies of Vendler (1967) and Smith (1997). The Vendlerian approach works well at the lexical level, but not at the sentential level. Conversely the approach of Smith (1997) works well at the sentential level but not at the lexical level. The two-level approach to situation aspect has sought to bridge this gap, operating at both lexical and sentential levels. While the two-level approach to modeling situation aspect has given a better account of the compositional nature of situation aspect by proposing a set of rules mapping verb classes at the lexical level onto situation types at the sentential level, it has also provided a more refined classification of situation aspect, most notably by distinguishing between two types of states. As the new model of aspect is based on and verified by corpus data from English and Chinese, it is more explanatory of attested language usages in the two distinctly unrelated languages. Indeed, as Xiao/McEnery (2002) observe, situation aspect is language independent. Our two-level model of situation aspect represents an extension of Smith's (1997) two-component aspect theory.

Our model of situation aspect has drawn evidence from both corpus data and intuitions. In theory-driven corpus research of this kind, both types of evidence are indispensable because they interact with each other in theory building. On the one hand, classifying verbs at the lexical level and situations at the clause level on the basis of semantic features is a task that is virtually impossible without recourse to one's intuitions, because the feature values such as telicity cannot be determined reliably using

linguistic co-occurrence tests (see Xiao/McEnery 2006b for a discussion of using compleative and durative temporal adverbials such as *in/for an hour* as telicity tests). The indispensable role of intuitions in theorization as demonstrated in this example shows that such a purely empiricist approach as taken by corpus-driven linguists, that rejects everything outside a corpus, is merely wishful thinking (cf. section 2.2.). On the other hand, corpora are not only a valuable resource that helps to test old hypotheses and formulate new ones, they are also a touchstone for our intuitions.

3.2. Aspect in Mandarin Chinese: A corpus-based model

Mandarin Chinese as an aspect language has played an important role in the development of aspect theory. Nearly all of the major works on aspect theory make reference to Chinese (e.g. Comrie 1976; Smith 1997). Nevertheless, while a few aspect markers have been studied intensively in Chinese linguistics for decades, little attention has been paid to date to the question of systematically describing the linguistic devices that the language employs to express aspectual meanings. Still less attention has been paid to the inherent temporality of situations denoted by utterances in Chinese. But aspect markers that signal different perspectives from which a situation can be presented are only one component of aspect. Worse still, there has been no generally agreed account even of the three most frequently studied aspect markers *-le*, *-zhe*, and *-guo*. For example:

Should the verb-final *-le* be distinguished from the sentence-final *le* and the modal particle *le*? Does the verb-final *-le* indicate the termination or completion of a situation? Does the verb-final *-le* interact with stative and atelic situations? Should the sentence-final *le* be covered in a study of aspect in Chinese? If so, what is its aspectual meaning? Is it necessary to distinguish between the experiential *-guo* and the RVC (resultative verb complement) *guo*? Does the imperfective *-zhe* indicate resultativeness or durativeness? How should the interchange between *-le* and *guo* be accounted for? Under what conditions is the perfective *-le* interchangeable with the imperfective *-zhe*? While intuitions are essential for answers to questions like these, proposals based on introspective evidence alone cannot account for the complexities existing in authentic language data.

Xiao/McEnery (2004b) presents a corpus-based study of aspect in Chinese, which demonstrates how corpora and linguistic theory can interact. All of the above issues have been addressed in this book. More importantly, the book explores aspect at both the semantic and grammatical levels. The two levels correspond to the two components of aspect, namely, situation aspect and viewpoint aspect. Situation aspect operates at the semantic level while viewpoint aspect operates at the grammatical level, but the two also interact with each other, thus explaining why some aspect markers are incompatible with some situation types while other aspect markers show a preference for other situation types. The corpus-based model of aspect in Chinese represents a systematic and structured exploration of linguistic devices which Chinese employs to express aspectual meanings. In addition to situation aspect, which is inherent in linguistic expressions of situations in human languages, this book has identified, on the basis of corpus data, four perfective viewpoints (the actual aspect marked by *-le*, the experiential aspect marked by *-guo*, the delimitative aspect marked by verb reduplication, and the compleative aspect marked by RVCs) and four imperfective ones (the durative aspect marked by *-zhe*, the

progressive aspect marked by *zai*, the inceptive aspect marked by *-qilai*, and the continuative aspect marked by *-xiagu*) in Chinese, and has discussed the characteristic features of each of them in exhaustive detail on the basis of their behavior in attested language use. Barring the three most studied aspect markers mentioned earlier, the aspectual values of the others have been overlooked in most research to date. For example, while RVCs were found in this book to be the most productive perfective markers indicating the completeness of a situation, their aspectual meanings have rarely been discussed elsewhere. While cursory discussions of some of these markers can be found scattered around a number of studies, they have mostly been misunderstood. Kang (1999, 223–243), for example, correctly treats *-qilai* as an aspect marker, yet she conflates its resultative and completive meanings together with its inceptive meaning. The current work has overcome these problems and defined the meaning and form of each aspect marker, thus giving a consistent account of viewpoint aspect in Mandarin Chinese. In addition, the book has corrected many intuition-based misconceptions and associated misleading conclusions readily found in the literature (see below).

Of particular importance is that the model of Chinese aspect focuses on the interaction between situation aspect and viewpoint aspect, which can only be explored reliably using a corpus-based approach because of the gradient nature of this interaction. For example, quite contrary to many intuition-based proposals in the literature (e.g. Pan 1993; Smith 1997; Li 1999), the perfective *-le* in Chinese is not sensitive to the features [\pm dynamic] or [\pm telic]. Rather, as our corpus data shows, *-le* can interact with all situation types in Chinese but it strongly prefers spatially or temporally bounded situations, which account for about 90 percent of the situations co-occurring with *-le*. With unbounded states, *-le* demonstrates the feature of ingressive dynamicity and coerces these situations into derived activities at the clause level. As a perfective marker, *-le* only indicates the actualization and focuses on the entirety of a situation but does not provide any final endpoint as the English simple aspect does. The interaction between the progressive *zai* and achievements is also not as simple as has traditionally been assumed. It has been asserted (e.g. Smith 1997; Yang 1995; Kang 1999) that achievements never occur with the progressive marker *zai*. Nevertheless, our corpus data shows that achievements of different types demonstrate different degrees of compatibility with the progressive aspect in Chinese. While simplex achievements and complex achievements with completive RVCs are strictly incompatible with the progressive, those with result-state and directional RVCs show some tolerance to the progressive aspect. These examples not only demonstrate that corpus data can correct biased intuitions, they also show that quantitative data readily available from corpora “can decide issues that less empirically minded researchers could debate endlessly without ever reaching a conclusion” (Stefanowitsch 2006, 98).

Xiao/McEnery (2004b) has sought to achieve a marriage between theory-driven and corpus-based approaches to linguistics through a study of aspect in Chinese. The use of corpus data as an input to the semantic analysis of aspect represents something new. Previous approaches to the semantics of aspect have rarely used corpus data. Yet the marriage of the corpus-based approach and traditional intuition-based semantic analysis has enabled this book to produce a more realistic account of situation aspect and viewpoint aspect in Chinese in a way that has not been attempted previously. As such, we believe that the book is a powerful demonstration of the way in which corpus data may lead to more accurate linguistic descriptions and hence theories.

3.3. Aspect marking: Contrastive and translation studies

The corpus-based aspect model established in Xiao/McEnery (2004b), which was first developed in Xiao (2002), also demonstrates its value as a unified language-independent framework not only for analyzing a single language, but also for contrasting two or more languages and explaining shifts of situation and viewpoint aspect which often occur in translations, hence helping us to explore the process of translation.

McEnery/Xiao/Mo (2003), for example, used the aspect model to contrast aspect marking in Chinese, British and American English on the basis of three comparable corpora, namely, the Lancaster Corpus of Mandarin Chinese (LCMC) and its matches for British and American English FLOB and Frown (see article 20). The study shows that while Chinese and English are distinctly different, aspect markers in the two languages show a strikingly similar distribution pattern, especially across the two broad categories of narrative and expository texts, as shown in Figure 46.1. In both LCMC and FLOB/Frown, the text categories where the frequency of aspect markers is above the average are the five fiction categories (text categories L, M, N, P, and K) plus humor (R), biography (G), and press reportage (A). The text categories where aspect markers occur least frequently include reports/official documents (H), academic prose (J), skills/trades/hobbies (E), press reviews (C), press editorials (B), religion (D), and popular lore (F). In both Chinese and British/American English, there is a great difference in usage between the first and second groups of texts, which indicates that the two are basically different. Text types like fiction, humor, and biography are narrative whereas reports/official documents, academic prose, and skills/trades/hobbies are expository. Press reportage appears to be a transitory category which is more akin to narrative texts. Statistic tests show that in both Chinese and the two varieties of English, the differences between the distribution of aspect markers in narrative and expository texts are significant. According to Hopper (1979), among many others, the discourse functions of aspect

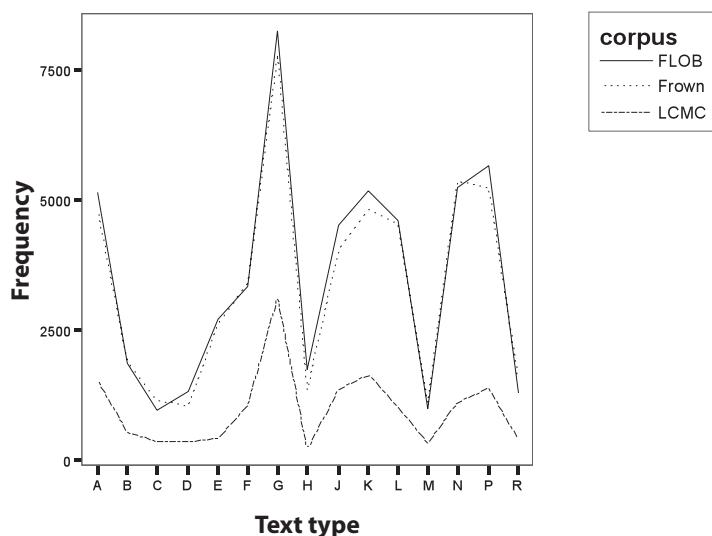


Fig. 46.1: Distribution of aspect markers (frequency)

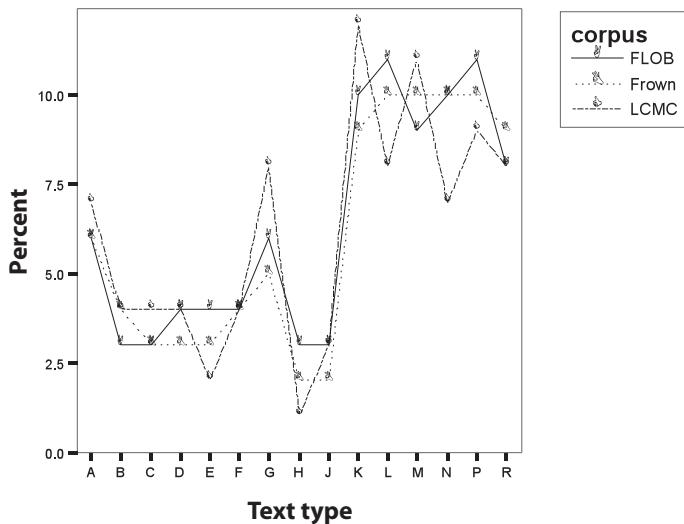


Fig. 46.2: Distribution of aspect markers (percentage)

marking are theoretically linked to foregrounding/backgrounding in narration. Consequently, it is hardly surprising to find that cross-linguistically, aspect markers are significantly more common in the “active, event-oriented discourse” than the “more static, descriptive or expository types of discourse” (Biber 1988, 109). Note that variation studies of this kind are important in linguistic theorizing because variation is inherent in language (cf. Meyer 2002, 3). As Elliot/Legum/Thompson (1969, 52) observed, there are facts about linguistic theory “whose existence will be obscured unless variation is taken into account”. While intuitions are particularly useful in a semantic domain such as aspect, they are of little use in language variation research, which relies heavily upon corpus data (cf. section 2.2.).

The contrastive study also reveals some important differences in the distribution of aspect markers in Chinese vs. English and British vs. American English across fifteen text categories, as shown in Figure 46.2. The figure shows the distribution of aspect markers (expressed as percentages) across the fifteen text categories in the three corpora. As can be seen, by comparison to the two major varieties of English, aspect markers in Chinese occur more frequently in text categories G (biography) and K (general fiction) but less frequently in N (adventure fiction), L (mystery fiction), H (official documents), and E (hobby/skill/trade). British English and American English also differ in that the latter variety does not show such a marked fluctuation in aspect marking in narrative texts, notably in biography and the five types of fiction.

Further analysis of the corpus data reveals that in expository texts, perfective aspect markers in Chinese generally occur more frequently than those in English whereas in narrative texts, perfective markers in English are generally more frequent than those in Chinese. The relatively high frequency of perfective markers in narrative texts and their lower frequency in expository texts in English can be accounted for by the fact that aspect markers in English express both temporal and aspectual meanings because the aspect and tense markers in English combine morphologically. For example, over 80

percent of perfective markers in FLOB and Frown are simple past forms. Given that narrative texts usually relate to what happened in the past whilst expository texts are typically non-past, the relatively high frequency of perfective markers in narrative as opposed to expository texts in English is hardly surprising.

In marked contrast, imperfective aspect markers show a totally different distribution pattern from perfective markers. In expository texts, imperfective markers in both varieties of English typically occur more frequently than those in Chinese, whereas imperfective markers in Chinese are generally more frequent than those in English. This phenomenon can be explained as follows. First, the Chinese progressive marked by *zai* can only signal progressiveness literally. In contrast, “the progressive in English has a number of other specific uses that do not seem to fit under the general definition of progressiveness” (Comrie 1976, 37). Although the different uses of the progressive in English and Chinese account for the slightly higher frequency of the English imperfective markers in expository texts, this cannot explain the relatively low frequency of these markers in narrative texts. Nevertheless, we can find an answer in the Chinese imperfective marker *-zhe*, which accounts for 88 percent of the imperfective markers in the Chinese corpus. This marker has three basic functions: to signal the durative nature of a situation, to serve with a verb as an adverbial modifier to provide background information, and to occur in locative inversion to indicate existential status (Xiao/McEnery 2004b). Of the three functions of *-zhe*, only the first is used in expository texts. Hence, in spite of the high overall frequency of *-zhe* in LCMC, only about 20 percent of all examples of *-zhe* occur in expository texts. In contrast, all of the three functions of *-zhe* apply to narrative texts. Furthermore, in addition to inducing a background effect, *-zhe* can also be used in an apparently “foregrounded” situation to move narration forward (see Xiao/McEnery 2004b). As such, it is only natural to find that Chinese imperfective markers occur more frequently in narrative texts than English imperfective markers.

Using the same analytic framework, McEnery/Xiao (2002) and Xiao/McEnery (2005) explore, on the basis of aligned parallel corpora, how aspectual meanings in English are translated into Chinese. It is found that in English-Chinese translation, most progressives in English (over 58 percent) do not undergo a shift in viewpoint aspect, though some of the translations (about 15 percent) may take the unmarked form. Whether a viewpoint aspect shift occurs in translation depends largely on the specific use of the progressive in the English source data, and on the interaction between situation aspect and viewpoint aspect in the Chinese target language. This means that on the one hand, when progressives in the English source data that indicate habitual situations or anticipated happenings are translated into Chinese, they necessarily undergo a viewpoint aspect shift, because the progressive in Chinese does not indicate habituality or futurity. On the other hand, when a translation triggers a situation type shift into individual-level states (ILSs) or achievements in the Chinese translations, a viewpoint aspect shift is expected, because these two types of situations do not normally take the prototypical progressive.

When English perfect constructions are translated into Chinese, they more often than not depend on context to indicate the perfect meaning. This is because Chinese does not have a grammatical aspect marker for the perfect. In this case, however, aspect markers such as *-le* and *-guo* could be used to mark the perfect meaning. Whether the translations take overt aspect markers or imply the perfect meaning contextually depends largely on the type of perfect, i. e. the perfect of result, the perfect of experience, the perfect of

recent past, and the perfect of persistent situation (see Comrie 1976) in the English source texts.

The perfect progressive is an interaction between the perfect and the progressive. Chinese translations of the perfect progressive may shift towards the progressive or the perfect meaning, depending on the situation type involved and the translator's choice of viewpoint. But in most cases both perfect and progressive meanings can be retained, with the perfect being lexicalized by temporal adverbs and the progressive being signalled by the progressive aspect marker *zai* or implied by the context. The pluperfect progressive is similar to the perfect progressive with the exception that it signals progressiveness with a relatively past time reference. Situations referred to by the English pluperfect can be translated into Chinese with the progressive or the durative aspect unless the translator chooses to present them perfectly or there is a shift in situation type which prohibits them taking the progressive or the durative aspect.

Situations marked by the English simple aspect are mainly presented perfectly and most of them take the covert form in Chinese translations. The high frequency of perfectives in translations of the simple aspect can be accounted for by the fact that the simple forms in English, the simple past in particular, are basically perfective in nature (cf. Brinton 1988). Translations of the simple past show a marked/unmarked ratio twice as high as that in translations of the simple present. A natural explanation for this contrast is that the simple present typically denotes states, which do not have to be marked aspectually. Translations of the simple future frequently take modal auxiliaries or adverbs that lexicalize future time references. This is because modal and future meanings are closely related (cf. Comrie 1976).

In conclusion, the research on situation aspect, on aspect in Chinese, and the contrastive and translation studies reported in this section have demonstrated that corpus data can indeed be used to inform aspect theory. Our corpus-based aspect model has not only provided an explanatorily adequate account of aspect in Chinese, it is also a useful framework for contrastive language study and translation research. Methodologically, the case studies presented in this section show that intuitions and corpora are complementary rather than antagonistic. The two types of data must complement each other so as to circumvent their weaknesses if as broad a range of research questions as possible are to be addressed by linguists (cf. section 2.2.).

4. Conclusions

This article explored theory-driven corpus research, as exemplified by the case studies in section 3 as well as articles 42 and 43 in this volume. The discussion shows that if linguistics is defined as the study of language as reflected by our knowledge as well as use of language – which it should be – instead of as the study of “idealized language”, corpus data can indeed contribute to linguistic theory, because corpora can provide attested, contextualized and quantitative language data. As noted in section 2.2., intuitions and corpora have their own strengths and weaknesses, and the two are not mutually exclusive. Different research questions require different kinds of data. The key is to find the balance between the use of corpus data and the use of intuitions to suit the needs of the research question under consideration. It is also clear from the discussion

that the sharp distinction between the “corpus-based” and “corpus-driven” approaches is in reality overstated and that the theory-free corpus-driven linguistics is at best an idealized extreme. The different approaches may be more appropriate in different areas of studies (consider, for example, the roles played by intuitions and corpora in collocation and aspect studies; see article 58 for discussion of collocations), but corpora are what they are. They can be used to verify and revise existing linguistic theories, and they can also be used to provide what intuitions alone cannot discern, on the basis of which entirely new linguistic theories can be developed. The corpus-based research of aspect reported in this article demonstrates that the theory-driven and data-driven approaches can, and indeed should, complement each other in linguistic analysis to make linguistic theory true and interesting at the same time. As Fillmore (1992) expects, the best practice is for the armchair linguist and the corpus linguist to exist in the same body.

5. Literature

- Aarts, J. (2002), Review of *Corpus Linguistics at Work*. In: *International Journal of Corpus Linguistics* 7(1), 118–123.
- Abney, S. (1996), Statistical Methods and Linguistics. In: Klavans, J./Resnik, P. (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press, 1–26.
- Adger, D. (2002), Why Theory is Essential: The Relationship between Theory, Analysis and Data. In: Gallagher-Brett, A./Dickens, A./Canning, J. (eds.), *Guide to Good Practice for Learning and Teaching in Languages, Linguistics and Area Studies*. Southampton: The British Higher Education Academy Subject Centre for Languages, Linguistics and Area Studies. Available at: <http://www.lang.ltsn.ac.uk/resources/goodpractice.aspx?resourceid=405>.
- Andor, J. (2004), The Master and his Performance: An Interview with Noam Chomsky. In: *Intercultural Pragmatics* 1(1), 93–111.
- de Beaugrande, R. (1997), Theory and Practice in Applied Linguistics: Disconnection, Conflict, or Dialectic? In: *Applied Linguistics* 18(3), 279–313.
- de Beaugrande, R. (2002), Descriptive Linguistics at the Millennium: Corpus Data as Authentic Language. In: *Journal of Language and Linguistics* 1(2), 91–131.
- Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Brinton, L. (1988), *The Development of English Aspectual System*. Cambridge: Cambridge University Press.
- Chomsky, N. (1957), *Syntactic Structures*. The Hague: Mouton.
- Chomsky, N. (1962), Explanatory Models in Linguistics. In: Nagel, E./Suppes, P./Tarski, A. (eds.), *Logic, Methodology, and Philosophy of Science*. Stanford: Stanford University Press, 528–550.
- Chomsky N. (1965), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1969), Language and Philosophy. In: Hook, S. (ed.), *Language and Philosophy: A Symposium*. New York: New York University Press, 51–94.
- Chomsky, N. (1986), *Knowledge of Language*. New York: Praeger.
- Comrie, B. (1976), *Aspect*. Cambridge: Cambridge University Press.
- Elliot, D./Legum, S./Thompson, S. (1969), Syntactic Variation as Linguistic Data. In: Binnick, R./Davison, A./Green, G./Morgan, J. (eds.), *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 52–59.
- Fillmore, C. (1992), “Corpus Linguistics” or “Computer-aided Armchair Linguistics”. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin/New York: Mouton de Gruyter, 35–60.

- Firth, J. (1968), A Synopsis of Linguistic Theory. In: Palmer, F. (ed.), *Selected Papers of J.R. Firth 1952–59*. London: Longman, 168–205.
- Francis, G./Hunston, S./Manning, E. (1996), *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins.
- Francis, G./Hunston, S./Manning, E. (1998), *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.
- Gast, V. (2006), Introduction. In: *Zeitschrift für Anglistik und Amerikanistik. Special Issue on the Scope and Limits of Corpus Linguistics* 54(2), 13–20.
- Gries, S. (2006), Some Proposals towards More Rigorous Corpus Linguistics. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 191–202.
- Hoey, M. (1997), From Concordance to Text Structure: New Uses for Computer Corpora. In: Lewandowska-Tomaszczyk, B./Melia, J. (eds.), *PALC '97: Proceedings of Practical Applications in Linguistic Corpora Conference*. Łódź: University of Łódź, 2–23.
- Hoey, M. (2000), A World beyond Collocation: New Perspectives on Vocabulary Teaching. In: Lewis, M. (ed.), *Teaching Collocations*. Hove: Language Teaching Publications, 224–243.
- Hopper, P. (1979), Aspect and Foregrounding in Discourse. In: Givon, T. (ed.), *Syntax and Semantics* (Volume 12) – *Discourse and Syntax*. New York: Academic Press, 213–241.
- Johansson, S. (1995), Mens sana in corpore sano: On the Role of Corpora in Linguistic Research. In: *ESSE Messenger* IV(2), 19–25.
- Kang, J. (1999), The Composition of the Perfective Aspect in Mandarin Chinese. PhD thesis, Boston University.
- Kennedy, G. (1998), *An Introduction to Corpus Linguistics*. London: Longman.
- Klein, W./Li, P./Hendriks, H. (2000), Aspect and Assertion in Mandarin Chinese. In: *Natural Language and Linguistics Theory* 18, 723–770.
- Krishnamurthy, R. (2000a), Collocation: From *silly ass* to Lexical Sets. In: Heffer, C./Sauntson, H./Fox, G. (eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, 31–47.
- Krishnamurthy, R. (2000b), Size Matters: Creating Dictionaries from the World's Largest Corpus. In: *Proceedings of KOTESOL 2000 – Casting the Net: Diversity in Language Learning*. Taegu, Korea, 169–180.
- Labov, W. (1972), *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, W. (1975), *What is a Linguistic Fact?* Lisse: Peter de Ridder Press.
- Leech, G. (1991), The State of the Art in Corpus Linguistics. In: Aijmer, K./Altenberg, B. (eds.), *English Corpus Linguistics*. London: Longman, 8–29.
- Leech, G. (1992), Corpora and Theories of Linguistic Performance. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter, 105–122.
- Leech, G. (2000), Grammars of Spoken English: New Outcomes of Corpus-oriented Research. In: *Language Learning* 50(4), 675–724.
- Léon, J. (2005), Claimed and Unclaimed Sources of *Corpus Linguistics*. In: *Henry Sweet Society Bulletin* 44, 36–50.
- Li, M. (1999), Negation in Chinese. PhD thesis, University of Manchester.
- Louw, B. (1991), Classroom Concordancing of Delexical Forms and the Case for Integrating Language and Literature. In: Johns, T./King, P. (eds.), *Classroom Concordancing. ELR Journal* 4. Birmingham: CELS University of Birmingham, 151–178.
- Louw, B. (1993), Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker, M./Francis, G./Tognini Bonelli, E. (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 157–176.
- Louw, B. (2000), Contextual Prosodic Theory: Bringing Semantic Prosodies to Life. In: Heffer, C./Sauntson, H./Fox, G. (eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, 48–94.

- McCarthy, M./Carter, R. (2001), Size isn't Everything: Spoken English, Corpus and the Classroom. In: *TESOL Quarterly* 35(2), 337–340.
- McEnery, A./Wilson, A. (2001), *Corpus Linguistics* (1st ed. 1996). Edinburgh: Edinburgh University Press.
- McEnery, A./Xiao, R. (2002), Domains, Text Types, Aspect Marking and English-Chinese Translation. In: *Languages in Contrast* 2(2), 51–69.
- McEnery, A./Xiao, R./Mo, L. (2003), Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study. In: *Literary and Linguistic Computing* 18(4), 361–378.
- McEnery, A./Xiao, R./Tono, Y. (2006), *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
- Meyer, C. (2002), *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Meyer, C./Nelson, G. (2006), Data Collection. In: Aarts, B./McMahon, A. (eds.), *The Handbook of English Linguistics*. Oxford: Blackwell, 93–113.
- Murison-Bowie, S. (1996), Linguistic Corpora and Language Teaching. In: *Annual Review of Applied Linguistics* 16, 182–199.
- Newmeyer, F. (2003), Grammar is Grammar and Usage is Usage. In: *Language* 79(4), 682–707.
- Pan, H. (1993), Interaction between Adverbial Quantification and Perfective Aspect. In: Stvan, L./Ryberg, S./Olsen, M. B./Macfarland, T./DiDesidero, L./Bertram, A./Adams, L. (eds.), *Proceedings of the Third Annual Formal Linguistics Society of Mid-America Conference*. Northwestern University, Bloomington, IN: Indiana University Linguistics Club Publications, 188–204.
- Partington, A. (1998), *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Partington, A. (2004), “Utterly content in each other's company”: Semantic Prosody and Semantic Preference. In: *International Journal of Corpus Linguistics* 9(1), 131–156.
- Radford, A. (1997), *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.
- Sampson, G. (2001), *Empirical Linguistics*. London: Continuum.
- Saussure, F. (1916 [1966]), *Course in General Linguistics*. New York: McGraw-Hill.
- Schütze, C. (1996), *The Empirical Base of Linguistics*. Chicago: University of Chicago Press.
- Seuren, P. (1998), *Western Linguistics: A Historical Introduction*. Oxford: Blackwell.
- Shei, C. (2004), Corpus and Grammar: What it isn't. In: *Concentric: Studies in Linguistics* 30(1), 1–18.
- Sinclair, J. (1997), Corpus Evidence in Language Description. In: Wichmann, A./Fligelstone, S./McEnery, T./Knowles, G. (eds.), *Teaching and Language Corpora*. London: Longman, 27–39.
- Smith, C. (1997), *The Parameter of Aspect* (1st ed. 1991). Dordrecht: Kluwer.
- Stefanowitsch, A. (2006), Negative Evidence and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 2(1), 61–77.
- Stubbs, M. (1995), Collocation and Semantic Profiles: On the Cause of the Trouble with Quantitative Methods. In: *Function of Language* 2(1), 23–55.
- Stubbs, M. (2001a), Texts, Corpora, and Problems of Interpretation: A Response to Widdowson. In: *Applied Linguistics* 22(2), 149–172.
- Stubbs, M. (2001b), *Words and Phrases*. Oxford: Blackwell.
- Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Vendler, Z. (1967), *Linguistics in Philosophy*. New York: Cornell University Press.
- Verkuyl, H. (1993), *A Theory of Aspectuality*. Cambridge: Cambridge University Press.
- Wasow, T./Arnold, J. (2005), Intuitions in Linguistic Argumentation. In: *Lingua* 115, 1481–1496.
- Widdowson, H. (2000), The Limitations Of Linguistics Applied. In: *Applied Linguistics* 21(1), 3–25.
- Xiao, R. (2002), A Corpus-based Study of Aspect in Mandarin Chinese. PhD thesis, Lancaster University.
- Xiao, R./McEnery, A. (2002), Situation Aspect as a Universal Aspect: Implications for Artificial Languages. In: *Journal of Universal Language* 3(2), 139–177.

- Xiao, R./McEnery, A. (2004a), A Corpus-based Two-level Model of Situation Aspect. In: *Journal of Linguistics* 40(2), 325–363.
- Xiao, R./McEnery, A. (2004b), *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins.
- Xiao, R./McEnery, A. (2005), A Corpus-based Approach to Tense and Aspect in English-Chinese Translation. In: Pan, W./Fu, H./Luo, X./Chase, M./Walls, J. (eds.), *Translation and Contrastive Studies*. Shanghai: Shanghai Foreign Language Education Press, 114–157.
- Xiao, R./McEnery, A. (2006a), Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. In: *Applied Linguistics* 27(1), 103–129.
- Xiao, R./McEnery, A. (2006b), Can Completive and Durative Adverbials Function as Tests for Telicity? Evidence from English and Chinese. In: *Corpus Linguistics and Linguistic Theory* 2(1), 1–21.
- Yang, S. (1995), The Aspects System of Chinese. PhD thesis, University of Victoria.

Richard Xiao, Ormskirk (UK)

47. Corpora and spoken language

1. Introduction: Evolution of spoken corpora
2. Corpora for studying language varieties and types of discourse
3. Size, representativeness, transcription, and other issues
4. Research: Important findings
5. Applications of spoken corpus research
6. Directions in spoken corpus linguistics
7. Literature

1. Introduction: Evolution of spoken corpora

Spoken corpora have evolved over the last four decades from early attempts at corpus-building for the purposes of better understanding such phenomena as first-language acquisition, social variation and conversational structure, to the large, general spoken corpora of today, which have found applications in a variety of contexts from speech recognition, lexicography, sociolinguistics and first and second language acquisition. In this article we focus on spoken corpora and their applications in linguistics and applied linguistics, rather than on ‘speech corpora’, which are typically collected for the purposes of improving technology, a distinction discussed at greater length by Wichmann in article 11; see also article 30.

Some of the earliest spoken corpora were developed within the field of child language acquisition, an example of which was the child-language word-frequency analyses described in Beier/Starkweather/Miller (1967). Another example, which included informal spoken language by adults, as well as by selected age groups of children from six years upwards in a corpus of some 84,000 words, is described in Carterette/Jones (1974). A

- Xiao, R./McEnery, A. (2004a), A Corpus-based Two-level Model of Situation Aspect. In: *Journal of Linguistics* 40(2), 325–363.
- Xiao, R./McEnery, A. (2004b), *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins.
- Xiao, R./McEnery, A. (2005), A Corpus-based Approach to Tense and Aspect in English-Chinese Translation. In: Pan, W./Fu, H./Luo, X./Chase, M./Walls, J. (eds.), *Translation and Contrastive Studies*. Shanghai: Shanghai Foreign Language Education Press, 114–157.
- Xiao, R./McEnery, A. (2006a), Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective. In: *Applied Linguistics* 27(1), 103–129.
- Xiao, R./McEnery, A. (2006b), Can Completive and Durative Adverbials Function as Tests for Telicity? Evidence from English and Chinese. In: *Corpus Linguistics and Linguistic Theory* 2(1), 1–21.
- Yang, S. (1995), The Aspects System of Chinese. PhD thesis, University of Victoria.

Richard Xiao, Ormskirk (UK)

47. Corpora and spoken language

1. Introduction: Evolution of spoken corpora
2. Corpora for studying language varieties and types of discourse
3. Size, representativeness, transcription, and other issues
4. Research: Important findings
5. Applications of spoken corpus research
6. Directions in spoken corpus linguistics
7. Literature

1. Introduction: Evolution of spoken corpora

Spoken corpora have evolved over the last four decades from early attempts at corpus-building for the purposes of better understanding such phenomena as first-language acquisition, social variation and conversational structure, to the large, general spoken corpora of today, which have found applications in a variety of contexts from speech recognition, lexicography, sociolinguistics and first and second language acquisition. In this article we focus on spoken corpora and their applications in linguistics and applied linguistics, rather than on ‘speech corpora’, which are typically collected for the purposes of improving technology, a distinction discussed at greater length by Wichmann in article 11; see also article 30.

Some of the earliest spoken corpora were developed within the field of child language acquisition, an example of which was the child-language word-frequency analyses described in Beier/Starkweather/Miller (1967). Another example, which included informal spoken language by adults, as well as by selected age groups of children from six years upwards in a corpus of some 84,000 words, is described in Carterette/Jones (1974). A

notable early spoken corpus project of the kind that has since become quite common was the *Oral Vocabulary of the Australian Worker* (*OVAW*), for which Schonell et al. (1956) give a full account of the data and its collection. The *OVAW* corpus consisted of some 500,000 words of spoken language and was used for, among other things, the study of idiomatic words and phrases in speech. A decade after *OVAW*, the Davis-Howes Count of Spoken English (Howes 1966) in the USA brought together half a million words of interviews with university students and hospital patients, and presented statistics for spoken usage. An influential early spoken corpus of British English was the London-Lund corpus (LLC). This corpus derives from two projects: the Survey of English Usage (SEU) at University College London, launched in 1959 by Randolph Quirk, and the Survey of Spoken English (SSE), which was started by Jan Svartvik at Lund University in 1975. The London-Lund corpus, which is the spoken component of the Survey of English Usage, comprises half a million words. Its goal is to provide a resource for accurate descriptions of the grammar of adult educated speakers of English (Svartvik 1990; Edwards/Lampert 1993). The spoken English component comprises both dialogue and monologue and was collected over a 30-year period towards the end of the last century.

Several other early spoken corpora for English were developed as appendages to much larger written corpora, a reflection of the time and expense involved in collecting such data relative to written texts. Major corpus projects such as the COBUILD Bank of English (see Moon 1997) and the British National Corpus (see Crowley 1993, 1994; Rundell 1995a, 1995b) contain large spoken segments, including broadcast speech as well as everyday unrehearsed conversation. The British National Corpus (BNC) contains over 100 million words of data, with the spoken component accounting for around ten million words. The spoken data consist of unscripted informal conversations recorded by volunteers selected from different ages, regions and social classes in a demographically balanced way. It is designed to represent as wide a range of modern British English as possible.

In the USA, work by Chafe and his colleagues, initially based on the British London-Lund spoken corpus design (Chafe/Du Bois/Thompson 1991), developed into larger corpus enterprises such as the five-million-word Longman Spoken American Corpus (see Stern 1997). Informal Australian spoken English has also been subjected to corpus analysis more recently by Eggins/Slade (1997), who look at everyday conversational activities such as gossiping. Also important is the ICE (International Corpus of English) project, which plans to bring together parallel corpora of one million words from 18 different countries where English is either the main language or an official language. The samples in the ICE corpus include 300 spoken texts, although these include many scripted samples, and broadcast interviews and discussions, with only 90 samples being face-to-face informal conversations (see Nelson 1996; also Fang 1995).

In 1984, Knowles, Alderson, Williams, Taylor, Leech and Kaye embarked on a joint research project between the University of Lancaster and the Speech Research Group at IBM UK Scientific Centre into the automatic assignment of intonation. The first aim of the project was to collect samples of natural spoken British English which could be used as a database for analysis and for testing the intonation assignment programs. The result was the Spoken English Corpus (SEC), a machine-readable corpus of approximately 52,000 words of spoken British English. The majority of texts in the corpus were obtained from the BBC and include news broadcasts, commentary, religious broadcast,

magazine-style reporting as well as fiction, poetry and dialogue (see Knowles 1990). Leech (2000) notes that while the LLC and the SEC benefited from careful and detailed prosodic transcription, they suffer from restrictions owing to the data they contain. The LLC, for example, used heavy reel-to-reel tape recorders, and a considerable portion of the spontaneous dialogue data is restricted to academic settings among staff and students at London University, and so academic topics of conversation prevail, while the SEC is even narrower in range as its recordings are mostly confined to scripted speech such as radio broadcasts.

In 1993, Stenström and Breivik set up the Bergen Corpus of London Teenage Language (COLT). The aim of the project was to create a corpus of British English teenage talk and make it available for research. The corpus designers believed that studying spontaneous teenage talk would yield insights into language development and language change, especially as regards grammaticalisation (see Breivik/Hasselgren 2002). The reason for restricting the corpus collection to London was the assumption that new trends predominate among teenagers in the capital, from where they can be expected to spread to the rest of the country, and even further afield. Stenström/Andersen/Hasund (2002) provide an extensive study of the COLT data, outlining the most prominent features of the teenagers' talk including 'slanguage', speech reporting, non-standard grammatical features, intensifiers, tags, and interactional behaviour in terms of conflict talk.

In 2000, Leech noted that "it may seem strange that the United States, where the age of English electronic corpora began with the Brown Corpus (in 1961), has held back from the development of wide-coverage spoken corpora" (Leech 2000, 684). This may be due to the long shadow cast by the general rejection of corpus data by Chomskyan linguists, Leech (2000) surmises. Offsetting the earlier lack of a major spoken corpus project, the American National Corpus (ANC) was set up as a comparative corpus to the BNC (Ide/Macleod 2001; Ide/Reppen/Suderman 2002). In 2003, a pilot sample of 11 million words was released. This comprised over 3 million words of spoken data and over 8 million words of written texts. The spoken data came from three sources: 1.5% from 'Callhome' (10 minute segments of telephone conversations), 95% from 'Switchboard' (2320 spontaneous telephone conversations averaging six minutes in length and comprising about 3 million words by over 500 speakers) and 3.5% from 'Charlotte Narratives' (95 narratives, conversations and interviews representative of the residents of Mecklenburg County, North Carolina and surrounding North Carolina communities). The full corpus, consisting of (at least) 100 million words annotated for part of speech, together with search and retrieval software, was expected to be in place in the fall of 2005 (see the ANC website <http://americanationalcorpus.org/> 2004).

As in so many other aspects of linguistic study, English tended to dominate spoken corpus building in the earlier years, but spoken corpora for many other languages now exist, including Bulgarian, French (both European and Canadian), Mandarin Chinese, Vietnamese, Egyptian Arabic, Farsi, German, Greek, European Spanish, Hindi, Japanese, Korean, Tamil and Vietnamese, amongst others. Many of these are available from the Linguistic Data Consortium at the University of Pennsylvania (see www.ldc.upenn.edu); ELDA, the Evaluations and Language resources Distribution Agency in Europe also makes available a number of spoken corpus resources in different languages (see www.elda.org); see also section 5.

2. Corpora for studying language varieties and types of discourse

The International Corpus of English (ICE) project was launched in 1991 by Sidney Greenbaum (see Greenbaum 1991, 1992). His initial goal was to gather at least 15 regional components from countries where English was the ‘native language’ as well as countries where it was an “official non-native language – India, Nigeria and the Philippines” (Greenbaum 1992, 171). Each corpus would comprise one million words of spoken and written material and the same template would be used throughout in the compilation and collection of data. The goal was “to provide the means for comparative studies” and for the first time provide “the resources for systematic study of the national variety as an end in itself” (Greenbaum 1992, 171). This project has led to the collection of spoken and written data for the Englishes of Hong Kong (Bolton et al. 2003), New Zealand (Holmes 1996), Singapore (Ooi 1997), Great Britain (Nelson/Wallis/Aarts 2002), Ireland (Kallen/Kirk 2001), Nigeria (Banjo 1996), East Africa (Schmied/Hudson-Ettele 1996) and the Caribbean (Nero 2000), with others under development.

In recent years a number of spoken corpora have been assembled with the express purpose of the study of aspects of spoken discourse in both formal and informal settings. The design principles of such corpora differ from spoken corpora collected for more general purposes. One such example is the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five-million-word collection of spoken data. It is designed so as to represent spoken language in different contexts of use, genres of speech and between different speaker relationships across the islands of Britain and Ireland (see McCarthy 1998). The corpus design focuses on representing a range of discourse contexts and speech genres across different speaker relationships with the aim of informing research and language pedagogy in the fields of lexis, grammar and discourse. Using the same design matrix, the Limerick Corpus of Irish English (LCIE) comprises one million words of Irish English conversations (see Farr/Murphy/O’Keeffe 2002). Other discourse-oriented corpora include that described by Cheng/Warren, who oversaw the collection of the two-million-word Hong Kong Corpus of Spoken English (HKCSE) (see Cheng/Warren 1999, 2000).

Spoken corpora focusing on institutional settings include the Michigan Corpus of Academic Spoken English (MICASE) (Simpson/Lucka/Ovens 2000), offering by 2004 online access to more than 150 transcripts of academic speech events recorded at the University of Michigan, USA (totalling 1.8 million words). MICASE was established in 1997 with the goal of describing the characteristics of contemporary academic speech and any potential differences across academic disciplines and different classes of speakers. The MICASE data consist of speech within the microcosm of the University of Michigan at Ann Arbor. Speakers represented in the corpus include faculty, staff, and all levels of students, and both native and non-native speakers. The contexts in which the recordings were made include large lectures, discussions, seminars, student presentations, advising sessions, dissertation defences, interviews, meetings, office hours, service encounters, study groups, tours and tutorials. Farr (2003) looks at a corpus of spoken encounters in the context of teacher education consisting of post-observation trainer-trainee interactions (the POTTI corpus) in a university setting. The Cambridge and Nottingham Business English Corpus (CANBEC), a one-million-word corpus of conversations in business contexts (see Handford/McCarthy 2004; O’Keeffe/McCarthy/Carter 2007), and the Corpus of Spoken Professional American English (CSPA), a two-mil-

lion-word corpus consisting of 50 per cent White House press briefings and 50 per cent university academic council meetings (Barlow 1998) are also recent examples of specialised, targeted spoken corpora. Within the field of language pedagogy, learner spoken data have been collected, a notable example being the Louvain International Database of Spoken English Interlanguage (LINDSEI) set up in 1995 (see De Cock 1998, 2000), which provides spoken data for the analysis of the speech of second language learners (see also Granger/Hung/Petch-Tyson 2002).

3. Size, representativeness, transcription, and other issues

Spoken corpora, because of collection and transcription problems and financial and time constraints, inevitably tend to be much smaller than general written corpora. However, this is not always necessarily seen as a disadvantage. Leech (2000) notes that more important than size for assessing the research value of a corpus is its composition in terms of genres and other design features. Furthermore, a number of researchers have noted the value of small corpora for particular kinds of research (Carter/McCarthy 1995; McCarthy/Carter 2001a; Cameron/Deignan 2003). O'Keeffe/Farr (2003) suggest the following guidelines: for spoken corpora, anything over one million words is considered to be moving into the 'larger' range, for written anything below five million is quite small. McCarthy/Carter (2001a), arguing for more qualitative research (as opposed to the quantitative tradition) in corpus linguistics support the view that small spoken corpora can be used to great effect, especially where high-frequency linguistic items and features are concerned.

Various perspectives on how a corpus should be designed concur that it should be a principled collection of texts that is assembled for a specific purpose. Sinclair (1995) sees a corpus as something that is not a random assortment of data but a collection of pieces of language that are selected and ordered to explicit linguistic criteria to be later used as a sample of the language in question (see also Francis 1982; Atkins/Clear/Ostler 1992; Crowley 1994; Biber/Conrad/Reppen 1998; Tognini Bonelli 2001). Three criteria generally prevail in the literature as regards good corpus design: 1) authenticity of the texts, 2) representativeness of language included in corpus and 3) sampling criteria used in the selection of texts (Tognini Bonelli 2001, 54). Hunston (see article 9) offers, in addition to the criteria of size and the problematic notion of representativeness, the criterion of balance, that is to say ensuring equality in the sizes of the sub-corpora that make up the whole corpus (see Hunston's discussion of the MICASE spoken academic corpus). Decisions regarding the representativeness and balance of written corpora may be largely resolved by recourse to text typologies (see Crystal 1995; Lee 2001, Aston 2001) and ensuring that the corpus includes a broad coverage of text types in substantial and balanced quantities. In the case of the design of spoken corpora, however, not least of the problems is deciding precisely what constitutes a text. Written texts have clear orthographic boundaries, which spoken texts do not. In the case of casual conversation, topical segments blend into one another, paragraphs and sentences are a mere artefact of transcription and, except in the case of extended monologue, more than one speaker contributes to the text, often simultaneously. Two main non-text-based solutions to these problems are therefore commonly pursued. One is to collect demographically stratified

samples of undetermined (or arbitrarily chosen) length which may be to a greater or lesser extent clearly delineated in terms of boundary phenomena such as conversational openings and closings. For example some spoken corpora aim to represent a language variety, e. g. the British National Corpus (BNC), and therefore need to give careful attention to the collection of data across a representative balance of standard demographic sampling variables for example gender, age, region, social class, etc. The Corpus of London Teenage speech (COLT) on the other hand only sought to represent one age group in one region, so while COLT modelled its design principles on the BNC, it limited its scope to a sample of teenagers in the London area. During a three-week period, using a network of London schools, students carried a small recording device and a lapel microphone for a few days and recorded all the conversations they took part in, with friends of the same age who were not supposed to be aware of the recording. The recruits were also equipped with a logbook and instructed to write down information about the co-speaker(s) and the setting. In three weeks all 0.5 million words of spoken language were collected (see also article 11). The other, not mutually exclusive solution to the problem of delineating data samples is to take a context- or genre-based approach, in which spoken samples are collected based on a pre-determined set of situational parameters. Corpora such as CANCODE, LCIE and HKCSE set out to examine English spoken discourse in specified contexts rather than to describe a language variety. In such cases, a highly representative corpus is not necessarily one which adheres to demographic sampling principles, but rather one which is based on representing the genres of spoken language itself (article 11 gives further examples of genre-based approaches to spoken corpus design). The five-million-word CANCODE spoken corpus, for example, was designed so as to represent everyday spoken language across different genres and speaker relationships. The design of CANCODE as described in McCarthy (1998) was based on a matrix with two axes for classification: *context type* and *interaction type*. Context type distinguished texts that were predominantly collaborative and those that were non-collaborative. The collaborative types were classified as *ideas* (e. g. exchanging opinions) and *tasks* (engagement in some physical task, e. g. doing the washing up) whereas the non-collaborative types were more asymmetrical and were classified as *information provision*. The interaction types reflected the relationship between the conversational participants. These fell into five broad categories: intimate, socialising, professional, transactional and pedagogic. LCIE used the same design principle with the same goal, and because these two corpora use the same design principles they have lent themselves to comparisons across two varieties (e. g. McCarthy/O'Keeffe, 2003). The HKCSE is also genre-based and includes Hong Kong Chinese speakers of English and native speakers of English. It is made up of four sub-corpora each comprising 0.5 million words, under the headings of conversations, academic discourses, business discourses, and public discourses. The data are transcribed both orthographically and prosodically.

Transcription of spoken corpora is, as Holmes/Vine/Johnson (1998) put it, the art of making the ephemeral tangible in a consistent and practical manner. In reality the spoken word is very difficult to make tangible in written form as one immediately loses the audio and visual component in which it had its existence. Transcription has been the cause of much discussion and debate (see for example Ochs 1979; Edwards 1991; Cook 1990; Edwards/Lampert 1993; Bucholtz 2000; Hepburn 2004). Duranti (1997) suggests that transcripts are inherently incomplete and should be continuously revised to display features of an interaction that have been illuminated by a particular analysis and allow

for new insights that might lead to a new analysis. Much of the already extant detailed work by conversation analysts has informed corpus transcription techniques over the years. For example, Jefferson (1985) provided a comprehensive account of laughter using a corpus of phone calls to a child protection helpline; Hepburn (2004), building on the work of Jefferson, examines crying using a corpus of calls to Child Protection Officers at the British National Society for the Prevention of Cruelty to Children. Hitherto, she points out, crying was considered as a unitary and self-evident category where it was uncommon for transcription to try and capture its different elements. Her work makes explicit some different elements of crying and shows how these elements can be represented in transcription, for example sniffing, wobbly voice, high pitch, aspiration, sobbing and silence. Despite such detailed attention to potential features for transcription, however, large corpora tend to remain only broadly transcribed. Leech (2000) notes that many of the large spoken corpora were built primarily for the purpose of English language dictionaries and were transcribed quickly and at a low unit cost, which means a simple orthographic transcription. One of the consequences of such “basic” transcriptions” (Leech 2000, 678) is that while lexis and grammar can be investigated, key aspects of spoken language such as prosody and discourse cannot, due to the absence of accurate and detailed phonological, contextual and turn-taking information. For this reason Leech (*ibid.*) notes that “even at a time when the availability of machine-readable corpora has brought a vast increase of knowledge about the spoken language within our grasp, the influence and limitations of the written language continue to impinge on the spoken medium”. Cheng/Warren (2002) also note that while the orthographic transcription of spoken data is well established and the conventions quite well-known, the number of spoken corpora that are also prosodically transcribed is very small, a well-known exception being the London-Lund Corpus of Spoken English (Svartvik/Quirk 1980; Svartvik 1990). Cheng/Warren point out that the representation of prosodic features is less standardized, that it is notoriously difficult and time-consuming to prosodically transcribe naturally-occurring data, and that it ideally requires inter-rater reliability measures to ensure the quality of the transcription. Articles 11 and 30 offer further discussion of transcription and annotation issues, especially those generated by the different purposes and applications which speech databases and spoken corpora, as demarcated at the beginning of this article, typically serve.

4. Research: Important findings

One far-reaching impact of the availability of spoken corpora can be seen in the attempts to elaborate independent descriptions of spoken grammar (Leech 2000). The availability of spoken corpus data brought to light the fact that written models were not always adequate to describe spoken usage. While, in the case of English and many other languages, the actual forms of grammar are to a very great extent shared between the spoken and written media, and while, potentially, any grammatical form may occur in either medium, the distribution of forms in actual fact is often markedly different across the two media (see Blanche-Benveniste 1995; Fonseca-Greber/Waugh 2003 for examples from French). Phenomena such as so-called left- and right-dislocated items (otherwise known as pre- and post-posed items) and situational ellipsis (e. g. non-use of otherwise

obligatory forms such as verb subjects or determiners) are common in casual spoken data but extremely rare in most kinds of formal writing (Carter/McCarthy 1995, 2006). In this extract from the CANCODE spoken corpus, the speakers are looking at photographs; ellipsis of *the* occurs before *same* in <\$1>'s turn, ellipsis of *you've* occurs before *seen* in <\$4>'s turn, and *that* in <\$5>'s turn is post-posed: (<\$#> = speaker number in order of speaking; <\$?F> = speaker unidentifiable, probably female; <\$E> <\\$E> beginning and end of non-verbal event (e. g. laughter))

<\$4> Oh. I'm like my father there aren't I.
<\$?F> You can't do anything about that now.
<\$1> Yes. Mm. Same eyes look. Same shape.
<\$4> Seen that one of Jim haven't you.
<\$?F> <\$E> laughs <\$E>
<\$5> Yeah. It's a good one that.

Furthermore, the descriptive apparatus and terminology itself is called into question in the face of spoken corpus evidence. The notion of 'subordination' as it has derived from the intuition of grammarians or from the observation of written texts has come under close scrutiny (Blanche-Benveniste 1982). McCarthy (2001, 128) points out that metaphors such as 'left' and 'right' (as used to refer to dislocated elements) are western-culture, page-driven ones, and that a different metalanguage is called for when spoken data is described. Similarly, 'ellipsis' is based on a notion of the absence of obligatory items, whereas face-to-face interaction proceeds unproblematically with often only minimal use of so-called 'obligatory' elements. Leech (2000) however cautions against assuming that the grammars of spoken and written language are radically different. He argues that spoken and written language utilise the same basic grammatical repertoire, though its implementation may differ. Speech, according to Leech, shows a tendency to simplified, loosely integrated and disjunctive construction (see Chafe 1982; Halliday 1989), giving grammatical structure a lesser role in the overall communication process than is characteristic of writing, something which can only be fully implemented by corpus research.

Alongside and emerging from grammatical research, studies of the spoken lexicon have suggested that the core, heavy-duty vocabulary of everyday spoken interaction is smaller than that of mainstream written texts, but that, importantly, the phenomenon of 'chunking' (i. e. recurrence of strings of two or more words) is more widespread in spoken data. Chunks, or lexical bundles (Biber/Conrad/Reppen 1998; McCarthy/Carter 2002) are also different in kind across spoken and written corpora. While both types of corpora throw up common chunks characterised by syntactic fragments functioning as clause- or sentence-frames (e. g. *I don't know if ...*, *at a time when ...*), there are notable differences between spoken and written data. Predominantly, the two-, three-, four- and five-word chunks found in written corpora tend to be prepositional phrases referring to basic notional categories such as time, place, manner, etc., or else determiner phrases (e. g. *one of the*), or adverbial phrases expressing various inter-clausal relations (e. g. *on the other hand*). Spoken chunks are dominated by interactional discourse marking expressions such as *you know what I mean* and vague expressions such as *or something like that* (McCarthy/Carter 2001b; McCarthy/Carter 2002; O'Keeffe 2004). The ubiquitous evidence of chunking in spoken corpora has contributed to debates on key aspects

of language processing and the notion of fluency, not only in monolingual contexts (Schmitt/Carter 2004) but also across languages (Spötl/McCarthy 2004).

Grammatical and lexical studies based on spoken corpora have developed in tandem with studies of discoursal and pragmatic aspects of spoken language. Difficulties persist in areas such as the automatic coding and retrieval of features such as speech acts and figures of speech, but, nonetheless, spoken corpora have been effectively exploited to investigate the reality of the everyday performance of common speech acts (Aijmer 1996), in contrast to the previous tradition within pragmatics of using intuitive data or elicitation instruments such as discourse completion tasks (DCTs). Aspects of turn-management have been investigated quantitatively by Tao (2003), and vocative address terms have been described, based on corpus evidence (Leech 1999; McCarthy/O'Keeffe 2003). McCarthy (2003) used the CANCODE corpus to investigate short listener responses (e. g. *right, fine, great, that's true*), a discourse feature in large part automatically retrievable by searching for single-word or very brief speaker turns. Meanwhile Aijmer (2002) used the LLC to examine 'discourse particles' (e. g. *now, oh, just, sort of, and that sort of thing, actually*), showing how the methods and tools of corpus analysis can sharpen their description. Aijmer illustrated the importance of linguistic and contextual cues such as text type, position in the discourse, prosody and collocation in the analysis of these items, hence the need to use a corpus which incorporated a detailed prosodic transcription system. Farr/O'Keeffe (2002) looked at the pragmatics of hedging in spoken Irish English. More diffuse but equally fundamental linguistic phenomena such as metaphor (Cameron/Deignan 2003), irony (Clift 1999), hyperbole (McCarthy/Carter 2004) and general conversational creativity (Carter/McCarthy 2004) have also been investigated and described using spoken corpora analysed through a combination of automatic retrieval of items (e. g. transcribed laughter, coded turn-overlaps, etc.) and manual searching, see O'Keeffe/McCarthy/Carter (2007) for specific examples.

5. Applications of spoken corpus research

Spoken corpora are increasingly used in diverse areas. These include forensic linguistics, for example in relation to forensic phonetics (e. g. speaker identification), the language of police confession, interrogation and deception (Shuy 1998), courtroom discourse (Cotterill 2002a, 2002b, 2003, 2004). Boucher (2005), in his analysis of features of deceit in recounting, compared a corpus of 200 three-to-five minute discourses where half represented truthful and half inaccurate accounts. He was able to statistically describe significant differences in variables such as hesitation, lexical repetition and utterance length.

Given that corpora can be built around variables such as age, gender, level of education and socio-economic background, the area of sociolinguistics, not surprising, is one where there is increasing use for spoken corpora. For example, Ihlainen (1991a) looked at regional variation in verb patterns in south-western British English, while Ihlainen (1991b) compared the grammatical subject in educated and dialectal English in the London-Lund and the Helsinki Corpus of British English dialects. Kirk (1992, 1999) and Kallen/Kirk (2001) look at languages in contact in the context of Northern Ireland and Irish English, Ulster Scots, Irish and Scots Gaelic using a corpus-based approach.

Age-related research is prevalent especially in the context of teenager language. The Corpus of London Teenage Language (COLT) (see Haslerud/Stenström 1995; Stenström

1998) has provided the basis for numerous studies. Features such as discourse markers have been given particular attention, for example Andersen (1997a, 1997b) on the use of *like* in London teenage speech, Stenström (1995, 1997a) and Stenström/Andersen/Hasund (2002) on the use of tags and taboo language, Hasund (1998) on class-determined variation in the verbal disputes of London teenage girls, Hasund/Stenström (1997) on conflict talk using a corpus-based comparison of the verbal disputes of adolescent females. Other corpus-based studies on language and gender include Aijmer (1995) which looks at apologies, Holmes (2001) which examines linguistic sexism and Mondorf (2002), a study of gender differences in English syntax.

Lapidus/Otheguy (2005), in a New York corpus-based study, look at language contact in the context of English and Spanish. They focus on the use of non-specific *ellos* (English equivalent: *they*). One of Lapidus/Otheguy's main conclusions is that the susceptibility of language varieties to contact influence is primarily at the discourse-pragmatic level.

In the second language pedagogical context, studies often illustrate how far the spoken language presented in textbooks for learners can be at odds with evidence from spoken corpora. Boxer/Pickering (1995), for example, looked at speech acts in textbook dialogues in comparison with real spontaneous encounters found in a corpus, while Carter (1998) found that textbook dialogues lacked core spoken language features such as discourse markers, vague language, ellipsis and hedges when compared to spoken corpus data (see also Gilmore 2004). Likewise, Hughes/McCarthy (1998) look at a range of grammatical items from the stock-in-trade of English as a Second Language pedagogy and argue that their distributions and functions in spoken language, based on corpus evidence, are often different from those focused on in pedagogy.

Recent years have seen a debate over the use of native-speaker corpora versus learner corpora and non-native speaker corpora in the pedagogy of English as a second language (Prodromou 1997, 2003; Seidlhofer 2001; Gut 2006).

Written corpora tend to be more homogenous and usually include texts aimed at a very wide readership, whereas spoken corpora (especially informal conversational ones) inevitably reflect very localised conditions and reflect the high levels of context-dependence and shared understandings typical of face-to-face speech. In the case of English, the issue is further complicated by the fact that the language has acquired the status of an international lingua franca, where users are not necessarily interested in modelling their talk on native speaker norms. There have, as a result, been arguments presented in favour of non-native, lingua franca spoken corpora. Prodromou (1997), arguing from the evidence of his mixed native- and non-native spoken English corpus of some 200,000 words, pointed to the potentially undermining effect of native-speaker English corpora on non-native-speakers faced with the many varieties and cultures of the target language as captured in the extant native-speaker corpora. Reacting to similar concerns, Seidlhofer proposed a spoken corpus of English as a Lingua Franca (ELF) to profile ELF as robust and independent of English as a native language with pedagogical applications (Seidlhofer 2001).

It is worth pointing out that many of the large spoken language corpora are collected not primarily for linguistic research but for speech technology projects. While English data dominates both types of spoken corpora, there is a growing number of non-English corpora. For example, Portuguese: *Português Falado – Documentos Autênticos: Gravações áudio com transcrição alinhada* (Bacelar do Nascimento 2001), which includes

Portuguese varieties spoken in Portugal, Brazil, Goa and African countries; Italian: *Banca dati dell'italiano parlato*, which hosts the 490,000 word LIP corpus (Pusch 2002; Voghera 1996; Cresti 2000); Basque: *Basque Spoken Corpus*, a collection of forty two narratives (Aske 1997); Spanish: The *Corpus Oral de Referencia del Español Contemporáneo* (Ballester/Santamaria/Marcos-Marin 1993), over one million words of spoken Spanish and *Corpus de Referencia del Español Actual*, a 133-million-word corpus, 10% of which comprises spoken data (see <http://corpus.rae.es/creanet.html>); Czech: 800,000 words of spontaneous spoken language (Čermák 1997).

6. Directions in spoken corpus linguistics

At the present time, projects are underway to combine different media in the construction and exploitation of spoken corpora. Cauldwell (2002) combines sound files with on-screen textual displays of natural data, while the *Kids' Audio Speech Corpus* at the University of Colorado, Boulder, USA combines audio and video data with the aim of enabling the development of auditory and visual recognition systems (see http://cslr.colorado.edu/beginweb/reading/data_collection.html). The *British Academic Spoken English* (BASE), assembled at the Universities of Warwick and Reading in Great Britain, under the directorship of Nesi and Thompson, is a companion corpus to MICASE (see above) (see Creer/Thompson 2004 for further details and see http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/). The majority of the BASE recordings are on digital video. The corpus team also plans to edit and compress the video recordings, and to link transcripts and video/audio files on CD-ROM. The corpus construction aims to facilitate the analysis of features such as the pace, density and delivery styles of academic lectures and the discourse function of intonation. Alongside these, the Multimedia Adult ESL Learner Corpus (MAELC) at Portland State University, Portland, Oregon, USA is a corpus of 3600 hours of classroom interaction where transcripts, audio files and video clips are available for research into second language acquisition (Reder/Harris/Setzler 2003, see also <http://www.labschool.pdx.edu>). Further developments in voice recognition may lead to effective automatic transcription of spoken data, and shortcomings in automatic tagging and parsing may be expected to be resolved as techniques advance, and as the need for spoken corpora increases with the extension of research and applications in areas such as voice recognition for the control of machine- and computer-processes, and spoken databanks that are accessed automatically in service contexts such as tourism, financial services, telecommunications, and so on.

7. Literature

- Aijmer, K. (1995), Do Women Apologise More Than Men? In: Melchers, G./Warren, B. (eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell, 59–69.
- Aijmer, K. (1996), *Conversational Routines in English*. London: Longman.
- Aijmer, K. (2002), *English Discourse Particles – Evidence from a Corpus*. Amsterdam: John Benjamins.
- Andersen, G. (1997a), ‘They gave us these yeah, and they like wanna see like how we talk and all that’. The Use of *Like* and Other Discourse Markers in London Teenage Speech. In: Kotsinas,

- U.-B./Stenström, A.-B./Karlsson, A.-M. (eds.), *Ungdomsspråk i Norden*. (MINS 43.) Stockholm: MINS, 82–95.
- Andersen, G. (1997b), ‘They like wanna see like how we talk and all that’. The Use of *Like* as a Discourse Marker in London Teenage Speech. In: Ljung, M. (ed.), *Corpus-based Studies in English*. Amsterdam: Rodopi, 37–48.
- Aske, J. (1997), Basque Word Order and Disorder: Principles, Variation, and Prospects. PhD dissertation, Department of Linguistics, University of California, Berkeley.
- Aston, G. (2001), Text Categories and Corpus Users: A Response to David Dee. In: *Language Learning and Technology* 5(3), 37–72. Available at: <http://llt.msu.edu/vol5num3/pdf/aston.pdf>.
- Atkins, S./Clear J./Ostler N. (1992), Corpus Design Criteria. In: *Literary and Linguistic Computing* 7(1), 1–16.
- Bacelar do Nascimento, F. (2001), *Português Falado, Documentos Autênticos, Gravações audio com transcrições alinhadas*. CD-ROM. Lisboa: Centro de Linguística da Universidade de Lisboa e Instituto Camões.
- Ballester, A./Santamaria, C./Marcos-Marin, F. A. (1993), Transcription Conventions Used for the Corpus of Spoken Contemporary. In: *Spanish Literary and Linguistic Computing* 8(4), 283–292.
- Banjo, A. (1996), The Sociolinguistics of English in Nigeria and the ICE Project. In: Greenbaum S. (ed.), *Comparing English World-wide: The International Corpus of English*. Oxford: Oxford University Press, 239–248.
- Barlow, M. (1998), *Corpus of Spoken Professional American English*. CD-ROM. Houston, TX: Athelstan.
- Beier E./Starkweather J./Miller D. (1967), Analysis of Word Frequencies in Spoken Language of Children. In: *Language and Speech* 10, 217–227.
- Biber, D./Conrad S./Reppen R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D./Johansson S./Leech, G./Conrad S./Finegan E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Blanche-Benveniste, C. (1982), Examen de la notion de subordination. In: *Recherche sur le Français Parlé* 4, 71–115.
- Blanche-Benveniste, C. (1995), De la rareté de certains phénomènes syntaxiques en français parlé. In: *French Language Studies* 5(1), 17–29.
- Bolton, K./Gisborne, N./Hung, J./Nelson, G. (2003), *The International Corpus of English Project in Hong Kong*. Amsterdam: John Benjamins.
- Boucher, V. J. (2005), On the Measurable Linguistic Correlates of Deceit in Recounting Passed Events. Paper presented to: *International Association of Forensic Linguists 7th Biennial Conference on Forensic Linguistics/Language and Law*. Cardiff University, Cardiff, UK, 1st–4th July 2005.
- Boxer, D./Pickering L. (1995), Problems in the Presentation of Speech Acts in ELT Materials: The Case of Complaints. In: *English Language Teaching Journal* 49, 99–158.
- Breivik, L. E./Hasselgren A. (eds.) (2002), *From the Colt's Mouth ... and Others': Language Corpora Studies in Honour of Anna-Brita Stenström*. Amsterdam: Rodopi.
- Bucholtz, M. (2000), The Politics of Transcription. In: *Journal of Pragmatics* 32, 1439–1465.
- Cameron, L./Deignan, A. (2003), Combining Large and Small Corpora to Investigate Tuning Devices around Metaphor in Spoken Language. In: *Metaphor and Symbol* 18(3), 149–160.
- Carter, R. A. (1998), Orders of Reality: CANCODE, Communication and Culture. In: *English Language Teaching Journal* 52, 43–56.
- Carter, R. A./McCarthy, M. J. (1995), Grammar and the Spoken Language. In: *Applied Linguistics* 16(2), 141–158.
- Carter, R. A./McCarthy, M. J. (2004), Talking, Creating: Interactional Language, Creativity and Context. In: *Applied Linguistics* 25(1), 62–88.
- Carter, R. A./McCarthy, M. J. (2006), *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

- Carterette, E./Jones M. H. (1974), *Informal Speech*. Berkeley and Los Angeles: University of California Press.
- Caudwell, R. (2002), *Streaming Speech: Listening and Pronunciation for Advanced Learners of English*. CD-ROM. Birmingham: Speechinaction.
- Čermák, F. (1997), Czech National Corpus: A Case in Many Contexts. In: *International Journal of Corpus Linguistics* 2 (2), 181–197.
- Chafe W./Du Bois J./Thompson S. (1991), Towards a New Corpus of Spoken American English. In: Aijmer K./Altenberg, B. (eds.), *English Corpus Linguistics*. London: Longman, 64–82.
- Chafe, W. (1982), Integration and Involvement in Speaking, Writing, and Oral Literature. In: Tannen D. (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex Publishing Corporation, 35–53.
- Cheng, W./Warren, M. (1999), Facilitating a Description of Intercultural Conversations: The Hong Kong Corpus of Conversational English. In: *ICAME Journal* 23, 5–20.
- Cheng, W./Warren, M. (2000), The Hong Kong Corpus of Spoken English: Language Learning through Language Description. In: Burnard, L./McEnery, T. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang, 133–144.
- Cheng, W./Warren, M. (2002), // ↘ beef ball // → you like //: The Intonation of Declarative-mood Questions in a Corpus of Hong Kong English. In: *Teanga* 21, 151–165.
- Clift, R. (1999), Irony in Conversation. In: *Language in Society* 28, 523–553.
- Cook, G. (1990), Transcribing Infinity: Problems of Context Presentation. In: *Journal of Pragmatics* 14, 1–24.
- Cotterill, J. (ed.) (2002a), *Language in the Legal Process*. Basingstoke: Palgrave.
- Cotterill, J. (2002b), *Language and Power in Court: A Linguistic Analysis of the O. J. Simpson Trial*. Basingstoke: Palgrave.
- Cotterill, J. (2003), *Language and Power in Court*. Basingstoke: Palgrave.
- Cotterill, J. (2004), Collocation, Connotation, and Courtroom Semantics: Lawyers' Control of Witness Testimony through Lexical Negotiation. In: *Applied Linguistics* 25 (4), 513–537.
- Creer, S./Thompson, P. (2004), Processing Spoken Language Data: The BASE Experience. In: *Workshop on Compiling and Processing Spoken Language Corpora, 24th May, LREC 2004*. Lisboa, Portugal, 20–27. Available at: http://www.rdg.ac.uk/AcaDepts/l1/base_corpus/creer_thompson_final.pdf.
- Cresti, E. (2000), *Corpus di italiano parlato*. Firenze: Accademia della Crusca.
- Crowdy, S. (1993), Spoken Corpus Design. In: *Literary and Linguistic Computing* 8(2), 259–265.
- Crowdy, S. (1994), Spoken Corpus Transcription. In: *Literary and Linguistic Computing* 9(1), 25–28.
- Crystal, D. (1995), Refining Stylistic Discourse Categories. In: Melchers, G./Warren, B. (eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell International, 35–46.
- De Cock, S. (1998), A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-native Speakers of English. In: *International Journal of Corpus Linguistics* 3, 59–80.
- De Cock, S. (2000), Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing. In: Mair, C./Hundt, M. (eds.), *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*. Amsterdam: Rodopi, 51–68.
- Duranti, A. (1997), *Linguistic Anthropology*. Cambridge: Cambridge University Press.
- Edwards, J. A. (1991), Transcription of Discourse. In: Bright, W. (ed.), *Oxford International Encyclopedia of Linguistics, Vol. 1*. Oxford: Oxford University Press, 367–371.
- Edwards, J. A./Lampert, M. D. (eds.) (1993), *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, NJ: Lawrence Erlbaum.
- Eggins, S./Slade, D. (1997), *Analyzing Casual Conversation*. London: Cassell.
- Fang, A. C. (1995), Distribution of Infinitives in Contemporary British English: A Study Based on the British ICE Corpus. In: *Literary and Linguistic Computing* 10(4), 247–257.

- Farr, F. (2003), Engaged Listenership in Spoken Academic Discourse. In: *Journal of English for Academic Purposes* 2(1), 67–85.
- Farr, F./Murphy, B./O'Keeffe, A. (2002), The Limerick Corpus of Irish English: Design, Description and Application. In: *Teanga* 21, 5–29.
- Farr, F./O'Keeffe, A. (2002), *Would* as a Hedging Device in an Irish Context: An Intra-varietal Comparison of Institutionalised Spoken Interaction. In: Reppen, R./Fitzmaurice S./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 25–48.
- Fonseca-Greber, B./Waugh, L. R. (2003), On the Radical Difference between the Subject Personal Pronouns in Written and Spoken European French. In: Leistyna, P./Meyer, C. (eds.), *Corpus Analysis. Language Structure and Language Use*. Amsterdam: Rodopi, 225–240.
- Francis, N. (1982), Problems of Assembling and Computerizing Large Corpora. In: Johansson, S. (ed.), *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities, 7–24.
- Gilmore A. (2004), A Comparison of Textbook and Authentic Interactions. In: *English Language Teaching Journal* 58(4), 363–374.
- Granger, S./Hung, J./Petch-Tyson, S. (eds.) (2002), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.
- Greenbaum, S. (1991), ICE: The International Corpus of English. In: *English Today* 28, 3–7.
- Greenbaum, S. (1992), A New Corpus of English: ICE. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm 4–8 August 1991*. Berlin: Mouton de Gruyter, 171–179.
- Gut, U. (2006), Learner Speech Corpora in Language Teaching. In: Braun, S./Kohn, K./Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy*. Frankfurt: Lang, 69–86.
- Halliday, M. A. K. (1989), *Spoken and Written Language*. Oxford: Oxford University Press.
- Handford, M./McCarthy, M. J. (2004), 'Invisible to us': A Preliminary Corpus-based Study of Spoken Business English. In: Connor, U./Upton, T. (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: Benjamins, 167–201.
- Haslerud, V./Stenström, A.-B. (1995), The Bergen Corpus of London Teenager Language (COLT). In: Leech, G./Myers, G./Thomas, J. (eds.), *Spoken English on Computer*. London: Longman, 235–242.
- Hasund, K. (1998), From Woman's Place to Women's Places: Class-determined Variation in the Verbal Disputes of London Teenage Girls. In: Despard, A. (ed.), *A Woman's Place: Women, Domesticity and Private Life*. Kristiansand: Norwegian Academic Press, 187–199.
- Hasund, K./Stenström, A.-B. (1997), Conflict Talk: A Comparison of the Verbal Disputes of Adolescent Females in Two Corpora. In: Ljung, M. (ed.), *Corpus-based Studies in English. Papers from the 17th International Conference on English Language Research on Computerized Corpora. (Language and Computers. Studies in Practical Linguistics 20.)* Amsterdam: Rodopi, 119–133.
- Hepburn, A. (2004), Crying: Notes on Description, Transcription, and Interaction. In: *Research on Language and Social Interaction* 37(3), 251–290.
- Holmes, J. (1996), The New Zealand Spoken Component of ICE: Some Methodological Challenges. In: Greenbaum, S. (ed.), *Comparing English World-wide: The International Corpus of English*. Oxford: Oxford University Press, 163–178.
- Holmes, J. (2001), Ladies and Gentlemen: Corpus Analysis and Linguistic Sexism. In: Mair, C./Hundt, M. (eds.), *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, 141–156.
- Holmes, J./Vine, B./Johnson, G. (1998), *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Howes, D. H. (1966), A Word Count of Spoken English. In: *Journal of Verbal Learning and Verbal Behaviour* 5, 572–606.
- Hughes, R./McCarthy, M. J. (1998), From Sentence to Discourse: Discourse Grammar and English Language Teaching. In: *TESOL Quarterly* 32, 263–287.

- Ide, N./Macleod, C. (2001), The American National Corpus: A Standardized Resource of American English. In: Rayson, P./Wilson, A./McEnery, T./Hardie, A./Khoja, S. (eds.), *Proceedings of Corpus Linguistics 2001*, vol. 13. Lancaster: University of Lancaster, 274–280.
- Ide, N./Reppen, R./Suderman, K. (2002), The American National Corpus: More than the Web Can Provide. In: *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*. Las Palmas, Spain, 839–844. Available at: <http://americannationalcorpus.org/pubs.html>.
- Ihalainen, O. (1991a), A Point of Verb Syntax in South-western British English: An Analysis of a Dialect Continuum. In: Aijmer, K./Altenberg, B. (eds.), *English Corpus Linguistics*. London: Longman, 290–302.
- Ihalainen, O. (1991b), The Grammatical Subject in Educated and Dialectal English: Comparing the London-Lund Corpus and the Helsinki Corpus of Modern English Dialects. In: Johansson, S./Stenström, A.-B. (eds.), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter, 201–214.
- Jefferson, G. (1985), An Exercise in the Transcription and Analysis of Laughter. In: Van Dijk, T. (ed.), *Handbook of Discourse Analysis* (Vol. 3). London: Academic Press, 25–34.
- Kallen, J. L./Kirk, J. M. (2001), Convergence and Divergence in the Verb Phrase in Irish Standard English: A Corpus-based Approach. In: Kirk, J. M./Ó Baoill, D. P. (eds.), *Language Links: The Languages of Scotland and Ireland*. Belfast: Cló Ollscoil na Banríona, 59–79.
- Kirk, J. M. (1992), The Northern Ireland Transcribed Corpus of Speech. In: Leitner, G. (ed.), *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 65–73.
- Kirk, J. M. (1999), The Dialect Vocabulary of Ulster. In: *Cuadernos de Filología Inglesa* 8, 305–334.
- Knowles, G. (1990), The Use of Spoken and Written Corpora in the Teaching of Language and Linguistics. In: *Literary and Linguistic Computing* 5(1), 45–48.
- Lapidus, N./Otheguy, R. (2005), Contact Induced Change? Overt Nonspecific *Ellos* in Spanish in New York. In: Sayahi, L./Westmoreland, M. (eds.), *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics*. Somerville, MA: Cascadilla Proceedings Project, 67–75. Available at: <http://www.lingref.com/cpp/wss/2/paper1141.pdf>.
- Lee, D. (2001), Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. In: *Language Learning & Technology* 5(3), 37–72. Available at: <http://llt.msu.edu/vol5num3/lee/>.
- Leech, G. (1999), The Distribution and Function of Vocatives in American and British English Conversation. In: Hasselgård, H./Oksefjell, S. (eds.), *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi, 107–118.
- Leech, G. (2000), Grammars of Spoken English: New Outcomes of Corpus-oriented Research. In: *Language Learning* 50(4), 675–724.
- McCarthy, M. J. (1998), *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. J. (2001), *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. J. (2003), Talking Back: ‘Small’ Interactional Response Tokens in Everyday Conversation. In: *Research on Language in Social Interaction* 36(1), 33–63.
- McCarthy, M. J./Carter, R. A. (2001a), Size Isn’t Everything: Spoken English, Corpus and the Classroom. In: *TESOL Quarterly* 35(2), 337–340.
- McCarthy, M. J./Carter, R. A. (2001b), Ten Criteria for a Spoken Grammar. In: Hinkel E./Fotos S. (eds.), *New Perspectives on Grammar Teaching in Second Language Classrooms*. Mahwah, NJ: Lawrence Erlbaum Associates, 51–75.
- McCarthy, M. J./Carter, R. A. (2002), *This That and the Other: Multi-word Clusters in Spoken English as Visible Patterns of Interaction*. In: *Teanga* 21, 30–52.
- McCarthy, M. J./Carter, R. A. (2004), ‘There’s Millions of Them’: Hyperbole in Everyday Conversation. In: *Journal of Pragmatics* 36, 149–184.
- McCarthy, M. J./O’Keeffe, A. (2003), ‘What’s in a Name?’ – Vocatives in Casual Conversations and Radio Phone-in Calls. In: Leistyna, P./Meyer, C. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 153–185.

- Mondorf, B. (2002), Gender Differences in English Syntax. In: *Journal of English Linguistics* 30 (2), 158–180.
- Moon, R. (1997), Vocabulary Connections: Multi-word Items in English. In: Schmitt, N./McCarthy, M. J. (eds.), *Second Language Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 40–63.
- Nelson, G. (1996), The Design of the Corpus. In: Greenbaum, S. (ed.), *Comparing English Worldwide: The International Corpus of English*. Oxford: Oxford University Press, 27–35.
- Nelson, G./Wallis, S./Aarts, B. (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
- Nero, S. J. (2000), The Changing Faces of English: A Caribbean Perspective. In: *TESOL Quarterly* 34(3), 483–510.
- O'Keeffe, A. (2004), 'Like the Wise Virgins and all that Jazz': Using a Corpus to Examine Vague Categorisation and Shared Knowledge. In: *Language and Computers* 52(1), 1–20.
- O'Keeffe, A./McCarthy, M. J./Carter, R. (2007), *From Corpus to Classroom*. Cambridge: Cambridge University Press.
- O'Keeffe, A./Farr, F. (2003), Using Language Corpora in Language Teacher Education: Pedagogic, Linguistic and Cultural Insights. In: *TESOL Quarterly* 37(3), 389–418.
- Ochs, E. (1979), Transcription as Theory. In: Ochs, E./Schieffelin, B. B. (eds.), *Developmental Pragmatics*. New York: Academic Press, 43–72.
- Ooi, V. (1997), Analysing the Singapore ICE Corpus for Lexicographic Evidence. In: Ljung, M. (ed.), *Corpus-based Studies in English*. Amsterdam: Rodopi, 245–260.
- Prodromou, L. (1997), Global English and its Struggle against the Octopus. In: *IATEFL Newsletter* 135, 12–14.
- Prodromou, L. (2003), In Search of the Successful User of English. In: *Modern English Teacher* 12, 5–14.
- Pusch, C. D. (2002), A Survey of Spoken Language Corpora in Romance. In: Pusch, C. D./Raible, W. (eds.), *Romanistische Korpuslinguistik*. Tübingen: Gunter Narr Verlag, 245–264.
- Reder, S./Harris, K./Setzler, K. (2003), The Multimedia Adult ESL Learner Corpus. In: *TESOL Quarterly* 37(3), 546–557.
- Rundell, M. (1995a), The BNC: A Spoken Corpus. In: *Modern English Teacher* 4(2), 13–15.
- Rundell, M. (1995b), The Word on the Street. In: *English Today* 11(3), 29–35.
- Schmied, J./Hudson-Ettele, D. (1996), Analysing the Style of East African Newspapers in English. In: *World Englishes* 15(1), 103–113.
- Schmitt, N./Carter, R. A. (2004), Formulaic Sequences in Action: An Introduction. In: Schmitt, N. (ed.), *Formulaic Sequences*. Amsterdam: John Benjamins, 1–22.
- Schonell, F./Meddleton, I./Shaw, B./Routh, M./Popham, D./Gill, G./Mackrell, G./ Stephens, C. (1956), *A Study of the Oral Vocabulary of Adults*. Brisbane and London: University of Queensland Press/University of London Press.
- Seidlhofer, B. (2001), Closing a Conceptual Gap: The Case for a Description of English as a Lingua Franca. In: *International Journal of Applied Linguistics* 11, 133–158.
- Shuy, R. (1998), *The Language of Confession, Interrogation and Deception*. London: Sage.
- Simpson, R. C./Lucka, B./Ovens, J. (2000), Methodological Challenges of Planning a Spoken Corpus with Pedagogical Outcomes. In: Burnard, L./McEnery, T. (eds.), *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora (TALC)*. Frankfurt: Peter Lang, 43–49.
- Sinclair, J. (1995), Corpus Typology – a Framework for Classification. In: Melchers, G./Warren, B. (eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell International, 17–34.
- Spöttl, C./McCarthy, M. J. (2004), Comparing Knowledge of Formulaic Sequences across L1, L2, L3, and L4. In: Schmitt, N. (ed.), *Formulaic Sequences*. Amsterdam: John Benjamins, 191–225.
- Stenström, A.-B. (1995), Taboos in Teenage Talk. In: Melchers, G./Warren, B. (eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell International, 71–80.

- Stenström, A.-B. (1997a), Tags in Teenage Talk. In: Fries, U./Müller, V./Schneider, P. (eds.), *From Ælfric to the New York Times. Studies in English Corpus Linguistics*. Amsterdam: Rodopi, 139–148.
- Stenström, A.-B. (1997b), ‘Can I have a Chips Please? – Just Tell me what one you Want’ Nonstandard Grammatical Features in London Teenage Talk. In: Aarts, J./de Mönninck, I./Wekker, H. (eds.), *Studies in English Language and Teaching*. Amsterdam: Rodopi, 141–152.
- Stenström, A.-B. (1998), From Sentence to Discourse: *cos(because)* in Teenage Talk. In: Jucker, A./Ziv, Y. (eds.), *Discourse Markers: Descriptions and Theory*. Amsterdam: John Benjamins, 127–146.
- Stenström, A.-B./Andersen, G./Hasund, I. K. (2002), *Trends in Teenage Talk*. Amsterdam: John Benjamins.
- Stern, K. (1997), The Longman Spoken American Corpus: Providing an In-depth Analysis of Every-day English. In: *Longman Language Review* 3, 14–17.
- Svartvik, J. (ed.) (1990), *The London-Lund Corpus of Spoken English: Description and Research*. (Lund Studies in English 82.) Lund: Lund University Press.
- Svartvik, J./Quirk, R. (1980), *A Corpus of English Conversation*. Lund: Gleerup.
- Tao, H. (2003), Turn Initiators in Spoken English: A Corpus-based Approach to Interaction and Grammar. In: Leistyna P./Meyer, C. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 187–207.
- Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Voghera, M. (1996), Corpora dell’italiano. In: *Revue française de linguistique appliquée* 1, 131–134.

Michael McCarthy, Nottingham (UK) and Anne O’Keeffe, Limerick (Ireland)

48. Cross-lingual influence: The integration of foreign items

1. Introduction
2. Background
3. The influence of English
4. English influence on Swedish word formation
5. Discussion
6. Related research
7. Conclusions
8. Literature

1. Introduction

Contact between cultures, and thereby between languages, is clearly as old a phenomenon as language itself, and it is undoubtedly also one of the strongest driving forces behind linguistic change and evolution. This has led to bi- or even polylingualism constituting the normal linguistic situation in large parts of the world (Ladefoged/Maddieson 1996). In addition to this, the current rôle occupied by the English language since more

- Stenström, A.-B. (1997a), Tags in Teenage Talk. In: Fries, U./Müller, V./Schneider, P. (eds.), *From Ælfric to the New York Times. Studies in English Corpus Linguistics*. Amsterdam: Rodopi, 139–148.
- Stenström, A.-B. (1997b), ‘Can I have a Chips Please? – Just Tell me what one you Want’ Nonstandard Grammatical Features in London Teenage Talk. In: Aarts, J./de Mönninck, I./Wekker, H. (eds.), *Studies in English Language and Teaching*. Amsterdam: Rodopi, 141–152.
- Stenström, A.-B. (1998), From Sentence to Discourse: *cos(because)* in Teenage Talk. In: Jucker, A./Ziv, Y. (eds.), *Discourse Markers: Descriptions and Theory*. Amsterdam: John Benjamins, 127–146.
- Stenström, A.-B./Andersen, G./Hasund, I. K. (2002), *Trends in Teenage Talk*. Amsterdam: John Benjamins.
- Stern, K. (1997), The Longman Spoken American Corpus: Providing an In-depth Analysis of Every-day English. In: *Longman Language Review* 3, 14–17.
- Svartvik, J. (ed.) (1990), *The London-Lund Corpus of Spoken English: Description and Research*. (Lund Studies in English 82.) Lund: Lund University Press.
- Svartvik, J./Quirk, R. (1980), *A Corpus of English Conversation*. Lund: Gleerup.
- Tao, H. (2003), Turn Initiators in Spoken English: A Corpus-based Approach to Interaction and Grammar. In: Leistyna P./Meyer, C. (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, 187–207.
- Tognini Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Voghera, M. (1996), Corpora dell’italiano. In: *Revue française de linguistique appliquée* 1, 131–134.

Michael McCarthy, Nottingham (UK) and Anne O’Keeffe, Limerick (Ireland)

48. Cross-lingual influence: The integration of foreign items

1. Introduction
2. Background
3. The influence of English
4. English influence on Swedish word formation
5. Discussion
6. Related research
7. Conclusions
8. Literature

1. Introduction

Contact between cultures, and thereby between languages, is clearly as old a phenomenon as language itself, and it is undoubtedly also one of the strongest driving forces behind linguistic change and evolution. This has led to bi- or even polylingualism constituting the normal linguistic situation in large parts of the world (Ladefoged/Maddieson 1996). In addition to this, the current rôle occupied by the English language since more

than half a century ago as the lingua franca of international communication and the influential status of English-speaking cultures within entertainment, scientific and technical domains during the past several decades, raises the more specific question regarding which linguistic effects this influence has had on native speakers of other languages. In this article, we will discuss some of the issues involved in corpus-based studies of the integration of such foreign items in a language on the phonetic, phonological and morpho-syntactic levels. We will use our own studies of English influence on Swedish as an example and make reference to research on other languages and on the influence of foreign items on specific applications.

The rest of this article is structured as follows. We begin, in section 2, by discussing preliminaries regarding spoken vs. written language, frequency properties of vocabulary growth, and the influence of English on other languages. Section 3 reports on a study of the segmental aspects of English influence on Swedish, and section 4 investigates word formation aspects of such influence, using data from several spoken and written language corpora. In section 5, the results regarding both segmental and word formation aspects are summarized and discussed. In section 6, an overview of related research regarding other languages and applications is presented. The article is concluded, in section 7, with a brief outlook on the future.

2. Background

2.1. Spoken vs. written language

As discussed in articles 11 and 30, spoken language differs fundamentally from written language in several obvious ways: spoken language constitutes a primary means of communication, normally acquired in early childhood, whereas the skills of reading and writing need to be actively learnt at a later stage. Speech is also evasive by its very nature, since it is being conveyed by an ephemeral acoustic signal, which cannot be erased or edited by the speaker once it has been produced. Partly due to this fact, spoken and written language differ in many linguistic respects, e. g. syntactically, regarding structure, length and frequency of phrases, in terms of vocabulary, etc. Planning of spoken utterances also needs to be done on-line, often while speaking, which is one reason (among several) for the hesitations, restarts, repairs and other disfluencies which are typical of spontaneous speech (Shriberg 1994; Eklund 2004). Furthermore, spoken language is normally situated in a real-world setting – or in the case of telecommunications, in two or more – in which references to real-world objects can be made, and as a result the types and frequency of deictic expressions are fundamentally different from those of written language (Jungbluth 1999). The normal style of speech is interactive and conversational (except when reading aloud, reciting monologues etc.), while producers and consumers of written language are separated across both time and space but are in a sense always off-line. Because of these fundamental differences between the spoken and the written modalities of language, it can be expected that the type of cross-language influence discussed here should show more extensive, qualitatively different, and presumably earlier effects on spoken than on written language.

To further complicate matters, it can be questioned to what extent the human processing of written vs. spoken language are in fact distinct processes. This is one of the issues discussed and investigated in a series of cross-modal experiments, undertaken by Fitt (1998), who says that "... there is some evidence to suggest that the perception of a written word is not divorced from activation of the corresponding phonological form, that is to say, it is not perceived simply as a string of graphemes, but as a string (or word, if it is a word) with a corresponding pronunciation." (Fitt 1998, 21–22).

Fitt also quotes research which indicates that human subjects in fact cannot refrain from recoding written input into a phonological form, thus more or less compulsorily assigning a pronunciation to written words, even when the task does not require it. As regards what considerations underlie human subjects' classification of written input of possibly foreign origin, Fitt states that "any features of written words which are either not found in English, or are uncommon, or are limited to loanwords, can lead a reader to suppose that the word is non-native, even when the reader does not know enough to assign the word to a particular foreign language" (Fitt 1998, 30).

2.2. Vocabulary size

From a purely linguistic point of view, rare phenomena can be at least as interesting and relevant to study as those which are very common. From a technical and computational point of view, however, it is often assumed to be opportunistic to deal with high-frequency phenomena first, since these would constitute the bulk of observed linguistic patterns, or to attempt to draw the borderline between productive and non-productive patterns, which is not at all straightforward, as shown in article 41. With this in mind, we will now discuss some aspects of vocabulary growth and size, with the purpose of identifying whether foreign language influence on spoken language is "large" or not.

The issue of vocabulary size and how to estimate it is a question of long standing within computational linguistics, with one approach drawing on resemblance to frequency distributions of other naturally occurring phenomena (Good 1953). The differences in frequency distribution between spoken and written Swedish are examined by Allwood (1998, 19) who notes that "the most common words are more common in speech than in writing". He goes on to say that this is not as helpful as it may sound, since these most common words are not the ones carrying the bulk of the information, quite the opposite. Kornai (2002) argues against the "closed vocabulary assumption", which claims that, at least synchronically, there is only a fixed number of words in a given language. Kornai shows that it suffices to assume that new words are introduced only by a constant rate for vocabularies in general to be "open", e. g. infinitely large (see also article 41). Kornai discusses a model of word frequency, based on the assumption that words (tokens) are drawn at random (with replacement) from an urn, large enough to hold all word types in fixed proportions, even if he admits that such a model may not be psychologically realistic. He then shows how such a model actually needs to be modified to include two urns, one for high-frequency function words, and the other for low-frequency content words, because of the different behaviour exhibited by the two categories. In fact, as shown already by Zipf (1935), the number of items with exactly one token (hapax legomena), types with exactly two tokens, types with exactly three

tokens, etc. show a linear relation when plotted against $\log f$, where f is the frequency (i.e. 1, 2 and 3, respectively in the previous cases). Kornai (2002, 75) states that “hapaxes are not some accidental contamination of corpora” and notes that most large corpora collected for linguistic purposes today are of the type defined by Baayen (2001) as having a large number of rare events. As Kornai puts it, “the large number of hapaxes is only the tip of the iceberg as far as vocabulary growth is concerned” (Kornai 2002, 76). He goes on to analyze the make-up of this infinitely large section of the “low-frequency” vocabulary and concludes that the main source of this infinite vocabulary growth is productive generative morphology, in particular compounding, and seems to see no argument against a description in which foreign linguistic items form part of those processes.

3. The influence of English

The establishment over the past century of the English language as more or less the global lingua franca is indisputable (Melchers/Shaw 2003), although there are indications that this might change in the future (Graddol 2006). Be that as it may, there is no doubt that the post-war influence of American-English culture (film, music, entertainment and so on), and also the strong association between this and youth culture is most likely one of the most important factors behind the current rôle of English in the world, sometimes referred to as ‘Global English’ (Graddol 2006). Presently, the overwhelmingly strong influence of English as *the* language within science and technology is also a major factor at play here. Other factors include increase in technical mobility and ease of global access e.g. through broadcast media and the Internet, subtitling rather than dubbing foreign language films etc. Improved foreign language education is probably yet another factor that lies behind the influence that English has had on many languages, and it is estimated that by the year 2010 there will be around 2 billion learners of English in the world, although it is argued that the figure will most likely peak there, and then decline (Graddol 2006, 14, 98–101). As an indicator, Graddol points out that while 51.3% of Internet pages were in English in the year 2000, the figure had fallen to 32% in the year 2005. (For an in-depth discussion on the international rôle of English, and other ‘competing’ languages, as well as number of speakers (native, second-language, learners) and so on, see Graddol (2006) and Skutnabb-Kangas (2000)). Skutnabb-Kangas (2000) lists formal education and mass media as “direct main agents” for the declining number of languages in the world (Skutnabb-Kangas 2000, 5.) All of the above can be assumed to influence the usage of certain words and phrases with the sociolinguistic purposes of displaying group identity and establishing common ground. The effects of language contact occur at virtually all linguistic levels, from phonology to pragmatics (Weinreich 1953), but the degree as well as the rate of integration of English elements in different languages seems to vary as a function of exposure, linguistic structure, attitude, foreign language education etc. These issues have consequently received attention within several areas of linguistics during at least the past century (Jespersen 1902), spanning such diverse areas as historical linguistics, second language acquisition, generative phonology and Optimality Theory (Prince/Smolensky 1995).

The lexicographically oriented project “The English Element in the European Languages” (Görlach 2001) charted anglicisms in more than twenty different European lan-

guages and, as part of the project, published a number of dictionaries, including one for Swedish (Antunović 1999). Filipović (1996) underlines that the rôle of English as a word donor to other languages has literally exploded from the middle of the 20th century and onwards. He gives examples of how these imported elements tend to affect the linguistic system of the borrowing languages, e. g. by expanding the repertoire of allowable final consonant clusters in Croatian, and by extending the Russian use of non-palatalized consonants to certain phonological positions, where normally the corresponding palatalized consonant would be obligatory. He also notes how vowel reduction in unstressed syllables, which is another highly characteristic property of Russian, is inhibited in anglicisms.

Several phonological theories have sought to model and explain the patterns found in this type of language contact. For instance, earlier ideas regarding the “markedness” of universally infrequent speech sounds and their phonotactic combinations were taken further within Natural Phonology (Donegan/Stampe 1979), which claims that certain processes are inherently more natural than others, which would explain e. g. why devoicing of final obstruents occurs in loanwords even in some languages lacking such segments in that position. In a similar vein, some of the allegedly universal constraints proposed within Optimality Theory also originate from typological studies of interference and assimilation phenomena due to language contact. Constraints requiring fidelity towards the underlying form are generally in conflict with constraints related to criteria for well-formedness of the surface forms of the language, and especially so in the case where the underlying form originates in a foreign, donor, language. This may explain differences between languages in dealing with foreign loans, but also the apparent stratification of the vocabulary of certain (or even most) languages into “native vocabulary”, “assimilated loans” and “foreign vocabulary”, as claimed by Itô/Mester (1999). This reasoning also illustrates how it is in fact quite difficult to define the notions of “native” vs. “foreign” and also how that distinction, should it be possible to define, is more or less bound to change over time. Among the first to present a more comprehensive theory of language contact phenomena, that also included the effects of exposure, were Thomason/Kaufman (1988).

Crosslingual issues have recently received increasing attention within several research areas related to spoken language dialogue applications. As Billi (n.d.) points out, the development and deployment of some of the commercially most interesting services, such as automatic train time table information (Billi/Lamel 1997), stock market information, directory services (e. g. Carlson/Granström/Lindström 1989, Spiegel/Macchi/Gollhardt 1989), and call routing (Gorin/Riccardi/Wright 1997), all evoke various multi- and cross-lingual issues, e. g. handling of non-native speech, native speakers’ handling of nonnative items and names, etc. From a purely technical point of view, foreign features at different linguistic levels would not constitute much of a problem even if they were frequent, as long as they were fossilized and could be listed as exceptions in one way or the other. However, it is quite clear that foreign traits are often assimilated or integrated into the receiving language in such a way that they also attain generative properties, e. g. in word-formation, adding considerably to the complexity of analyzing new words in both spoken and written language, as noted for instance by Lüdeling/Schmid (2001).

Finally, it should also be pointed out that the current development, with the domination of English on the international arena, is not necessarily something that native speakers of English are happy about. As Graddol points out in the introduction to his “English Next”:

“Anyone who believes that native speakers of English remain in control of these developments will be very troubled /.../ it is native speakers who, perhaps, should be the most concerned.” (Graddol 2006, 12)

Graddol’s main point is, of course, that as English is incorporated into other languages it is also transformed, as we will look at in more detail in the remainder of this article.

3.1. English influence on Swedish segmentals

During 1981–1985, surveys of some 2,000 informants’ attitudes towards English loans and preferences regarding wording and grammatical constructions were made in the project “Engelska i svenska” (Ljung 1986). Recordings were made of a smaller number of subjects, reading sentences including fairly frequent English loans as well as names of (at the time) well-known athletes and politicians. Demographic data were also collected, and the recordings were labelled as either adhering to the “native” (English) pronunciation, or to some sort of Swedish approximation. Results showed that sociolinguistic factors, e. g. educational level, affected both attitudes and performance, but that even well-educated subjects made use of conventionalized adaptations to Swedish, e. g. substituting [s] for [z] when the English spelling used <rs> for the latter. In general, the younger and well-educated subjects had a more positive attitude towards anglicisms, and also used them to a higher degree. Frequency of anglicisms in newspapers was also studied, and was found to be in the order of 0.3 %.

In a more recent study by Sharp (2001), which also includes a good overview of the literature on English influence on Swedish, “code-switches” of English origin in two corpora of spoken Swedish were compared. One corpus was based on recordings made at business meetings in an international shipping company, while the other consisted of casual conversation of young adults drawn from a televised reality show. Sharp found differences between the two corpora in a number of respects, including frequency of occurrence, prosodic signalling and degree of integration or accommodation.

3.2. The *Xenophones* production study

Eklund/Lindström (1996) reported from a production study which was based on recordings made in 1995–1996 of nearly 500 Swedish subjects between the ages of 15–75, who each provided approximately one hour of computer-prompted speech, where sentences were read from a monitor (Eklund/Lindström 2001). The primary purpose of the recordings was to collect training material to build a Swedish speech recognizer, something all subjects were informed about. Included in the material were a dozen sentences with well-known foreign names and words, most of which were English, an example of which is given in (1).

- (1) *Veckopressens favoriter är verkligen Diana and Charles*
Diana and Charles are indeed the favourites of the tabloids

This set of twelve sentences took in the order of a minute for each subject to read. It can be assumed that the subjects were not aware of the specific object of this study,

namely to investigate the pronunciation of foreign sounds, since this was never mentioned to the subjects, who were instead given the (truthful) explanation that the recordings were made for speech recognition purposes, and that, above all, they should provide as “normal” a rendering of the material as possible, and avoid hyperarticulation or similar.

The recordings were transcribed by phonetically trained native speakers of Swedish, all with a high degree of proficiency in English. A common subset of the transcriptions was later cross-checked for inter-transcriber consistency. For each subject, every allophonic transcription in 33 target positions – like the ones indicated by curly brackets in Example (2) – within the 12 sentences were then assigned to one of three different categories along an axis, ranging from near-source-language (CATEGORY I) via partly accommodated (but clearly not “Swedish”) (CATEGORY II) to rephonematized (CATEGORY III). Through this process, approximately 23,750 manually transcribed and classified tokens were collected.

(2) *Veckopressens favoriter är verkligen D{i}{a}na and {Ch}arle{s}*

Detailed results have been provided in several publications (Eklund/Lindström 1996, 1998, 2001; Lindström/Eklund 1999, 2000, 2002; Lindström 2004a, 2004b), so we will only summarise some of the main findings here:

Nearly all subjects either used or made an attempt to use “foreign” speech sounds, termed *xenophones* (Eklund/Lindström 1998). For the present study, 19 target instances (of the 33 available) were selected, the selection criteria being of practical rather than linguistic nature. Just as in Ljung’s (1986) study, the frequency of occurrence differed considerably across different lexical items and different positions within words or phrases, even regarding the “same” sound, as can be seen in Figure 48.1.

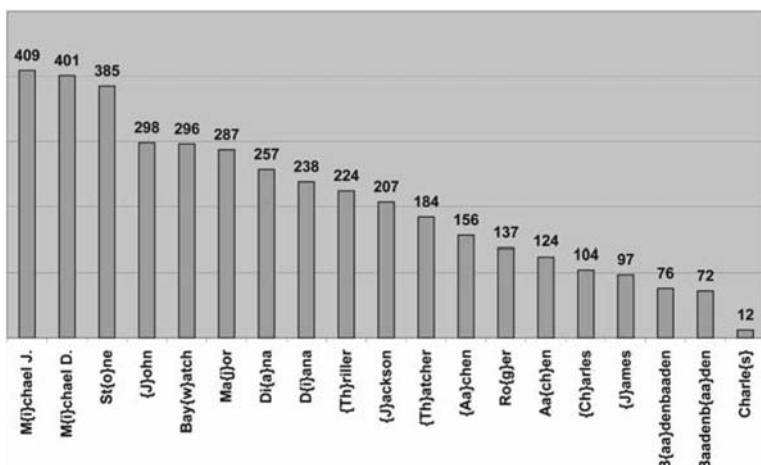


Fig. 48.1: Number of subjects (out of 422) with non-rephonematized productions (corresponding to CATEGORIES I and II in section 3) for 15 potential xenophone instances, positionally indicated by curly brackets in the orthography

Eklund/Lindström found that almost all vowel segments featured pronunciations that were close to the source language, notably also including the diphthong [øʊ], which does not resemble any native speech sound in Swedish. The consonant segments [w], [z] and [ʒ] were produced by most subjects as [v], [s] and [ʃ], respectively. Of these, [w] occurred word-initially, while the fricatives occurred in medial position (with one exception, occurring word-finally). On the other hand, virtually all subjects produced [tʃ], and quite a few also [dʒ], in a fashion very close to the source language. A more scattered distribution along the near-native-to-rephonematized axis was displayed by e.g. [ð] and [θ], which were often replaced by the corresponding stops. These results largely confirm those of Ljung (1986), even if no direct comparison of the production data was possible, since Ljung's subjects were probably aware of the object of study, and the labelling conventions and instructions also differed between the two studies.

As regards explanatory underlying factors, education and age showed significant effects (Pearson chi-square, two-tailed), in the sense that higher education yielded a larger share of near-English pronunciations. That share was also significantly higher for subjects between 25 and 45 years of age. Gender and dialectal background, however, did not turn out to be quite as easy to identify as explanatory factors where there seemed to be considerable variability, seemingly due to a host of other factors than the demographic variables recorded. Due to the nature of some of these sources of variability, such as diachronic change, they would be difficult – if at all possible – to remove by clever design of the data collection procedure. An alternative analysis procedure was therefore undertaken with the purpose of reducing the scattering of the data, rather than multiplying the sources of variability with each other. In order to artificially construct an abscissa along which an ordinate of conflated frequencies could be plotted, all data from the 19 xenophone instances were subjected to the following analysis: the individual data from each subject i who had a complete set of data points for the 19 xenophone instances were selected in order to make comparison possible. The number of such subjects was 422. The next step was to pool the data across all 19 instances for each subject, by calculating the number N_i of non-rephonematized productions, corresponding to CATEGORIES I and II in section 3, out of these 19 potential instances for each subject. Next, the whole set of individual data points, still associated with their demographic variables, was sorted according to N_i along this newly constructed abscissa. Orthogonal to this, the sum S of non-rephonematized productions for each xenophone instance was also calculated, making it possible to rank the instances along that dimension. By this non-destructive procedure two new axes were obtained, hopefully more suitable for analysis of the demographic variables and practically without data loss. A scatter plot of the entire data set, reorganized according to the procedure described above, is shown in Figure 48.2 (oriented so as to match the layout in Figure 48.3 and by rotation also Figures 48.1 and 48.4).

As is apparent in Figure 48.2, most of the non-rephonematized productions are shifted towards one corner of the plot. In this way it was possible to study to what extent this shifting happened because of the tilting just described – and also whether it can be shown to depend on the demographic parameters – since every data point is associated with the demographic parameter of the individual subjects.

As was mentioned earlier, education and age showed significant effects. This effect can now be illustrated on a semi-continuous scale across the whole data set, rather than only per xenophone instance. Such a visualization is shown in Figure 48.3, which plots

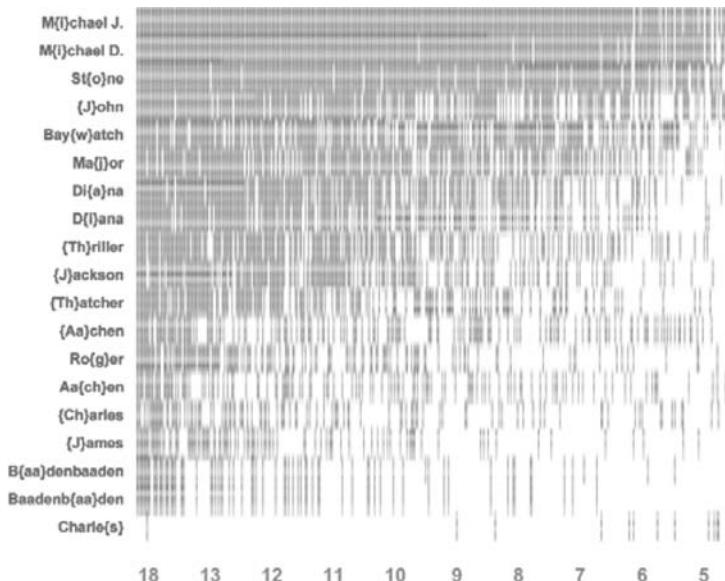


Fig. 48.2: An overview of the entire data set representing the 19 xenophone instances for 422 subjects, using the same abscissa as that in Figure 48.3. The ordinate used in this figure is the same as the abscissa in Figure 48.1 and Figure 48.4. Each data point represents a non-rephonematised production corresponding to CATEGORIES I or II in section 3.1.

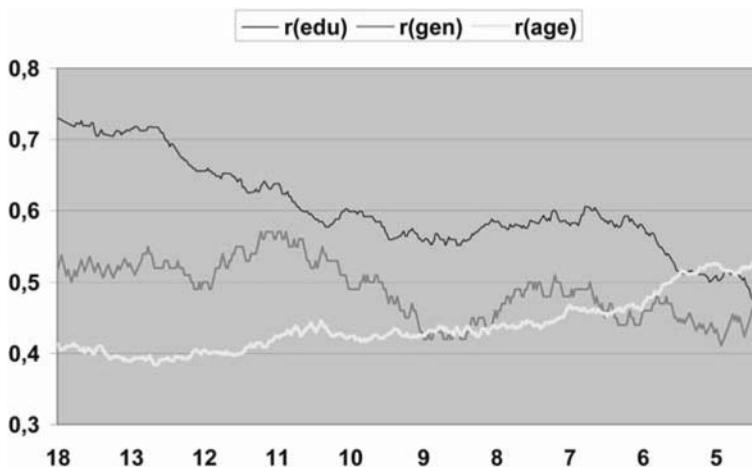


Fig. 48.3: Education, gender and age ratios, normalized to the interval [0,1] and plotted as a smoothed average (over ± 50 data points) for 422 subjects against each subject's number of non-rephonematised productions (corresponding to CATEGORIES I and II in section 3.1.)

the dimensionless ratios r_{edu} , r_{gen} and $r_{\text{age}} \in [0,1]$ for education, gender and age, respectively.

What seems to be a slight tendency towards higher degree of xenophone production for female subjects can also be observed. The effects of educational level are shown in

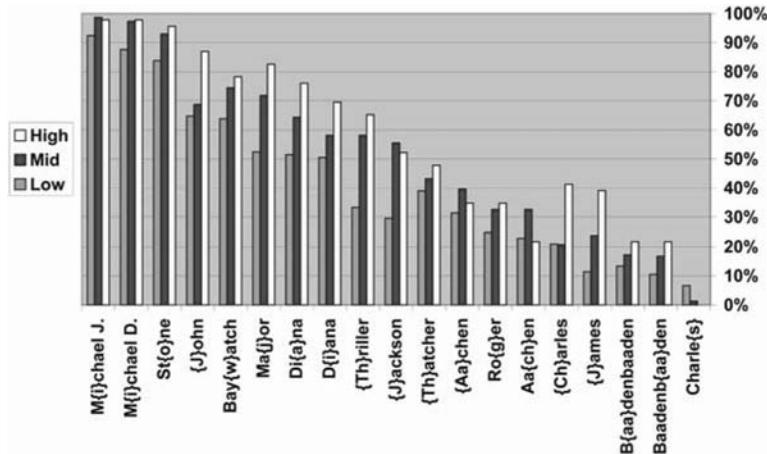


Fig. 48.4: Percentage of subjects per educational level with non-rephonematized productions (corresponding to CATEGORIES I and II in section 3.1.) plotted for the xenophone instances in Figure 48.1. The educational level is coded as *Low* (≤ 9 years of school), *Mid* (10–13 years of school) or *High* (university education). Bars pertaining to *Aachen* and *Charles* are not significant

more detail in Figure 48.4, where the subjects have been divided into three levels of education and results are shown for 19 selected xenophone instances.

In order to check that any differences discerned in Figure 48.4 were larger than the contribution of random or systematic sources of error, the same data were plotted against the quarter of year of birth of each subject. This check indicated that the results for *Aachen* and *James* should be disregarded, while the others would remain valid. In all instances but two, subjects with *Low* education (up to 9 years of school) produced the smallest share of CATEGORIES I and II productions. It also seems to be the case that the differences between educational groups are very small for segments which were produced using xenophones (or good approximations) by a very large share of the subjects, as seen e. g. in the first vowel in *Michael* and in the vowel in *Stone*. This difference also seems to get accentuated as the overall share of xenophone productions was negatively correlated as a function of N .

4. English influence on Swedish word formation

As pointed out by Schmid et al. (2001), it is the productive nature of non-native elements which calls for modelling, both from a linguistic and a computational point of view. In order to investigate such properties, extensive (transcribed) corpora of spoken Swedish are called for, but unfortunately such resources do not exist in abundance. However, earlier results reporting on the influence of English even on written Swedish would suggest that such investigations regarding spoken Swedish would in all likelihood be fruitful (Söderberg 1983). Söderberg (1983) also points out the relative influence of different languages (e. g. Dutch, German and French) on Swedish at different points in time, thus highlighting the rôle of cultural influx. Lindström (2004a, 2004b) analysed several cor-

pora of both spoken and written Swedish of relevance to the problem of word formation. Among those were the GSLC (Allwood et al. 2002), the GSM (Andersson et al. 1999; Norrby/Wirdeñäs 1998), and material from the Talbanken corpus (Einarsson 1976). The morphosyntactically annotated Parole corpus (Parole Consortium (n.d.); Gronostaj (n.d.)) consisting of around 19 million running words from different genres of written Swedish, was also included for reference, and the frequency distribution of this corpus compared to those of Parole and Talbanken is shown in Figure 48.5.

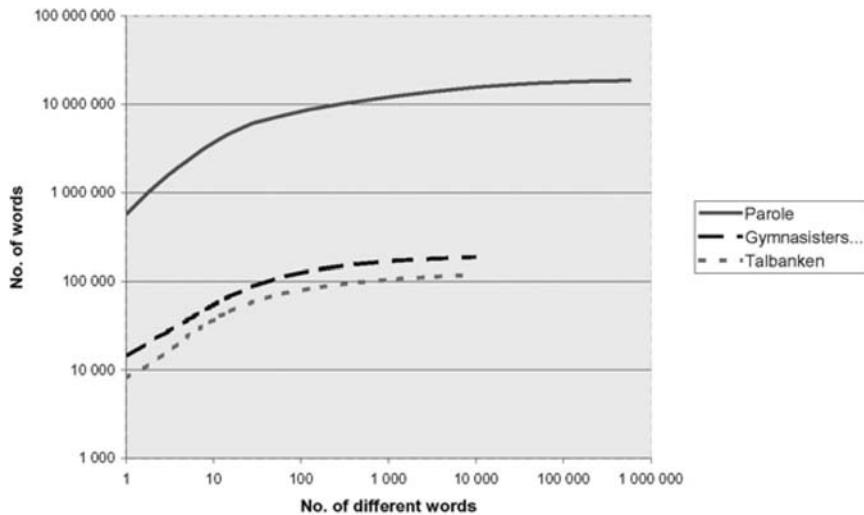


Fig. 48.5: Cumulative frequency distribution for the text corpus Parole and the spoken language corpora Gymnasisters språk- och musikvärldar (Andersson et al. 1999) and Talbanken (Einarsson 1976)

An excerpt across all genres of GSLC consisting of the 24,544 most frequent types was scanned for foreign items, and after manual checking of a conservative candidate list, 125 types remained, including many creative examples of crosslingual compounding, e.g. *brownspråk* ('brown language'), *skoboard* ('shoe board'), but also of more or less straightforward integration of English roots with the Swedish inflectional system, e.g. *approachen* ('approach' + definiteness suffix). The examples, drawn from conversational speech, also included many interjections of English origin.

Within the project GSM (Andersson et al. 1999; Norrby/Wirdeñäs 1998), 27 group conversations, amounting to approximately 20 hours of speech have been collected and orthographically transcribed. The cumulative frequency distribution of this corpus is shown in Figure 48.1, and some examples drawn from that corpus are shown in Table 48.1, where each word is shown along with its rank, frequency, relative frequency and cumulative coverage in the actual corpus. One thing to note is the high productivity featured by *skate[board]*. Another observation, which relates directly to the phonological level, discussed in section 3.2., is that several of the example words are highly likely to elicit pronunciations using xenophone extensions, e.g. by [dʒ] in *energy* and [əʊ] in *snow-*.

Within the Talbanken project (Einarsson 1976), more than 115,000 words of interviews, conversation and debate were recorded and transcribed, as described by Tele-

Tab. 48.1: Example words, along with rank, frequency, relative frequency and cumulative coverage in the corpus Gymnasisters språk- och musikvärldar (Andersson et al. 1999)

Word	Rank	<i>f</i>	Rel. <i>f</i> [%]	Cumul. <i>f</i> [%]
de	1	14148	7.48	7.5
e	2	7743	4.09	11.6
ja	3	5473	2.89	14.5
jag	4	4748	2.51	17.0
så	5	4470	2.36	19.3
på	6	4244	2.24	21.6
inte	7	3819	2.02	23.6
:	:	:	:	:
energy	447	35	0.02	83.1
skateare	682	20	0.01	86.4
coolt	724	19	0.01	86.8
cool	866	15	0.01	88.0
coola	1016	12	0.01	89.1
skate	1098	10	0.01	89.5
coolio	1677	6	< 0.01	91.8
coolhetsstatus	4388	2	< 0.01	96.0
speedar	6360	1	< 0.01	97.2
speeda	6361	1	< 0.01	97.2
snowboardtävlingar	6428	1	< 0.01	97.2
snowboardare	6429	1	< 0.01	97.2
skatesvängen	6659	1	< 0.01	97.4
skatepunkarna	6660	1	< 0.01	97.4
skatepunk	6661	1	< 0.01	97.4
skatekulturen	6663	1	< 0.01	97.4
skateboardmen	6664	1	< 0.01	97.4
skateboardkulturen	6665	1	< 0.01	97.4
skateboardare	6666	1	< 0.01	97.4
skatearmusik	6667	1	< 0.01	97.4
skatearaktigt	6668	1	< 0.01	97.4
skatearkläder	6669	1	< 0.01	97.4
skateaktigt	6670	1	< 0.01	97.4
:	:	:	:	:
Total	11,635	189,246		100.0

man (1974) and Einarsson (1978). The cumulative frequency distribution of this corpus is also included in Figure 48.5. As expected, the cumulative frequency distributions of the three corpora show that the two spoken language corpora are quite similar, while the much larger text corpus Parole behaves differently. In order to cover 90 % of each of the three corpora, it takes 14 % of the 8,289 different words in Talbanken, 10 % of 11,635 different words in GSM, but only 5 % of the 573,546 different words in Parole. The share of hapax legomena, however, is roughly the same across all three corpora, namely 56 % in Talbanken, 60 % in GSM and 54 % in Parole.

The Talbanken interviews were part of a sociological study regarding attitudes towards labour immigration, which (as expected) elicited topics such as ethnicity, foreign language learning, culture etc. Despite this, when exactly the same directed semi-auto-

matic search procedure that was used with the Parole and GSM corpora was applied to Talbanken, only two examples turned up, one being a book title quotation and the other being the use of the verbal form *toucha* ('touch' + Swedish verbal infinitival ending -*a*).

This almost complete lack of anglicisms in the Talbanken corpus could of course also be due to other causes than language contact, e. g. factors related to the interview situation. To eliminate the risk that any perceived distance between interviewer and interviewee inhibited crosslingual word formation processes, a set of highly informal conversations, recorded in a project by Bengt Nordberg and others in 1967–1968, were also studied (Pettersson/Forsberg 1970). Transcriptions of two hours of conversational-style interviews with five subjects, between the ages of 17–23, plus a conversation between two female subjects, aged 20 and 21, comprising a total of 16,250 running words, were analyzed by one of the present authors. The topics of conversation were school, language education, hobbies, sports, travel, TV shows, etc. Only one single example of English influence was found, namely in the unsupervised conversation between the two female subjects, who quote an English song title ("You are the only one"). These results appear to support the hypothesis that English influence on spoken Swedish is much more widespread today than a couple of decades ago. One must, however, bear in mind that the corpora analyzed here were collected by other researchers for completely different purposes, which may make it difficult to draw such strong conclusions.

5. Discussion

From the corpora studied, we can conclude that word formation processes, like derivation and compounding, are highly productive in contemporary spoken Swedish also when incorporating material (morphemes, simple lexemes or even complex nominals) borrowed from English. In addition, such material also integrates well with several Swedish inflectional paradigms. In the spoken language corpora from the 1960s, hardly any examples of these processes are found, despite the fact that the topics of conversation should, if anything, elicit precisely that. In the productive processes we see today, the foreign material can or has to undergo adaptations in order to fit with morphotactic or morphonological constraints, as e. g. in the case of stress pattern and word accent restrictions in Swedish compounding. Sometimes virtually no adaptation occurs, e. g. when retaining English plural endings, instead of employing one from an appropriate Swedish declination paradigm. We have also revisited existing data from a production study of the segmental aspects, which indicates that there are numerous cases in everyday communicative situations where the phonological system simply needs to be extended with xenophones in order to model, produce or perceive contemporary and socially acceptable spoken Swedish.

The effects of educational level in the production study, coupled with well-known sociolinguistic grounding mechanisms, seem to suggest that selecting the appropriate level of xenophone inclusion should be of importance for the perceived persona e. g. in the generation of synthetic speech. It is also worth noting that the differences between educational groups seem to get accentuated with decreasing overall share of xenophone productions. One way of interpreting this is that in some cases pronunciations involving xenophones of English origin are already conventionalized, and consequently produced

by virtually all subjects. While this may be technically interesting since it will require special treatment e. g. in spoken language dialogue systems, it is probably of less interest from a strictly linguistic perspective. Borrowing the terminology from Optimality Theory, one might simply say that constraints requiring faithfulness towards the underlying forms of English origin completely outrank conflicting well-formedness constraints, requiring re-phonematization. In less conventionalized cases, however, education, age, and possibly also gender, seem to determine to what *degree* faithfulness constraints are allowed to outrank conflicting well-formedness constraints. This fits well with Davidson (2001), who claims that studying the phonotactics of a vocabulary in equilibrium gives little insight regarding the interaction between conflicting constraints, compared to what can be extracted from the type of situation we are dealing with here.

One question often raised in other studies is whether the English influence on (particularly written) Swedish is large or not. This question is often associated with a debate where some regard this type of influence as a problem or even perceive it as a threat to relatively small languages, or specific domains within those languages, e. g. computing, engineering, etc. It should be borne in mind, however, that linguistic borrowing, boosted by cultural contact of various kinds, is (and always has been) one of the most fundamental driving forces in linguistic development, with the English language itself being a very obvious result of such a process. Also, as can be seen from the many examples we have given, although terminology borrowed from English is often allowed to expand the phonotactic repertoire, spoken Swedish is still subject to seemingly stable “native” morphological, morphonological and prosodic constraints. These processes need to be further studied and descriptions of contemporary Swedish need to be revised and extended to take them into account, rather than treating them as a marginal phenomenon. The relative stability of some of these well-formedness constraints does not mean that Swedish speech and language technology applications will face no problems related to English and other foreign linguistic elements, quite the opposite. At first glance, the problem may seem to be marginal and of minor importance – after all, each item in Table 48.1 accounts for a relatively small fraction of the entire corpus. However, one needs to remember that while the most frequent items (many of which belong to the closed word classes) rapidly yield high coverage in terms of cumulative frequency, the productive nature of Swedish morphology in connection with English items, as we have seen, in fact makes that section of the vocabulary infinitely large. As we have seen, the share of hapaxes is 54% in Parole. However, that corresponds to no less than 310,973 items. Even if these unique word forms probably also include a few inevitable typos that have escaped the meticulous annotation process, they are primarily formed through the productive morphological processes, of which we have just seen numerous examples. The two spoken language corpora in Figure 48.5 are not really very different from Parole in terms of growth, they are only a lot smaller, with approximately 5,000 hapaxes each. Chances are that the next time a speech corpus of a similar size is collected, its set of hapaxes will not have much overlap with any of these corpora (Good 1953). It has also proven necessary to develop and evaluate any lexical component against functional criteria, rather than using data-driven methods (Lindström 2003). Furthermore, these items may cause a disproportionate amount of trouble for spoken language applications when they are mispronounced and/or misrecognized, e. g. when repeating someone’s given name in a dialogue situation. What is intended as an act of clarification may then well

be perceived as an insult, since errors in pronouncing proper names are especially prone to cause serious identification mistakes, or possibly offend the bearer of the name, or both.

6. Related research

It goes without saying that the phenomena discussed in this article are not limited to Swedish, but also occur in other languages. They also have certain consequences for spoken language applications, as we will illustrate briefly by example.

Weiss (2003) partially replicated the Xenophone experiment in an experiment using English words and German subjects. The linguistic material was a short physics text and Weiss interestingly reports results very similar to those obtained by Eklund/Lindström in their original study.

A thorough study on anglicisms in both German and Swedish was carried out by Adler (2004), who used both newspaper texts and radio and TV programs, while Morset Størseth (2005) studied the inclusion of English affricates and fricatives in Norwegian.

Abresch/Breuer (2004, 1284) investigated the perceptual aspect of xenophones in German listeners, and concluded that “[in order to] provide for an appropriate pronunciation of anglicisms, [they] have to integrate English xenophones into a German TTS system”.

Foreign items can be expected to cause quite different classes of problems for spoken language applications such as speech recognition, speech synthesis, and spoken language dialogue systems. For some applications, such as speaker verification and identification, they might even prove helpful, in the sense that any prompt involving foreign linguistic material would elicit a more scattered set of individual acoustic productions, thereby aiding any algorithm responsible for categorization.

In speech recognition, part of the problem is that foreign items will lead to increased variability, unless said variability can be understood and modeled. Many commercially very interesting applications are also truly multilingual, as already mentioned, with car navigation systems constituting a good example (Fitt 1995; Trancoso et al. 1999). Trancoso/Nunes/Neves (2006) bring up another interesting class of applications and discuss the problem of cross-lingual influence in the context of a speech-based classroom application for Portuguese. A speech database, the “Cross-Towns Corpus” (Schaden/Jekosch 2006), that includes non-native pronunciations of European city names is described in Schaden (2002a, 2003, 2006). The database is to be used in a system designed to cover directions in a total of 16 languages, including German, French, Spanish, Italian, English and Dutch.

Another study that focuses on the modelling of foreign items from a speech recognition perspective is Stemmer/Nöth/Niemann (2001). They conclude that certain domains are more prone than others to include foreign phones, like e.g. automatic services that handle film titles, where normally more than 50% of the titles are given (in Germany) in their original (untranslated) English form. Since there is a general lack of training data for the recognizer, they merged foreign and German phoneme models and report a 16.5% word error rate reduction, compared to a baseline system.

Other studies that report modelling, lexicon adaptation and generation of foreign items in speech technology applications include Mayfield Tomokyo (2001), Adda-

Decker et al. (2003), Schaden (2002b), Cremelie/ten Bosch (2001), ten Bosch/Cremelie (2002), Goronzy/Kompe/Rapp (2001), and Goronzy/Sahakyan/Wokurek (2001). The more widespread importance of properly handling foreign items in general PC applications is discussed in more detail by Junqua (2000).

For speech synthesis, the main problem related to foreign items is how to design the proper phone set, and given that, how to choose a contextually appropriate level in each and every case, so that listeners are neither left behind, nor think the synthesizer sounds ill-educated. Woodward (2003) aims at synthesizing the *Oxford English Dictionary* and reports on studies where native speakers of English pronounced words of French, Spanish, Italian and German words. In contrast to the results of Eklund/Lindström, Woodward finds a rather high degree of nativized pronunciations, especially among British speakers. Other authors who explicitly point to the use of xenophones in speech synthesis are Duit for Swedish (2004) and Steingrimsson for Icelandic (2004).

7. Conclusions

In this article, we have examined the influence of foreign (mainly English) items in Swedish, from both a synchronic and diachronic perspective. At the phonetic level, it was shown that native speakers of Swedish to a large extent expand their phonetic inventory when pronouncing foreign items, that some such xenophones are more or less required, and when that is not the case, that their usage seems to depend on language users' educational level, age, and possibly also on gender. At the morphological level it was shown that English clearly influences contemporary spoken Swedish word formation, while hardly any such examples could be found in corpora from the 1960s. Restrictions at the prosodic and morphotactic levels seem more difficult to violate, however. We have also given references to contemporary research concerning other languages, and regarding several different application domains, and it is evident that this issue extends beyond English–Swedish, although the extent to which this problem needs to be handled most certainly varies across languages as a function of a plethora of different phenomena, linguistic and non-linguistic.

It is also important to point out that not only are the observations made in this article of theoretical interest, they also have practical consequences for speech technology applications, as is also illustrated by the number of other studies devoted to the handling of foreign items in speech recognition, speech synthesis and other speech technology systems in general. Although corpora are being created that focus on the foreign item issue (notably Schaden/Jekosch 2006), there is still a general lack of available language resources within this particular field, which definitely calls for further corpus-based studies in this area.

8. Literature

- Abresch, J./Breuer, S. (2004), Assessment of Non-native Phones in Anglicisms by German Listeners.
In: *Proceedings of Interspeech 2004*, 4–8 October 2004. Jeju Island, Korea, 1281–1284.
- Adda-Decker, M./Antoine, F./Boula de Mareuil, P./Vasilescu, I./Lamel, L./Vaissiere, J./Geoffrois, E./Liénard, J.-S. (2003), Phonetic Knowledge, Phonotactics and Perceptual Validation for Auto-

- matic Language Identification. In: *Proceedings of ICPHS*, Barcelona, 3–9 August 2003. Barcelona, Spain, 747–750.
- Adler, M. (2004), Form und Häufigkeit der Verwendung von Anglizismen in deutschen und schwedischen Massenmedien. PhD thesis, Jena Thüringer Universitäts- und Landesbibliothek.
- Allwood, J. (1998), Some Frequency Based Differences between Spoken and Written Swedish. In: Haukioja, T. (ed.), *Proc. of the 16th Scandinavian Conference of Linguistics*. Turku, Finland, 18–29.
- Allwood, J./Grönqvist, L./Ahlsén, E./Gunnarsson, M. (2002), Göteborgskorpusen för talspråk. In: *Nydarne Sprogstudier* 30, special issue on “Korpuslingvistik”, 39–58.
- Andersson, L.-G./Edström, K.-O./Lilliestam, L./Norrby, C./Widenäs, K. (1999), *Gymnasisters språk- och musikvärldar*. Available at: <http://svenska.gu.se/~svekw/gsmidx.html>.
- Antunović, G. (1999), *A Dictionary of Anglicisms in Swedish*. Zagreb: University of Zagreb.
- Baayen, R. H. (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.
- Billi, R. (n.d.) Interview. Available at: http://www.hltcentral.org/usr_docs/LeJournal/I-phones_PhonemesMultimodality.pdf.
- Billi, R./Lamel, L. F. (1997), Railtel: Railway Telephone Services. In: *Speech Communication* (23), 63–65.
- Bosch, L. ten/Cremelie, N. (2002), Pronunciation Modeling and Lexical Adaptation Using Small Training Sets. In: *Proceedings of ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA) 2002*, 14–15 September 2002. Estes Park, CO, 111–116.
- Carlson, R./Granström, B./Lindström, A. (1989), Predicting Name Pronunciation for a Reverse Directory Service. In: *Proceedings of Eurospeech*. Paris, France, vol. 1, 113–116.
- Cremelie, N./Bosch, L. ten (2001), Improving the Recognition of Foreign Names and Non-native Speech by Combining Multiple Grapheme-to-phoneme Converters. In: *Proceedings of ITRW on Adaptation Methods for Speech Recognition, 29–30 August 2001*. Sophia Antipolis, France, 151–154.
- Davidson, L. (2001), Hidden Rankings in the Final State of the English Grammar. In: Horwood, G./Kim, S.-K. (eds.), *Ruling Papers*. New Brunswick, NJ: Rutgers University, vol. II, 21–48.
- Donegan, P. J./Stampe, D. (1979), The Study Of Natural Phonology. In: Dinnsen, D. A. (ed.), *Current Approaches to Phonological Theory*. Bloomington, IN: Indiana University Press, ch. 6, 126–173.
- Duit, C. (2004), Development of LottaPron – an Automatic Pronunciation Generator. MA thesis, Språktechnologiprogrammet, Dept. Of Linguistics and Philology, Uppsala University.
- Einarsson, J. (1976), *Talbankens talspråkskonkordans*. Lund University, Department of Scandinavian Languages, on CD-Rom.
- Einarsson, J. (1978), *Talad och skriven svenska. Sociolinguistiska studier*. (Lundastudier i nordisk språkvetenskap Serie C, Studier i tillämpad nordisk språkvetenskap 9.) Lund: Walter Ekstrand Bokförlag.
- Eklund, R. (2004), Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues. PhD thesis, Department of Computer and Information Science, Linköping University, Sweden.
- Eklund, R./Lindström, A. (1996), Pronunciation in an Internationalized Society: A Multi-dimensional Problem Considered. In: *Proceedings of the Swedish Phonetics Meeting (Fonetik 96) (TMH-QPSR 2/1996)*. Nässlingen, Stockholm, Sweden, 123–126.
- Eklund, R./Lindström, A. (1998), How to Handle “Foreign” Sounds in Swedish Text-to-speech Conversion: Approaching the ‘Xenophone’ Problem. In: *Proceedings of the 5th ICSLP*. Sydney, Australia, vol. 7, 2831–2834.
- Eklund, R./Lindström, A. (2001), Xenophones: An Investigation of Phone Set Expansion in Swedish and Implications for Speech Recognition and Speech Synthesis. In: *Speech Communication* 35(1–2), 81–102.
- Filipović, R. (1996), English as a Word Donor to Other Languages of Europe. In: Hartmann, R. (ed.), *English Language in Europe*. (European Studies Series.) Oxford: Intellect, 37–46.

- Fitt, S. [E.] (1995), The Pronunciation of Unfamiliar Native and Non-native Town Names. In: *Proceedings of Eurospeech 95*, 18–21 September 1995. Madrid, Spain, 2227–2230.
- Fitt, S. E. (1998), Processing Unfamiliar Words: A Study in the Perception and Production of Native and Foreign Placenames. PhD thesis, University of Edinburgh.
- Good, I. J. (1953), The Population Frequencies of Species and the Estimation of Population Parameters. In: *Biometrika* 40, 237–264.
- Gorin, A. L./Riccardi, G./Wright, J. H. (1997), How May I Help you? In: *Speech Communication* 23(1/2), 113–127.
- Görlach, M. (ed.) (2001), *A Dictionary of European Anglicisms*. Oxford: Oxford University Press.
- Goronyz, S./Kompe, R./Rapp, S. (2001), Generating Non-native Pronunciation Variants for Lexicon Adaptation. In: *Proceedings ITRW on Adaptation Methods for Speech Recognition*, 29–30 August 2001. Sophia Antipolis, France, 143–146.
- Goronyz, S./Sahakyan, M./Wokurek, W. (2001), Is Non-native Pronunciation Modelling Necessary? In: *Proceedings of Eurospeech 2001*, 3–7 September 2001. Aalborg, Denmark, 309–312.
- Graddol, D. (2006), *English Next*. British Council. Available at: <http://www.britishcouncil.org/learning-research-englishnext.htm>.
- Gronostaj, M. T. (n.d.) *The Swedish PAROLE Lexicon. A Language Engineering Resource with Access to Morphological and Syntactic Information in Swedish, Developed by Språkdata*. Gothenburg University. Available at: <http://spraakbanken.gu.se/>.
- Itô, J./Mester, A. (1999), *The Handbook of Japanese Linguistics*. Malden, MA and Oxford, UK: Blackwell Publishers, chapter “The Structure of the Phonological Lexicon”, 62–100.
- Jespersen, O. (1902), Engelsk og nordisk: En afhandling om läneord. In: *Nordisk tidsskrift för vete-*
skap, konst och industri, 500–514.
- Jungbluth, K. (1999), Two- and Three-dimensional Deictic Systems between Speech and Writing – Evidences from the Use of Demonstratives in Romance Languages. In: André, E./Poesio, M./Rieser, H. (eds.), *Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief. Work-*
shop held at the 11th European Summer School in Logic, Language and Information (ESSLLI), ESSLLI. Utrecht University, Utrecht, the Netherlands, paper 4, 12–19.
- Junqua, J.-C. (2000), *Robust Speech Recognition in Embedded Systems and PC Applications*. Boston: Kluwer Academic Press.
- Kornai, A. (2002), How Many Words are There? In: *Glottometrics* (4), 61–86.
- Ladefoged, P./Maddieson, I. (1996), *Sounds of the World's Languages*. Oxford: Blackwell.
- Lindström, A. (2003), Non-native Linguistic Elements in Spoken Swedish. In: *Proceedings of the 15th ICPHS*. Barcelona, Spain, 2353–2356.
- Lindström, A. (2004a), English and Other Foreign Linguistic Elements in Spoken Swedish. PhD thesis, Department of Computer and Information Science, Linköping University.
- Lindström, A. (2004b), English Influence on Swedish Word Formation and Segmentals. In: *Nordic Journal of English Studies* 3(2), 115–142.
- Lindström, A./Eklund, R. (1999), Xenophones Revisited: Linguistic and Other Underlying Factors Affecting the Pronunciation of Foreign Items in Swedish. In: Ohala, J. J./Hasegawa, Y./Ohala, M./Granville, D./Bailey, A. C. (eds.), *Proceedings of the 14th ICPHS*. San Francisco, CA, 2227–2230.
- Lindström, A./Eklund, R. (2000), How Foreign are “Foreign” Speech Sounds? Implications for Speech Recognition and Speech Synthesis. In: *ESCA-NATO Tutorial and Research Workshop on Multi-lingual Interoperability in Speech Technology (ETRW MIST'99)*. (NATO Research and Technology Organization (RTO) Meeting Proceedings 28.) Hull, Québec, Canada, AC/323(IST)TP/4, 15–19.
- Lindström, A./Eklund, R. (2002), Xenophenomena: Studies of Foreign Language Influence at Several Linguistic Levels. In: *Proceedings of 24. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft: Mehrsprachigkeit Heute*. Universität Mannheim, 132–134.
- Ljung, M. (1986), *Skinheads, hackers och lama ankor. Engelskan i 80-talets svenska*. Stockholm: Trevi.

- Lüdeling, A./Schmid, T. (2001), Does Origin Determine the Combinatory Properties of Morphological Elements in German? In: Booij, G./DeCesaris, J./Ralli, A./Scalise, S. (eds.), *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting*. Barcelona: IULA, 255–265.
- Mayfield Tomokiyo, L. (2001), Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in LVSCR. PhD thesis, School of Computer Science, Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA.
- Melchers, G./Shaw, P. (2003), *World Englishes. An Introduction*. London: Edward Arnold.
- Morset Størseth, T. (2005), Language Context as Phonetic Discriminator of English Xenophones in Norwegian. MA thesis, Department of Modern Foreign Languages, Faculty of Arts, NTNU Dragvoll, Trondheim, Norway.
- Norrby, C./Wirdeñäs, K. (1998), Language and Music Worlds of Senior High School Students. In: Pedersen, I.-L./Scheuer, J. (eds.), *Sprog, kon og kommunikation. Rapport fra 3. Nordiske konference om sprog og kon*. Copenhagen: C. A. Reitzels forlag, 155–163.
- Parole Consortium (n.d.) <http://www.ub.es/gilcub/SIMPLE/simple.html>.
- Pettersson, P. A./Forsberg, K. (1970), *Beskrivning och register över Eskilstunainspelningar*. (FUMS Rapport nr. 10, del 1–2.) Forskningskommittén i Uppsala för Modern Svenska, Uppsala, Sweden.
- Prince, A./Smolensky, P. (1995), *Optimality Theory: Constraint Interaction in Generative Grammar*. (Linguistic Inquiry Monograph Series.) Cambridge, MA: MIT Press.
- Schaden, S. (2002a), A Database for the Analysis of Cross-lingual Pronunciation Variants of European City Names. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC) 2002*, 29–31 May 2002. Las Palmas de Gran Canaria, Spain, vol. 4, 1277–1283.
- Schaden, S. (2002b), Regelbasierte Generierung fremdsprachlich akzentgefärbter Aussprachevarianten. In: *Proceedings of ESSV 2002*, 25–27 September 2002. Dresden, Germany, 289–293.
- Schaden, S. (2003), Non-native Pronunciation Variants of City Names as a Problem for Speech Technology Applications. Text, Speech and Dialogue. In: *Proceedings of the Sixth International Conference TSD 2003*. (Lecture Notes in Artificial Intelligence 2807.) České Budějovice, Czech Republic/Berlin etc.: Springer, 229–236.
- Schaden, S. (2006), Evaluation of Automatically Generated Transcriptions of Non-native Pronunciations Using a Phonetic Distance Measure. In: *Proceedings of LREC 2006*, 24–26 May 2006. Genova, Italy. Available at: http://www.shameacademy.de/publications/lrec2006_evaluation_preprint.pdf.
- Schaden, S./Jekosch, U. (2006), “Casselberveetovallarga” and Other Unpronounceable Places: The CrossTowns Corpus. In: *Proceedings of LREC 2006*, 24–26 May 2006. Genova, Italy. Available at: http://www.shameacademy.de/publications/lrec2006_crosstowns_preprint.pdf.
- Schmid, T./Lüdeling, A./Säuberlich, B./Heid, U./Möbius, B. (2001), DeKo: Ein System zur Analyse komplexer Wörter. In: Lobin, H. (ed.), *Proceedings of GLDV*. Gießen, Germany, 49–57.
- Sharp, H. (2001), *English in Spoken Swedish. A Corpus Study of Two Discourse Domains*. (Stockholm Studies in English 95.) PhD thesis. Stockholm: Almqvist & Wiksell International.
- Shriberg, E. (1994), Preliminaries to a Theory of Speech Disfluencies. PhD thesis, Department of Psychology, University of California, Berkeley.
- Skutnabb-Kangas, T. (2000), *Linguistic Genocide in Education – or Worldwide Diversity and Human Rights?* Mahwah, NJ: Lawrence Erlbaum.
- Söderberg, B. (1983), *Från ryttars och cowboys till tjuvstrykers. S-pluralen i svenska. En studie i språklig interferens*. Stockholm: Almqvist & Wiksell.
- Spiegel, M. F./Macchi M. J./Gollhardt, K. D. (1989), Synthesis of Names by a Demisyllable-based Speech Synthesizer (Spokesman). In: *Proceedings of Eurospeech*. Paris, France, vol. 1, 117–120.
- Steingrimsson, S. (2004), Bilingual Voice for Unit Selection Speech Synthesis. MA thesis in Speech and Language Processing, Theoretical and Applied Linguistics, School of Philosophy, Psychology and Language Sciences, University of Edinburgh.

- Stemmer, G./Nöth, E./Niemann, H. (2001), Acoustic Modeling of Foreign Words in a German Speech Recognition System. In: *Proceedings of Eurospeech 2001*, 3–7 September 2001. Aalborg, Denmark, 2745–2748.
- Teleman, U. (1974), *Manual för grammatsk beskrivning av talad och skriven svenska (Mamba)*. Lund: Studentlitteratur.
- Thomason, S. G./Kaufman, T. (1988), *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
- Trancoso, I./Nunes, R./Neves, L. (2006), Classroom Lecture Recognition. Computational Processing of the Portuguese Language. In: *7th International Workshop, PROPOR 2006*. Itatiaia, Brazil, 190–199.
- Trancoso, I./Viana, C./Mascarenhas, I./Teixeira, C. (1999), On Deriving Rules for Nativised Pronunciation in Navigation Queries. In: *Proceedings of Eurospeech 1999*, 5–9 September 1999. Budapest, Hungary, 195–198.
- Weinreich, U. (1953), *Languages in Contact*. The Hague: Mouton.
- Weiss, B. (2003), Are Native English Words Different from Neoclassical English Words for German Speakers? Unpublished document.
- Woodward, G. (2003), Synthesizing the Oxford English Dictionary. MSc thesis, Dept. of Speech and Language Processing, University of Edinburgh.
- Zipf, G. K. (1935), *The Psycho-biology of Language*. Boston, MA: Houghton Mifflin.

Anders Lindström and Robert Eklund, Farsta (Sweden)

49. Corpora and discourse analysis

1. Introduction
2. Tracing the joint history
3. The present: Studies and findings
4. Discourse-oriented applications of corpora
5. Methodological issues
6. Conclusion: New vistas for discourse analysis?
7. Literature

1. Introduction

Corpora and discourse analysis have a troubled relationship. Yet, it is a steady relationship going well back in corpus-linguistic time, and one that both parties are highly motivated to keep up despite its many hazards and challenges. Corpora can be of many kinds, as shown by the entries in this volume; some will obviously turn out to be more suitable for the purposes of discourse analysis than others. Discourse analysis, again, is here interpreted in a very wide sense, encompassing text linguistics, discourse analysis/studies, conversation analysis and pragmatics. Some of the observations made below are also relevant to several adjacent areas of study such as rhetoric, narratology and linguistic stylistics. Of related interest are articles 6, 29, 38, 51, 50 and 47.

- Stemmer, G./Nöth, E./Niemann, H. (2001), Acoustic Modeling of Foreign Words in a German Speech Recognition System. In: *Proceedings of Eurospeech 2001*, 3–7 September 2001. Aalborg, Denmark, 2745–2748.
- Teleman, U. (1974), *Manual för grammatsk beskrivning av talad och skriven svenska (Mamba)*. Lund: Studentlitteratur.
- Thomason, S. G./Kaufman, T. (1988), *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
- Trancoso, I./Nunes, R./Neves, L. (2006), Classroom Lecture Recognition. Computational Processing of the Portuguese Language. In: *7th International Workshop, PROPOR 2006*. Itatiaia, Brazil, 190–199.
- Trancoso, I./Viana, C./Mascarenhas, I./Teixeira, C. (1999), On Deriving Rules for Nativised Pronunciation in Navigation Queries. In: *Proceedings of Eurospeech 1999*, 5–9 September 1999. Budapest, Hungary, 195–198.
- Weinreich, U. (1953), *Languages in Contact*. The Hague: Mouton.
- Weiss, B. (2003), Are Native English Words Different from Neoclassical English Words for German Speakers? Unpublished document.
- Woodward, G. (2003), Synthesizing the Oxford English Dictionary. MSc thesis, Dept. of Speech and Language Processing, University of Edinburgh.
- Zipf, G. K. (1935), *The Psycho-biology of Language*. Boston, MA: Houghton Mifflin.

Anders Lindström and Robert Eklund, Farsta (Sweden)

49. Corpora and discourse analysis

1. Introduction
2. Tracing the joint history
3. The present: Studies and findings
4. Discourse-oriented applications of corpora
5. Methodological issues
6. Conclusion: New vistas for discourse analysis?
7. Literature

1. Introduction

Corpora and discourse analysis have a troubled relationship. Yet, it is a steady relationship going well back in corpus-linguistic time, and one that both parties are highly motivated to keep up despite its many hazards and challenges. Corpora can be of many kinds, as shown by the entries in this volume; some will obviously turn out to be more suitable for the purposes of discourse analysis than others. Discourse analysis, again, is here interpreted in a very wide sense, encompassing text linguistics, discourse analysis/studies, conversation analysis and pragmatics. Some of the observations made below are also relevant to several adjacent areas of study such as rhetoric, narratology and linguistic stylistics. Of related interest are articles 6, 29, 38, 51, 50 and 47.

A prerequisite for a discussion of the relationship between corpora and discourse analysis is a consideration of two fundamental aspects of discourse. The first of these is concerned with text and discourse as product and process, the second with the primary notion of context.

Corpora consist of (fragments of) texts and discourses presented as products. These products are the outcome of various dynamic processes which have taken place between interlocutors, and which were potentially affected by onlookers (or lurkers), in particular and always unique communication situations. Such dynamic processes include situated decisions concerning, for instance, information structuring, the linear and hierarchical organization of discourse, singling out important portions of text from the rest, indicating and interpreting affect, and making connections between the words and the world. Linguistic choices pertaining to such discourse processes are not made in a vacuum; rather, they emerge from joint discourse work, conscious or not, between interlocutors in a particular communication situation. The form and function of a text or piece of discourse are thus influenced by the situated discourse practices in which interlocutors as subjects are engaged at a given point or period of time. Their form and function are also affected by the shared intertextual and interdiscursive repertoire of knowledge that is emergent in particular contexts. And the unfolding texts and discourses provide impetus for the construction, maintenance or alteration of contexts and cultures; in a sense, they serve to create contexts and cultures. Texts and discourses thus constitute an important meeting point between individual and distributed cognition: they are affected by the communicative goals of the interlocutors in given situational and socio-cultural contexts, their activated encyclopaedic knowledge, and the dynamism of the recursive processes of co-construction, negotiation and adaptation that contribute to the form and function of the end-products, presented to analysts as corpus data.

The second prerequisite for the discussion of corpora and discourse analysis has to do with the primary notion of context. The two-way traffic between texts and contexts constitutes a nexus which discourse linguists wish to study. Traditionally, a distinction has been made between textual (linguistic) context – or co-text – situational context and the context of culture. Studying entire texts in context is one of the *raisons d'être* of text linguistics and discourse analysis, and the context to be taken into account in such studies has expanded maximally over the past forty years. This is evident if we view the five dimensions of text and discourse presented in section 2 as a chronology of text-linguistic and discourse-analytic enquiry. In line with the dynamic nature of discourse dealt with above, it is the expanding notion of context that proves to be extremely problematic in corpus studies: co-text can be increasingly made available to analysts using modern software but the situated aspects of discourse constitute a real challenge.

Yet, corpus studies and their applications often happily employ words such as ‘authentic’ or ‘real-life’ to characterize the data and to promote the chosen method. The notion of authenticity, in fact, became fashionable in linguistics and the study of languages during the last quarter of the 20th century: individual texts studied by linguists of various orientations, examples offered in dictionaries and grammar books as well as multi-purpose corpora planned for linguistic study were increasingly given this epithet. But the problem is that what is authentic – or real-life – data for other kinds of linguistic study is not necessarily so for the discourse linguist.

Recent discussions in discourse studies and sociolinguistics indicate a need for a critical reanalysis of the idea of authenticity (see e.g. the special issues of *Discourse Studies*

3:4 (2001) and *Journal of Sociolinguistics* 7:3 (2003); cf. also Gill 2008). In its simplest sense, ‘authentic language’ is language that has been attested and is attestable – as is the ‘real language’ stored in a corpus, or the texts on the discourse linguist’s screen or desk, sometimes available as recordings. But even this commonsense notion of authenticity as naturally occurring language is in itself problematic. Authenticity turns all the more tricky when we recall that discourse serves to index situational and socio-cultural contexts. Preserving the ‘authentic’ spontaneity of impromptu speech in a transcription is an impossible task as has been noted by conversation analysts who continuously worry about this process of interpretation (cf. e. g. the classic article on transcription as theory by Ochs 1979). Making the criteria for authenticity stronger leads to a complex notion that might serve the discourse linguist better; yet, such a concept perforce lies beyond the reach of the corpus compiler. In corpus studies, authenticity of data has been to some extent discussed by linguists working with historical materials (e. g. Kytö 2000).

Some contextual information is inherently present in textual data. Accordingly, many corpora include as much co-text (i. e. linguistic context) as has been deemed expedient for the kinds of study for which they have been planned. Corpora may also provide analysts with a coding of some of the main variables of the situational and socio-cultural contexts from which the texts included in the corpus have been extracted. Software, too, has grown more friendly towards the needs of the discourse-oriented corpus linguist: several programs now allow the retrieval of items with more co-text than was previously possible, and contextual codes and tags can be used as a point of departure for automatic searches. Contextual coding involves interpretation of the typically decontextualized data by external observers who have not normally been participants in the particular communication situation or who access the texts in ways and for purposes other than those in which most interlocutors would. The corpus linguist analysing coded data is therefore faced with relatively static contextual interpretations which may need to be questioned in the light of the analysis. Some corpus linguists, in fact, argue for the importance of using uncoded data instead. What is important, however, is to include entire texts where possible.

In discourse linguistics the investigation of text and discourse is motivated by the wish to study processes, irrespective of whether such processes are primarily linguistic, cognitive or social in character. The decontextualized products (or parts of such products) included in a corpus, in a recontextualized form, bear traces of the various dynamic processes that influenced the form they came to adopt, and it is these traces that we can thus hope to study with hindsight using large bodies of data. Corpus data have perforce been decontextualized as they have been separated from their situational and socio-cultural contexts, and often their overall textual context, too. And they have been recontextualized in the sense that they have been given a new context in the corpus itself. For instance, a news story which first appeared in a printed paper can be included as such in a corpus consisting of texts published in the particular newspaper over a given period of time. The new context of the text is thus very different from the original one, in which it appeared in another shape in a given place on the page, adjacent to other verbal and visual materials. The particular section of the paper and other contextual aspects deemed important by the corpus compilers may be coded in the corpus. Yet, these data are presented to the analyst in the new corpus context, rather than the original newspaper context. This will have a bearing on the kinds of analyses the discourse linguist may hope to undertake using the given corpus data.

In contrast to the in-depth study of an individual text or piece of discourse in context, corpora are expected to provide us with an opportunity to draw parallels between large numbers of such texts. The idea is that the contribution of the informed discourse linguist and the insights of a corpus linguist, residing in one and the same person, might reveal patterns of discourse that we have not previously been aware of. Yet, the dynamic nature of text and discourse processes cannot be retrieved from any standard corpus, irrespective of its size or whether the corpus is finite or one that is continuously added to. And turning processes into counts means that this dynamism is irretrievably lost. But corpora can still be planned so that they allow investigations of discourse phenomena: in what follows we will review some of the central areas of study where corpora have been exploited to come to grips with aspects of text and discourse. It is, however, crucial to always keep in mind the inherently static nature of these computerized data; the study of corpora should, indeed, invite us to look beyond the thus recontextualized products and use them, to the extent that this is possible, as evidence for the multifaceted processes that have contributed to their construction in the first place. Seen in this way, corpus data can ideally also help us understand conventionalized and emergent contextual phenomena which contribute to systematic variation within and across texts and discourses in given socio-cultural contexts and through time.

In this light, one of the basic software requirements should be to allow analysts to keep track of the unfolding nature of discourse data, a requirement which has not yet proved to be self-evident or straightforward. The strained relation between a product and process approach to corpus data and the possibilities of adopting a dynamic view of discourse including context constitute the biggest stumbling block for the future development of the relationship between corpora and discourse analysis.

In what follows we will first take a look back in time, to trace the joint history of corpora and discourse analysis from the perspectives of each of the two parties to the relationship. The main bulk of this article will be devoted to an overview of corpus work done in the fields of text and discourse linguistics and pragmatics, including conversation analysis. There is also a section on discourse-oriented applications of corpora and another on methodological concerns.

2. Tracing the joint history

To appreciate the troubled nature of the relationship between corpora and discourse analysis, it is instructive to trace their joint history starting from each separately. The discussion below is built around five complementary dimensions of text and discourse; not all of them will be equally accessible to users of corpus methods. Another distinction that will be repeatedly referred to is that between corpus-based and corpus-driven approaches.

Building on the four models of the study of text and discourse distinguished by Enkvist (1984), we can today speak of five different dimensions of discourse. While reflecting the chronology of approaches in text and discourse linguistics, they are all still in use, each contributing to our understanding of discourse phenomena from different but equally important points of view. To start with, we can speak of a ‘structural dimension’, i. e. the approach to text and discourse as structure that is the outcome of a pro-

cess, a structure that is constructed using lexico-grammatical tools. The second dimension is ‘content-oriented’, primarily focusing on the propositional content to be communicated. Thirdly, we can single out a ‘cognitive dimension’, where the concern is with the production and interpretation of discourse. The fourth dimension is ‘interactional’: here the point of departure is the co-construction of discourse. And finally, we can speak of a ‘social dimension’ of discourse, where the starting point is the socio-cultural context shaping discourses and being in turn shaped by them. (For further discussion, see Virtanen 1997.)

Another methodologically important distinction that is needed if we are to understand the joint history of corpora and discourse analysis is the one that Sinclair (2004, 12) draws between ‘corpus-based’ linguistics and ‘corpus-driven’ linguistics. Corpus-based studies are concerned with topics and problems predicted from advances in other fields of language study, and the concepts and tacit assumptions tend to come with the package. Corpus-driven studies, Sinclair (2004, 191) states, demand very large corpora and essentially an open mind; texts and discourses are here approached from non-conventional angles and caution is taken not to adopt the pre-existing categories of linguistics as such. The metaphor in this case is one of ‘uncontaminated text’.

2.1. From corpora to discourse analysis

Linguists engaged in corpus compilation and analysis of corpus data have traditionally been interested in issues of grammar and the lexicon. The motivation has often been a need to account for linguistic variation in a given period of time or through time (i. e. linguistic change). But words and structures indicating variation and highlighting change inevitably lead to texts and discourses as this is the environment in which they are used; it is in a particular context that they are attested as examples of particular linguistic phenomena. And conflicts and conspiracies between grammar and text, or lexical elements and text, suggest discourse-linguistic perspectives and tools. Using corpora in such research has predominantly led to what was above characterized as corpus-based studies. Such investigations have typically come to a stop where the possibilities offered by existing corpora and the software selected for use meet their limits in terms of discourse analysis. One option has then been to switch over to discourse analysis of the traditional kind – and another to try to collect additional, focused data and write programs to single out linguistic items and relations that are of interest for discourse-analytic purposes. This is not, however, a simple task and studies may therefore in the end risk showing what we already knew about discourse; yet, letting the technology decide what is possible to investigate is obviously not the solution linguists wish to opt for, either.

Corpus-driven studies, again, promise to relieve the worry felt by the context-sensitive discourse linguist. The idea of corpus-driven investigation is one in which the so-called very large corpora invite us to contemplate discourse phenomena in novel ways: we are offered an opportunity to investigate the interplay of linguistic aspects of different kinds by examining concordance data and hence approaching discourse in an unorthodox order (as compared to the processes of reading texts or transcriptions and listening to spoken discourse) and investigating textual characteristics with the help of statistical

information. Still, it is a fact that the dynamism and the authentic situational and socio-cultural context of the ‘uncontaminated text’ is here, too, irreversibly lost, even if we can try to reconstruct contexts and cultures for the data (i. e. interpret discourse from a particular point of view and for a particular purpose). The refusal to work with coded text and the insistence on having an open mind are reminiscent of some of the ideas of conversation analysts concerning the refusal to work with predetermined categories and their discussions of the status of texts and discourses as evidence, as separate from the social contexts in which they emerged. It is also worth noting that uncontaminated text cannot, by definition, undergo hybridization, in terms of observed interdiscursivity; the notion is thus incompatible with contextual concerns related to genre. Hybridization refers to processes of combining elements of different kinds of discourse, the notion thus presupposing the existence of such categories in the first place (see the discussion in section 3.4.).

In terms of the five dimensions of text and discourse touched upon above, it is possible to categorize the development from corpora to discourse analysis as follows. If we start from (standard) corpora, we are likely to view discourse as structure, rather than process as we are searching for the formal exponents of whatever discourse processes we would like to get at – or the lack of them, for that matter. Basically what we need to do is to decide what counts as, say, a hedge before studying the delicate process of hedging in corpus data. We wish to find formal exponents of the functions we are interested in, to expand our generalizations to what will be possible on the basis of corpus evidence, as compared to the in-depth study of individual texts. However, if we start from discourse, we would like to study functions and processes without distorting the multifaceted and dynamic nature of its (co-)construction as we know this to be a fundamental characteristic of discourse. While opening new – and possibly unexpected – avenues to the study of text and discourse, corpus-driven methods still suffer from the lack of contextual dynamism that might help us to complement the structural and content-oriented analyses with investigations that truly take into account the bidirectional relation of these corpus data and their situated and socio-cultural aspects. This is the main problem on the road from corpora to discourse analysis, and the solutions offered range from discourse-sensitive tagging to the compilation of focalized corpora consisting of entire texts where possible. Opinion concerning the cognitive status of corpus data is divided between, on the one hand, the view that assigns them the same kind of idealized role as discussed in terms of the myth of the native speaker in applied linguistics, the idea that very large corpora somehow provide access to ‘langue’ (see e. g. Tognini Bonelli 2001, 3), or alternatively, that they reflect processes of distributed cognition (see 2.2. and 3.3.), and on the other hand, the standpoint that the issue is not relevant and can therefore be ignored in corpus linguistics.

2.2. From discourse analysis to corpora

Text and discourse linguists have been interested in electronic corpora from the very beginning. The text categories of the early corpora (Brown, LOB and others modelled in the same fashion) were meticulously planned to provide a high degree of balance and representativeness to these archives of written language (cf. article 9 for a discussion of

the ‘representativeness’ and ‘balance’). The SEU and the LLC opened new avenues to the study of spoken discourse (cf. article 47). Section A of the present volume offers insights into the impressive technological development that has taken place in corpus linguistics.

Yet, discourse linguists have often been disappointed with the outcome of their corpus studies; accepting the lack of sufficient ‘precision’ in their area of study, they have performed time-consuming searches on aspects of discourse in which the outcome has either been something that they were already familiar with or which they could not trust because of the insufficient ‘recall’ of the search procedure (for the two terms, see Brodda 1991). The exercise has in due course turned them into better detectives who practise indirect methods to get what they want out of a corpus (i. e. sufficient recall), a procedure which has opened their eyes to new aspects of discourse. The road has, however, been winding because discourse linguists doing corpus work usually suspect that there is more to the phenomenon they are investigating than meets the eye, and more importantly, that this may indeed be more to the point than what their searches have allowed them to find. They also know that frequency is not all: the most interesting discourse phenomena need not be particularly frequent and yet interlocutors may well be aware of them and use them as heuristics when they co-construct, adapt to and interpret discourse in context. But despite the lack of high hopes in this area of study, corpora have a fascination: in addition to their obvious function as a test ground for hunches, the expectation is that they might offer new insights simply because analysts are forced to enter the texts in an unconventional way. This aspect will come to the fore in section 3.3.

Apart from the lack of contextual information, the main obstacles for the discourse linguist interested in the use of corpus data have been the following: (i) many standard corpora include fragments of text, rather than entire texts; (ii) the sampling and excerpting procedures have not usually been planned with discourse studies in view; (iii) the typological classification may be based on heterogeneous or ad hoc criteria; (iv) the sheer amount of output provided by triggers of textual information can be overwhelming (i. e. lack of precision); and (v) discourse linguists face serious problems in trying to find what they want to find in a corpus (i. e. insufficient recall). In addition, there is the issue of representativeness. When discourse linguists know that a given corpus is representative enough for the study of text and discourse, they are likely not to need it at all; when they do not know whether it is representative, they cannot use it (cf. the ‘representativeness paradox’ discussed in Mair 1990, 14).

Very large corpora may delude us into believing that some of these problems will be solved as soon as we let the corpus steer our investigations. For instance, Tognini Bonelli (2001, 3) argues that, in contrast to texts, “[t]he type of information gathered from a corpus is evaluated as meaningful in that it can be generalized to the language as a whole, but with no direct connection with a specific instance”. Hence for her, “the patterns shown up by corpus evidence yield insights into *langue*”. For many text linguists and discourse analysts, used to working with what Tognini Bonelli calls instances of *parole*, ‘language as a whole’ cannot ever be more than a myth (see e. g. the discussion in Enkvist 1991).

In corpus studies, then, it is difficult enough to study text or discourse structure using corpora: often the items to be searched are far too frequent in the corpus or they perform too many different functions in texts for a corpus-based study to be meaningful. The outcome of such studies might simply be that the items searched for are ‘multifunc-

tional', which does not take us very far on our way to understanding the workings of texts and discourses. But the opposite may also be true: the items searched for can also be too few to show in other than very large corpora and yet important enough to provide us with a key to some central issue in the study of discourse organization. Counts can also be misleading in themselves: unless a manual check-up is performed, we run the risk of drawing conclusions that prove to be false when the actual texts of the corpus are analysed individually.

When the main focus is on the content, as in rhetorically oriented studies of text and discourse, lexical key words will be helpful and the investigation of lexical clusters and collocation rewarding. However, it will be difficult to be aware of instances where the propositional and structural dimensions of a text diverge, and the recall of the search procedure will thus not be as good as it should. In other words, here, too, the predetermined set of key words may lead to a deterministic outcome. Yet, the possibility of rerunning searches in light of findings is part of the fascination: while looking for key words related to particular content, we learn about other, perhaps more important ways of conveying aspects of such content in the data.

Whether a corpus in some sense represents the shared memory of some socio-cultural community is an issue worth pondering. It appears to be naive, however, to argue for a straightforward link between conceptualization and the results of discourse-oriented metaphor analysis of corpus data. Yet, the thought that at least the largest corpora would provide us with access to distributed cognition, however limited, is not without attraction to many. In this light, it is of interest to note that psycholinguists have started to use corpora to complement experimental data (cf. e.g. Pander Maat/Sanders 2001) and cognitive linguists occasionally rely on corpus data as evidence for what seems difficult to investigate otherwise, as in advances in construction grammar proposed by Charles J. Fillmore. As always in the study of discourse and cognition, we need to approach them using nothing less than discourse and cognition (see e.g. contributions to Virtanen 2004b). It is, however, obvious that discourse-oriented cognitive linguistics will make use of corpus data for purposes that are very different from mainstream corpus linguistics. Yet, many of these endeavours are linked to attempts to understand how discourse works.

The interactional dimension of discourse has been approached through corpora of spoken language, even though central pragmatic topics have also been studied using written data. Section 3.2. is devoted to interaction. At this point it is in order to note that these studies, too, have been predominantly corpus-based, rather than corpus-driven, and the motivation for examining particular particles or expressions from the perspective of their interactional functions has usually stemmed from studies that have started out from individual texts and discourses in context. It is, however, in this area that some of the most interesting corpus work has taken place as concerns discourse analysis and pragmatics.

The socio-cultural dimension of discourse poses the biggest problems as any attempt to reduce such holistic phenomena to a set of search items, without distorting the variability and contextual dynamics of discourse, is by definition doomed to failure. Yet, corpora are increasingly and successfully used in sociolinguistic and pragmatic enquiry as several of the articles in this volume suggest and as we will see in section 3.

Irrespective of whether we start from corpora or from discourse analysis, there will be problems. But problems are there to be solved: the dialogue can be turned into a

constructive one for both parties once we realize the significance of the different historical roots and the assumptions that come with them. While corpus linguists have often been primarily interested in describing what is there and what is not there in a corpus, the point of discourse analysis is to understand how discourse functions. The practitioners of each direction have not always understood and perhaps fully appreciated the motives behind the preoccupations of the other. Still, there are signs that this need not be the case in the future. The common denominator is the reliance on data in both corpus linguistics and discourse linguistics, even though the size and type of data vary from individual texts in context to large decontextualized corpora, which has important implications for their status in the investigation. There are clear ontological and epistemological differences (see section 5). With hindsight, however, we can see that corpora have contributed in important ways to the study of linguistic variation across texts when it has been possible to attain sufficient recall and tolerable precision in the search procedure. Secondly, corpora have let us approach organizational and interactional aspects of impromptu speech to an extent which would not have been possible otherwise. Thirdly, corpus-driven studies appear to offer new ways of approaching discourse; what kinds of new insights the future has in store for the followers of this road cannot as yet be fully envisioned.

Having dealt with some of the principled differences that make the relationship between corpora and discourse analysis an uneasy one, we are now in a position to go on to contemplate some of the findings of discourse-linguistic studies of corpus data.

3. The present: Studies and findings

Corpus-based studies of discourse can be broadly divided into four areas according to whether their main concern is with (i) discourse organization and text structure, (ii) discourse-pragmatic aspects of interaction, (iii) textual and pragmatic collocation, or (iv) variation across texts and discourses. The last of these has turned out to be an area where corpora can be used not only to investigate variation, and to some extent variability, in a large body of data but also to reveal dimensions of variation not anticipated before the analysis. These latter findings can ultimately take us from corpus-based studies to corpus-driven investigations. Such studies have so far mainly concerned lexical and phrasiological issues, in particular collocation, which have disclosed intriguing aspects of discourse phenomena. Another area that has profited from corpus-based investigation is the study of the explicit signalling of interactional aspects of discourse. It is perhaps in the core area of coherence, text structure and discourse organization, that the impact of corpora has proven to be most problematic as the study of these phenomena is difficult and liable to error without access to entire texts in context.

In what follows the discussion starts with some of the main topics in discourse linguistics, to see how corpora can facilitate their study and lead to new findings. One of the advantages of corpora in discourse analysis is the fact that they can be used in ways that discard established divisions of linguistics. The concluding section accordingly explores some of the shortcuts between the different dimensions of discourse that are conspicuous in studies adopting corpus-linguistic methodology.

3.1. Discourse organization, text structure

Three main profiles have been singled out in studies of text structure and discourse organization: (i) cohesion and coherence, (ii) information dynamics, and (iii) (fore)-grounding, which is closely tied to point of view. These are umbrella concepts: each of them consists of a large number of recursively activated parameters which can be observed to conspire towards one and the same communicative goal or enter into conflict with one another. The realizations of the three have been shown to vary across the different modes of speech and writing, different discourse types, genres, registers and styles. Each of the three discourse profiles is closely linked to context and each, in turn, serves to create context. Such inherently dynamic characteristics of discourse organization make their investigation difficult unless we have recourse to entire texts in context. The main motivation for even trying to come to grips with these phenomena using corpus data thus lies in the promise of corpora to permit more general statements about the nature of discourse organization than might be possible on the basis of an in-depth analysis of individual texts.

Written texts allow us to explore textuality through corpus-based analyses of topic continuity. Once a hypothesis of the mechanisms of topic continuity has been formulated on the basis of the study of texts and discourses, it is to some extent possible to obtain quantitative data, to test the validity of such hypotheses in a text corpus. Hence, the distance of mentions of a referent in a text can be counted and the linguistic realizations of the references can be categorized and analysed (cf. e. g. the early studies included in Givón 1983; cf. also Biber et al. 1998, 106–132). Text-linguistic studies have shown that the placement of discourse-strategic markers of various kinds in the clause, sentence, typographic paragraph, textual unit or text can serve the functions of indicating coherence and text segmentation; the choice of placing them at the beginning, end or somewhere in the middle of a text or textual unit has been shown to have implications for text structure and discourse organization (cf. e. g. Enkvist 1984; Virtanen 1992, 2004a). To conduct a meaningful corpus-based analysis of the exponents of various discourse strategies, however, we need a corpus which consists of entire texts or long stretches of speech and adequate information about the origins of the data. It is also helpful to know what text categories are represented in the corpus; for instance, genres and registers selected on the basis of non-linguistic criteria allow us to link our findings to different categories of text. Alternatively, we might investigate references to continuous participants (topics), which we detected through word frequency analysis, and come up with a categorization of texts in that corpus on the basis of the realization forms of the references. Text-linguistic studies have shown that participant reference works in a different way in narrative and non-narrative and that the main participants in narrative tend to be singled out from the minor ones by the linguistic form used to refer to them. Such tendencies should be exposed in suitable corpora consisting of different genres and types of discourse. As such abstract categories of discourse are, however, normally realized through blends of exponents that combine with one another in a text, the tendencies we uncover in a corpus may well be signs of some other discourse phenomenon while essential signals may go unnoticed because we did not come to think of searching for them in the texts. Discourse-analytic methods are thus needed in parallel when corpus data are used for investigations of discourse organization. Genres and discourse types are not of equal value in a particular context. Also, narrative is a strong type of discourse whose

explicit signals in a text can therefore be few. Serious problems faced in corpus-based studies of coherence are caused by its implicitness and context-sensitivity, which may not be revealed to the analyst through corpus data.

Linguistic signals of cohesion and coherence also serve to indicate text segmentation. Identification of elements that mark boundaries between units of text of various kinds has been central in the study of text and discourse. The textual scope to be taken into account varies from local, micro-level units to global relations at the macro-level of the text. There is a remarkable body of discourse-linguistic literature showing how grammar and lexis can serve such purposes. Corpus studies attempting to assist in this task start from lexical items or grammatical tags to investigate patterns in a larger body of data than is possible in the context-bound analysis of text and discourse. Descriptions of what linguistic material is found, for instance, at the outset of sentences in text corpora can be conducted but explanation of the findings demands discourse-linguistic expertise.

Impromptu speech has been shown to include a much larger number of situationally anchored references than writing, which should show in a corpus study. To examine the sequentiality of spontaneous conversation or (semi-)planned speech, corpus linguists have often focused on explicit signals of discourse organization. Hence, discourse markers indicating topic shift, turn-taking, closing and so forth have been popular objects of investigation in this area of corpus studies (see e. g. Aijmer 1996; Stenström 1994; Swales/Malczewski 2001). In section 3.2. we will return to discourse markers as signals of interaction.

Metadiscursive elements such as connectors, which are easy to find, have been eagerly studied in professional non-narrative discourse. Many studies have focused on academic writing or talk (cf. e. g. Bondi 2004; Dorgeloh 2004; Mauranen 2001). Connectors have also often been studied in learner language as they are considered a good diagnosis of proficiency level (see e. g. Altenberg/Tapper 1998; Granger/Tyson 1996; Wikborg 1985). The idea of clearly separating metadiscourse from the rest will, however, turn problematic as soon as we study the complex mechanisms of reflexivity in discourse contexts.

Biber et al. (2004) explore the possibilities of automatically identifying what they call 'vocabulary-based discourse units' in academic discourse. From the other studies of discourse organization included in Partington/Morley/Haarman (2004) we can see that while the approach seems promising, the corpora used are still relatively restricted, and predetermined choices, combined with manual analysis, are necessary for the explanation of the findings. This avenue, however, strengthens the belief that the lexicon constitutes a natural meeting point for students of corpora and those of discourse. (For such a view, see article 45; cf. also Thomas/Wilson 1996.)

Discourse linguists know that the beginnings and ends of clauses, sentences and paragraphs or tone units and intonational patterns are important loci of textual information. Thematic and rhematic choices can be studied using corpora; what will be more difficult is to assign them an information value. Early text linguists devised typologies of theme progression, which might be to some extent operationalized to examine variation in text corpora. Miller/Weinert (1998) investigate focus constructions in different languages using a set of corpora. Kaltenböck (2004) uses corpus data to examine discourse functions of extraposition and its non-extrapository counterpart, which turn out to be used for very different purposes. Blanche-Benveniste (1986) explores the effects of context on the use of voice in spoken French.

Explicit markers of foregrounding or backgrounding can to an extent be investigated using corpus data (e. g. tense variation in narrative, some grounding signals; for an in-

ventory of grounding criteria in narrative, see Wårvik 1996). It is, however, very difficult to get at the relative foregrounding and backgrounding of events and ideas using such methods. Similarly, point of view lies beyond the reach of the corpus linguist; superficial analyses of pronominal signals or other markers do not add to our present knowledge of how point of view is indicated in texts and discourses.

It is thus obvious that corpus-based studies of discourse phenomena can primarily help us to come to grips with cohesion, rather than coherence. Also, aspects of a positionally defined thematic structure will be easier to investigate than the intricate interplay of given and new information. Vocabulary-based studies disclose rhetorical units based on structure and content. What this suggests is a focus on textuality rather than the dynamic nature of discourse (see e. g. Partington 1998; Stubbs 1996). Discourse-oriented studies have attempted to make use of standard multipurpose corpora or parts of them, specialized corpora which are typically smaller in size but may consist of entire texts of a particular kind, or corpora compiled for the purposes of verifying some discourse characteristic in a body of data that allows quantitative generalizations to be made.

3.2. Interaction: Discourse-pragmatic perspectives

Coherence, information structuring, (fore)grounding and point of view also contribute to interactive ends, for instance, in terms of constructing connections between discourse, interlocutors and other contextual phenomena or indicating a given point of view through affective foregrounding. Interactive aspects of discourse have, however, more commonly been studied under headings such as affect, evaluation, involvement, engagement, stance, intersubjectivity, politeness, and so forth.

It is predominantly in the area of spoken discourse that corpora have proven very useful as windows on interaction (cf. articles 11 and 47). Paradoxically, this is where corpus compilation is especially cumbersome, and ethical issues tricky to solve (see e. g. Rock 2001); corpora can at best be characterized as small (if transcribed in any systematic and linguistically oriented manner), and questions of authenticity are highlighted in the concrete procedure of transcription, which is essentially an interpretation of data involving application and construction of theory. Spoken corpora are, however, a rich source of insight as they provide linguists with access to the interactional aspects of impromptu speech and planned discourse – public, semi-public/semi-private and private.

For instance, linguistic elements identified as serving pragmatic or discursive functions of various kinds have been successfully studied in corpus data. In addition to exposing important patterns of use and several of the functions that they repeatedly serve in discourse, this strand of research has contributed to giving particles and routine expressions a central status in linguistic enquiry, extending it beyond the ground-breaking work by the early enthusiasts of discourse markers and pragmatic particles. Important corpus-based studies in this area include those originating in the Lund circle directed by Jan Svartvik, who computerized and analysed the LLC (cf. articles 11 and 47; see e. g. Svartvik 1979; Aijmer 1996; Altenberg 1987; Stenström 1994). Aijmer's (1996) studies on conversational routines were based on corpus data. Stenström's and her students' work has shed light on several aspects of spontaneous spoken discourse, including the use of

pragmatic particles (see e. g. Stenström 1994, 2004; Andersen 2001). For a corpus-based investigation of the collocational patterns of various pragmatic particles in spoken Dutch, see e. g. van der Wouden (2002) (for collocation, see section 3.3.).

The study of interactive particles and routine expressions has often started out from individual lexical items whose functions have been described in light of corpus data. Another starting point has been to select a particular discourse-organizing function, such as topic shift or closing, or a communicative function, such as disagreeing or making requests, and then proceed to the description of a set of linguistic items and expressions that serve the selected function in a given corpus or in comparable corpora (cf. e. g. Holmes/Stubbe 2003; Vine 2004 on control acts in workplace interaction). Similar methods have been used to study hedging in corpus data (see e. g. Mauranen 2004). Brinton (1996) and Culpeper/Kytö (1999) investigate discourse markers and hedging in historical data. Dietrich (2003) presents studies of communication in high-risk environments involving time pressure. Several of these studies are based on linguistic analyses of corpus data. Fernandez (1994) investigates the functions of pragmatic particles in discourse construction in a number of typologically different languages. Conversation analysts, too, often start their empirical analyses of spontaneous speech from given linguistic elements (in addition to the classic studies of this orientation, see e. g. Hakulinen 1998 on the Finnish particle *nyt*) or from particular discourse functions (see e. g. Sorjonen 2001 on responding in Finnish). In-depth investigations of conversational data have repeatedly served as an impetus for follow-up studies based on large electronic corpora (cf. e. g. McCarthy 2002 on response tokens).

Wichmann's investigations of intonation (e. g. 2004) show how demanding this kind of corpus-based study is as there is no way of making searches over large quantities of transcribed text and yet hoping to come up with findings such as those originating in her close-up, context-related observations of individual occurrences in the data. Others have studied, for instance, the frequencies and functions of interrupting, where corpus data more readily disclose overlapping speech than the normative phenomenon recognized by interlocutors as interruption in a speech event (see e. g. the discussion in Bilmes 1997). The need for well-formulated research questions at the outset of a corpus-based study is both an asset and a drawback in investigations of this kind where manifestations of interrupting in a given context can turn out to serve subtle functions in the situational and socio-cultural context.

Other discourse-pragmatic issues include the study of intersubjectivity. Du Bois's work on stance alignment, for instance, can be partly carried out using corpus methods while the identification of the phenomena and their discourse functions demand a command of the data that can only be acquired through in-depth manual analysis (cf. e. g. Du Bois 2007; see also Kärkkäinen 2003 on intersubjectivity in stance-taking; both using the SBCSAE: <http://projects.ldc.upenn.edu/SBCSAE/>). For instance, interlocutors tend to repeat words like *too* and *(n)either* as in (1) and (2), which suggests that suitable corpus data can profitably be checked for frequencies, patterns and variations in their indication of stance alignment.

- (1) A: I love the movie
B: Me *too*
- (2) A: I don't like the movie
B: Me *neither* / I don't *either*

Corpus-based investigations of stance, affect or evaluation can be found in Hunston/Thompson (2000). Hunston (2004) approaches evaluation in texts from two perspectives, those of the text and the corpus. Noting that findings from lexically oriented studies of text as structure and process cannot be readily operationalized to serve automatic corpus analysis, she sets out to explore the possibilities and limits of corpus studies of this complex, and typically implicit, area of language use. Starting from given hypotheses concerning some explicit form of evaluation, which have been derived from earlier studies and so-called ‘secondary corpus resources’ (i. e. corpus-based dictionaries, grammars and books of information about particular corpora), she proceeds to the description and interpretation of corpus evidence. Her conclusion is that “reliable automatic identification and quantification can be carried out on only a limited set of realizations of evaluation” (Hunston 2004, 186).

Biber (1988) used corpus data to investigate the relative involvement vs. detachment (combined with integration vs. fragmentation of information, cf. also Chafe 1982), relating the findings to the text categorizations of standard corpora of speech and writing. Petch-Tyson (1998) used corpus data to explore differences concerning this dimension in EFL student writing representing different mother-tongue backgrounds (ICLE, cf. article 15). Virtanen (1998) explored the relatively frequent use of questions in these data. Facchinetto (2003) argues that variation in the use of the modal *may* in corpus data can be related to rhetorical concerns such as audience design. Fløttum (2005) focuses on the linguistic signals of polyphony in the KIAP corpus of academic writing (<http://www.uib.no/kiap/KIAPCorpus.htm>), using the ScaPoLine framework of Nølke/Fløttum/Norén (2004). Finally, many studies of vocabulary frequencies and lexical patterns indicate stylistic tendencies which have a bearing on the interactive dimensions of discourse (see e. g. contributions to Reppen et al. 2002).

Studies of politeness have been problematic in corpus linguistics as their typically cross-linguistic focus has sometimes inadvertently strengthened their deterministic nature as the point of departure has perforce been similarities and differences in form. Socio-cultural aspects of politeness seem beyond the reach of corpus linguistics. Some scholars, however, argue for the possibility of investigating the range of social work that discourses perform by making use of insights from sociolinguistics and discourse studies and tying them to relevant sociolinguistic variables accompanying corpus data that is relatively specialized in terms of context (cf. the discussion of reflexivity in academic talk in Mauranen 2001; see also Yates 2001 for a discussion of corpus-analytic methods in the investigation of Internet interaction). Some of the chapters included in Hickey/Stewart (2005) adopt a corpus-based approach to politeness. Östman (2005) uses his PIA (‘pragmatics as implicit anchoring’) toolkit to investigate coherence, politeness and involvement in corpus data.

The kinds of studies discussed in this section are primarily corpus-based. Corpus study in this area has to work in close cooperation with manual analyses of the data. The research process can thus be characterized as one where the insights of the manual analysis, and previous studies of discourse, suggest search items for the corpus study. Once the corpus study has been initiated or conducted, this phase in the research process, in turn, often raises issues that are again best examined through close scrutiny of texts and discourses.

Sinclair’s (2004) discussions of interactive structures, too, can be traced back to other fields of language study; it remains to be seen, however, whether corpus-driven studies

of raw data will ultimately make us abandon some of our present ideas concerning the interactive aspects of discourse in favour of new ones as more work becomes available in this area. Raw data is, however, something of an illusion in the study of spoken discourse as such corpora have perforce been ‘contaminated’ by an interpretation process during transcription. Though not a solution to the problems of authenticity referred to above, corpora of spoken discourse which provide us with text and sound are obviously to be preferred. Corpus linguists can also gain insight into interaction by analysing corpora consisting of text-based computer-mediated discussions, which open windows on the linguistic construction of virtual identities and communities. Approaches to interaction informed by dialogism will, in general, benefit less from corpus study than the monological frameworks traditionally adopted in corpus linguistics.

3.3. Textual and pragmatic collocation

The existence of very large corpora – and the Internet – have resulted in something of a renaissance in the study of collocation. Texts and discourses provide access to the study of collocation in the very concrete sense of words that like each other’s company. Firth (1968) made a distinction between ‘collocation’ and ‘colligation’ to distinguish between lexical and syntactic patterns of meaning manifest on the syntagmatic axis. Such patterns have been investigated in corpus linguistics from an essentially sentence-grammatical perspective using word and tag sequences (cf. e.g. Kjellmer 1994; Aarts/Granger 1998). Sinclair (1991; 2004, 141) adopts an automatic view of collocation as “the co-occurrence of words with no more than four intervening words”. This view and method allow us to contemplate co-occurrences of morphemes, words and sequences of words in novel ways, starting from what is present in texts and discourses of various kinds and ignoring for a moment the constraints of grammar. Contextual issues come to the fore when we note that collocational patterns vary according to register, discourse type, genre, style and so forth. Indeed, Firth’s early interest in matters of context should encourage us to study collocation in relation to the context-of-situation and cultural context. (Of related interest are articles 43, 45, 18 and 58.)

For the analysis of meanings available to us through the study of collocation it is important to use very large corpora (cf. Sinclair 2004). This kind of analysis has been suggested as a starting point for cognitive text linguistics, too (see de Beaugrande 2004, 24–26). The advantage would seem to be the fact that a given meaning in context – the product – should reflect processes of multiple activations in networks with other meanings. De Beaugrande argues that collocations might indeed constitute the ‘missing link’ between language and text, between the emergent constraints manifest in discourse and the standing constraints specified in language. This would explain why people can say what a word means, out of context (cf. language), yet use and interpret it in a specific sense in a particular discourse context, in a particular meaning foregrounded in networks of multiple activations that help us to highlight other meanings in other contexts.

Collocations are of prime importance to discourse linguists. Tracking collocation across texts yields patterns of use that can be studied in terms of the traditional categories of register, discourse type, genre and style. But new categorizations of texts are also likely to emerge through the study of particular collocations in large bodies of data.

Many corpus linguists employ established broad categories such as speech vs. writing, narrative vs. non-narrative, academic discourse vs. newspaper texts vs. fiction. Categories like these reflect a need to rely on what we already know and perform corpus-based analyses on topics that we have taken for granted but which are suspected to show profiles of use different from what has been expected. Applications of such discourse-oriented studies include new types of discourse-sensitive dictionaries and grammar books, where corpus data help assign discourse domains to the description of linguistic items and constructions.

The possibility of varying the search span is interesting in the study of discourse. Many linguists, however, select a relatively narrow search span to avoid overwhelming problems of insufficient precision. While four intervening words permit observations that are independent of lexico-grammatical structures, widening the span between the node and its potential collocates allows discourse linguists to examine phenomena which operate over sentence boundaries. The study of collocation can, for instance, disclose genre-specific patterns of written argument both within and across sentences, as shown, for example, by Virtanen (2005). Hence, instances where a set of lexical key words systematically co-occurs with, say, a set of argumentative connectors, within or across textual units of various sizes and in a particular order, can be called ‘textual collocation’.

Sinclair (2004, 142) speaks of ‘semantic preference’, i.e. regular co-occurrence of words that share some similarity of meaning, irrespective of whether they constitute Firthian collocations or colligations. It is indeed here that corpus-driven approaches allow for fascinating discoveries to be made in text and discourse linguistics. It is crucial that the search span can be varied and the borderline between lexis and grammar ignored for observations to be made which cast new light on the workings of discourse.

For the analysis of affect Sinclair (2004, 144–146) makes use of the notion of ‘semantic prosody’:

“The semantic prosody of an item is the reason why it is chosen, over and above the semantic preferences that also characterize it. It is not subject to any conventions of linguistic realizations, and so is subject to enormous variation, making it difficult for a human or a computer to find it reliably. It is a subtle element of attitudinal, often pragmatic meaning and there is often no word in the language that can be used as a descriptive label for it”.

(Sinclair 2004, 144–145)

He points out the importance of looking for its presence as well as absence in discourse. As text linguists and rhetoricians know, the absence of, or silence concerning some textual aspect or form has important consequences for discourse meaning.

An attempt to account for ‘pragmatic collocation’ in a systematic way has been made by Östman (2005), who argues that all words can be thought of as having such implicit collocations. Widening the perspective further he notes that

“[i]f some implicit collocations cannot be retrieved nor even discovered except through corpus study, then there are probably a number of other similar pragmatic phenomena that can be fruitfully approached with this combination of methods” [i.e. implicit pragmatics and corpus study].

(Östman 2005, 209)

The study of textual and pragmatic collocation and similar phenomena is an obvious meeting point for corpus linguists and discourse linguists. Studies adopting such promis-

ing perspectives will profit from the relative ease with which we can search for unexpected collocational patterns in large bodies of data, including the Internet, combined with the interest in implicitness that has been prevalent in discourse analysis and pragmatics from the start.

3.4. Variation across texts and discourses

The discussion in sections 3.1. and 3.2. has clearly indicated that there are important links between discourse-organizational and interactive phenomena and categories of text of various kinds. Content analysis of topical lexis and textual collocations has also disclosed variation across texts and discourses. This section focuses on some of the established categorizations of text and discourse in light of corpus studies. As shown in several articles in this volume (e.g., 37, 38, 58), discovery of distributional patterns lies at the very heart of corpus linguistics. Variation is also an area where decisions concerning corpus design come to the fore in a very concrete fashion, namely whether or not they permit comparisons between categories of text and between corpora. The usual text classifications include text/discourse types, genres, registers, styles and modes. Fictionality can also constitute a dividing line between texts.

Two quotations concerning text/discourse type suggest two very different views of the goal of discourse analysis. Referring to the linguistic description of text and discourse structure, Sinclair (2004, 67) claims that “despite theoretical frameworks that are general enough, descriptions are too dependent on the text or discourse type”. In corpus-driven studies the goal is to uncover patterns and connections across huge quantities of raw text which have not already been partly interpreted through the use of predetermined linguistic categories. Singling out and explaining findings, however, seems to profit from prior knowledge of linguistics, even though ad hoc labels are sometimes employed to avoid reference to this body of knowledge. Also, decisions concerning the selection of data, however large the corpus, or the mechanisms of random selection, leave traces in the corpus, which can never represent ‘language as a whole’. Corpus-driven approaches are expected to disclose patterns we are not aware of and ultimately contribute to the emergence of a particular kind of linguistics, i. e. ‘corpus linguistics’.

The second quotation is from Longacre (1996, 7), who argues that “[s]o determinative of detail is the general design of a discourse type that the linguist who ignores discourse typology can only come to grief”. This view would thus rather seem to be favoured by those whose main goal is to use corpora to test hypotheses originating in prior qualitative studies and to provide evidence for claims made on the basis of in-depth studies of small numbers of entire texts in context. Corpus-based studies of this kind also aim at identifying the kinds of data where a given linguistic phenomenon is to be found, as a first step to the close analysis of such data. What is important is to make sure that what we compare are, in fact, comparable. If we compare narrative discourse with non-narrative discourse and state on those grounds that language has changed in particular ways, not noticing that the differences rather relate to discourse type, we are comparing apples with oranges, rather than one kind of apples with another kind.

Postulating collocation as the missing link between language and discourse, de Beaugrande (2000, 2004) suggests that it should be adopted as a tool in cognitive text linguistics.

tics. He notes that the new vistas offered by very large corpora and advanced corpus technology are still overshadowed by thorny problems concerning the choice of data and the methods of interpretation: "we must show how cognition adjusts the strengths of collocations to fine-tuned distinctions among text types, registers, or styles" (de Beaugrande 2004, 29).

Several standard corpora have been planned to allow comparisons between text categories of various kinds. Some corpora manifest categorial consistency to a higher degree than others. Category labels may remind users of, for instance, established genres (e.g. 'newspaper editorials', 'science fiction', 'interview'), content-based register differences (e.g. 'religious writing', 'medical writing') or classifications based on specific variables of the communication situation (e.g. the number, age, gender, level of education or linguistic background of the interlocutors, the channel, the relative privateness vs. publicness of the data, its degree of institutionalisation, spontaneity, etc). Some corpora allow for a selection of data on the basis of discourse-pragmatic and sociolinguistic variables. In addition to multi-category corpora, there are also specialized collections of a given type of data, representing only one genre, register or type of discourse. While writing is and has been the major source of corpus materials, there is a growing variety of speech corpora. Availability and ease of compilation are mainly responsible for the proliferation of text corpora but the channel of communication has also long been considered a major criterion for separating the two kinds of data. In fact, speech and writing have only recently started to appear in one and the same corpus. And multimodal corpora are likely to grow more numerous, not least because of the growing interest in computer-mediated discourse (see articles 12, 17, and 31). In written corpora a major distinction is often made between fiction and non-fiction, which can be further divided into established genres. For discourse analysts the notion of genre is, however, not straightforward.

Text-externally characterized genres are indicated in a growing number of corpora. They are sometimes called text types but more often text/discourse type refers to text-internal categories such as narrative vs. non-narrative. Many studies use very general categories of text (e.g. 'fiction', 'conversation', 'newspaper language'), for instance because it makes the amount of discourse included in each category bigger or because such distinctions are so well established in everyday usage that a discussion of text classification prior to the study need not be conducted. Instead it will be instructive to study variation within such categories in light of corpus findings. In Biber's (1988) study, the categories were the heterogeneous collection of registers and genres included in standard corpora (cf. also article 38). The purpose of the study, to explore variation across speech and writing, yielded as an outcome a new typology of the texts found in those data. This typology was based on a number of dimensions that had been suggested in discourse-linguistic literature. The study also disclosed important internal variation in several text categories. This raises the issue whether discourse linguists should compile corpora of their own where the relevant contextual parameters would be taken more fully into account than can be the case in standard corpora. It is obvious that fiction should manifest exponents of narrative to a higher degree than, say, official documents; however, where the text fragments selected to represent fiction come from has important consequences for the number and kind of indices of narrativity we can hope to find.

Other studies concentrate on adjacent genres found, for instance, in academic discourse (e.g. textbooks, research articles, seminars within a discipline; various kinds of

abstracts within and across disciplines) or computer-mediated discourse (e.g. various categories of home page discourse, chat, blogs and discussion boards). Such corpora allow for investigations of interdiscursive chains and relations, and it will be easier for the analyst to note links between findings and discourse context. Some corpora only include one text category (e.g. the genre of student essay in ICLE, see article 15). They serve to highlight linguistic variation within the chosen category, relating it to the contextual parameters included in the coding of the data.

Apart from Biber, other corpus linguists have investigated speech and writing and the relations between them. For instance, Miller/Weinert (1998) approach the interplay of syntax and discourse using corpora of spontaneous speech in several languages, which they compare to written texts. Kytö (2000) and Culpeper/Kytö (1999, 2000) study speech-related writing in historical data, profiting from insights from the analysis of spoken discourse and orality in contemporary genres. Using data from the Helsinki Corpus Wårvik (2003) points out that the linguistic features of orality and speech vs. literacy and writing figuring in the literature can more readily be related to genres than reception format. She suggests that “some of the features of orality that are still characteristic of certain genres have their origin in the oral roots of the genre” (Wårvik 2003, 45) and proposes that such features can, in due course, disappear, change or be maintained as conscious stylistic choices rather than genuine orality. It should be noted that the Helsinki Corpus was one of the pioneers to include a coding of text categories.

Large numbers of literary texts or excerpts from them are available for corpus linguists. Early studies used these data for dialogue analysis, in the absence of corpora consisting of impromptu speech. Today fictional prose is often used for the analysis of translated discourse (see section 4), simply because translations in many countries predominantly consist of fiction. Corpus studies of literary discourse can be found in, for instance, the journal *Literary & Linguistic Computing*. The new French journal *Corpus* devotes its second issue to intertextual distance in corpora of various kinds, some of which consist of literary works (Luong 2003). Studies of linguistic stylistics contribute to our understanding of how discourse works. When corpus linguists make statements about fiction, it is, however, usually for comparisons with non-fiction, such as academic discourse or newspaper discourse (cf. e.g. Short/Semino/Culpeper 1996). Conversational data and fictional dialogue have also been compared and contrasted with one another.

Style is a notion that has been used in different senses; many involve an idea of choice. If at all, the notion has entered the corpus-linguistic scene at the stages of labelling or explaining variation, rather than being in itself the focus of the study. In discourse analysis style can profitably be used to indicate variation within textual categories such as discourse types or genres (see e.g. Enkvist 1987). In this sense, it comes very close to what Verschueren (1995, 14) calls ‘variability’. Variability as used in pragmatics is a processual notion whereas style can be interpreted as the outcome or product of such processes.

Variation is an overarching theme in corpus studies. The main bulk of discourse-oriented studies of corpora have, in one way or another, contributed to our knowledge of variation across texts and discourses, or variation through time. In contrast to variation, variability does not lie in the nature of corpus studies, which rather seek to uncover phenomena that are common (or uncommon) in (particular kinds of) discourse.

It is at the level of large categorizations of text that corpora have profitably been used to disclose important differences. When such categories have been defined using

non-linguistic criteria, the analyst has been free to draw conclusions concerning their linguistic characteristics. When we start from established categories, a great deal of variation within a category raises the issue of ‘hybridization’ (cf. Fairclough 1992; see also Bhatia 2005). Texts manifest combinations and co-occurrences of linguistic elements of different categories such as discourse types and genres. Further, characteristics of other categories of text can be appropriated and different categories can merge or hybridize. In contrast, if the starting point is not a set of established categories, as in corpus-driven approaches, the findings may suggest another kind of categorization of the data, a new typology. Hybridization is then not an issue as ‘uncontaminated’ text cannot, by definition, be hybridized; it is only at the subsequent stages of the analysis of the data-driven emergent typology that the corpus linguist adopting this approach will have grounds for observing such processes. Some rather general, basic distinctions still tend to show across data of many kinds: not surprisingly, it is easy to pinpoint the distinction between narrative and non-narrative in many discourse-oriented corpus studies.

A final point concerns, in fact, hybrids. Many kinds of text categories but notably genres change through time. Genres and discourse types do not constitute static sets of categories of equal value. Such abstractions emerge from discourse because they are in many ways helpful for interlocutors engaged and reengaged in various discourse practices. They function as cognitive heuristics and influence, maintain or alter social practices in complex ways, relating them to speech communities, discourse communities and communities of practice. To be able to function in that way, such categories need to be prototypical, context-sensitive and tolerant of internal variation. But these are also important factors facilitating generic change. Genre dynamism is difficult to come to grips with in corpus studies. However, historical corpus linguists (cf. e. g. Dorgeloh 2005; Taavitsainen 1999) have traced patterns of conventionalization which help us understand the development and internal variation of given genres and registers. Corpora of computer-mediated discourse constitute another source of insight into genre formation. Specialized corpora such as MICASE can shed light on the interplay of related genres within a given social and professional context. Others, such as ICLE, manifest variation and variability within a single genre. However, a pertinent question to ask is whether textual hybrids – the usual kinds of texts and discourses – are indeed the default. Perhaps they are simply texts, without any ‘hybrid identity’ that our categorizations force on them. Both discourse analysts and corpus linguists do well in giving their decisions concerning text categorizations the attention that they need, questioning the categorizations they use while simultaneously viewing their findings in the light of potential traditional and novel categories of discourse that may be relevant in the study.

3.5. Shortcuts between dimensions of discourse

One of the main advantages of using corpora in the study of discourse is the many shortcuts between the different dimensions of discourse offered to the analyst. Combining methods from corpus linguistics and discourse analysis and devoting attention to what is possible should smooth the way for a happy relationship between the two areas of study. Exposure through concordance data to the discourse dimensions of structure, content and interaction all at the same time makes it possible to blur borderlines such as

those between the informational, expressive and phatic functions of discourse, between transmission and transaction in discourse practices, or between discourse and metadiscourse. Such shortcuts may thus have the advantage of presenting discourse data in ways that allow us to rethink concepts and models. Eclectic approaches are, not surprisingly, found in both corpus linguistics and discourse linguistics. The cognitive dimension of discourse can be highlighted through exposure to the structural, content-oriented and interactional aspects of the texts included in a corpus. The cognitive and social dimensions can meet in discourse-oriented corpus studies as we can try to contemplate large bodies of data in terms of distributed cognition. While it will not be possible to examine the dynamic aspects of the ‘discursive struggle’ in which individuals and communities of people engage and reengage throughout their lives, care can be taken to turn the often monologicistic and static manner of studying discourse phenomena in corpus data, into a more dynamic approach where dialogistic aspects are taken into account as far as possible. Faced with the products, it will also be important for analysts to recall that the discourse processes leading to them have not necessarily been rational or intentional in character. Being exposed to text in a non-orthodox way has already suggested new ways in which we can go about modelling discourse phenomena; there is reason to believe that we are only at the very start of a harmonious relationship between corpus linguists and discourse linguists.

4. Discourse-oriented applications of corpora

Corpus-linguistic methodology serves to foreground the exploratory nature of the study of language and it is therefore particularly well suited to a student-driven, problem-based pedagogy in linguistics and the study of languages (cf. article 7). Corpora can be used similarly in translation studies and translator training (cf. article 55). Language professionals such as teachers and translators profit from existing corpora in their work.

Further, linguists, social scientists and students of cognition sharing an interest in discourse profit from corpora as they set out to formulate and test hypotheses. Psycho-linguists and cognitive linguists have noticed the advantages of using corpus data instead of, or in addition to, experimentation when they wish to study discourse phenomena. Computer scientists, too, find corpora helpful (cf. articles 35, 56). Hunston (2002) offers an overview of the impact of corpora on applied linguistics. While applied linguists have long struggled with the problems connected to the definition and status of the ideal native speaker, characteristics of potential language-independent nonnative interlanguage, transfer and individual learner differences, discourse-oriented corpora of learner language of various kinds, or international use of English in academic and other professional settings where native and nonnative speakers regularly communicate with one another, promise to offer new grounds for such investigations and discussions (cf. e.g. Granger 1998; Swales/Malczewski 2001; cf. article 15). Also, corpora can be used not only to compare translations to and from a language or a set of languages of various kinds (cf. article 16 on parallel corpora) but also to compare the discourse of translated texts with that produced in the language(s) of the country (cf. e.g. the Savonlinna project on Finnish and translated Finnish, directed by Anna Mauranen). In this respect, countries such as Finland or Sweden where the majority of published books are translations

offer a striking contrast to English-language contexts where this is not so. Related to translational aspects of discourse are characteristics of special, professional or institutional discourse that people switch to and from in their daily lives. Discourse analysts are interested in the social consequences of such genres, motivations behind conventionalization and generic change, and so forth. Register-specific interdiscursive corpora which permit longitudinal studies would seem feasible and helpful for such projects.

5. Methodological issues

A striking difference between the two areas of corpus linguistics and discourse analysis is found in the form and content, and indeed, presence or absence, of an explicit account of the methods and materials used in a given study. While it is normally possible to find such statements in both kinds of study, corpus linguists tend to include an explicit and relatively detailed section on the materials, search procedures and other methods used. In discourse-analytic publications methods and materials have sometimes been left implicit and where they appear, their form and content vary a great deal. There are obvious reasons for this difference. Similarly, it is possible, and indeed usually necessary, to formulate research questions in an explicit manner in corpus studies. In discourse analysis, the point of departure is an acceptance of at least some degree of causal indeterminacy as concerns explanation, which affects the construction of the entire research article. While acutely aware of the impossibility of separating themselves from their data as a natural scientist would, both corpus linguists and discourse linguists aim at restricting subjectivity but they do this in different ways, which shows in their writing. Unlike discourse analysts, however, corpus linguists may be occasionally tempted to trust the myth of objectivity created by the quantitative results indicating distributional patterns in large bodies of data.

Corpus-based studies typically start from a set of formal exponents of a given discourse phenomenon, singled out by previous scholarship, only to note that the range of search items needs to be expanded, or that there is a central issue which should be studied in more detail. Here the focus can be on the role of a given search item as a pointer to a particular discourse phenomenon or the variety and patterns of its use in a large corpus or several comparable corpora. The most and least common occurrences attract the interest of the corpus linguist while phenomena whose frequency of occurrence is in no way conspicuous can raise fundamental questions in the mind of the discourse linguist. Further, the finding that something is infrequent or missing in particular kinds of data can be of prime importance to our understanding of discourse, and an infrequent item may well lead us to discover patterns in discourse of a kind that is only marginally represented in the corpora we are studying.

Both corpus linguists and discourse linguists investigate discourse in some form, irrespective of differences in the size and kind of data needed. Both can, indeed, start out from predetermined linguistic elements and subsequently expand or focus their study in light of findings. This is, however, much easier in corpus studies when new searches can be done quickly. Discourse analysts extend themselves to examine their data from various perspectives and in the process they gain an understanding of the data which is different from the one corpus linguists acquire. It is this combination of insights that gives corpus-oriented discourse studies their added value.

There are ethical issues that are common to both areas. Data collection involves a balance between, on the one hand, authenticity, naturalness and representativeness, and on the other hand, ethics, metalinguistic awareness and availability. In this respect corpus design can profit from the experience of linguists of different orientations.

One of the issues often dealt with is the representativeness of the data. Discourse linguists know that they can only investigate relatively few texts, that one single text can provide them with more information than can be exposed through their analyses, that contexts which are not necessarily available to them are an inseparable aspect of any piece of discourse that they are analysing, that they construct contexts through their own investigations, and any explanations they are offering are likely to be causally indeterminate and subject to change when new analyses and other texts and discourses are added to their study. In this light, it is obvious to discourse linguists that if they know that their data are representative of what they want to study, they do not need them, whereas in the usual instance, they cannot be sure that their materials are representative enough to warrant generalization. Corpus linguists, too, know that their data can never be representative of language as a whole but they put a good deal of effort into making sure that it is representative of some well-thought-out part of it. Such representativeness of even very large corpora will, however, always be more problematic in light of the goals of discourse analysis.

Corpus linguists worry about corpus design because their purpose is to quantify data. What is important are the frequencies of occurrence indicating distributional patterns. For the discourse analyst, frequency need not be of primary concern. While it is sometimes of great value to know whether something is common in some contexts, it is as often important to focus on unique aspects of particular kinds of discourse which may provide the analyst with a cue to understanding how some discourse phenomenon works. Individual texts are here an important source of information to the essentially qualitative analyses of the ways in which linguistic signals function in linking texts to contexts and contexts to texts, to merge the two into discourse. Also, linguists interested in the study of discourse and pragmatics are intrigued by textual silence and subtle means of indicating organization and interaction in discourse. They are hoping for new insights from corpora. As shown in article 31, many corpora that represent discourse and have complex annotations seem to be better served by the use of multi-layer systems, rather than flat-file systems in their corpus architecture.

In corpus linguistics description of data is a commonly accepted goal in itself. In discourse analysis describing would rather appear to be a trivial process of paraphrasing or interpreting a given text, which cannot be the goal of the study. Linguistic explanation is, however, a difficult task in both areas as it is not usually evident what actually causes the phenomena that one has observed in the data. Still, in discourse analysis it is mandatory to come to some non-trivial and non-speculative conclusions about the tendencies one has observed, the inferences needed to understand particular texts and discourses, the nature of the discursive struggle and the work that the discourse does for the interlocutors engaged in it in a given situational and socio-cultural context, and so forth. In corpus-based studies the relatively strict formulation of hypotheses, research questions, methods and materials facilitate the analyst's task of explanation. In corpus-driven studies, as in conversation analysis, reliance on the text may lead to unforeseen findings but it may also force the analyst to ad hoc categorization of the data which might profitably be linked to related insights from previous scholarship.

6. Conclusion: New vistas for discourse analysis?

We know today that the mere transformation of authentic discourse into a computerized corpus results in the loss of the dynamism inherent in the context from which it has been extracted. Even written texts thus recontextualized lose their authentic nature. Yet, we are also witnessing something new which seems to be the outcome of such procedures: it is hoped that corpus studies will help us to uncover new aspects of the five dimensions of discourse dealt with in this article. Corpora promise to take us beyond discoveries of what we already know, at least partly, to reconsideration of the established ‘facts’ in today’s discourse linguistics and our tacit assumptions that come with the package.

Two of the four areas where corpora and discourse analysis should be able to meet have proven to be problematic: it is difficult to come to grips with the dynamism of discourse organization and interaction using corpus-based or corpus-driven approaches. In contrast, another two areas are full of promise: variation across texts and discourses, and textual and pragmatic collocation. Both of these profit from the large-scale studies that are possible using corpora. Variation suggests established and emergent conventions. Collocation gives access to textual and pragmatic phenomena without prior decisions concerning lexis and grammar, or sentence boundaries.

Corpora and discourse analysis have a troubled relationship because of their fundamental ontological and epistemological differences. They set out to describe and explain very different realities, and even though both are concerned with discourse, there are principled differences in what constitutes evidence in each, what kinds of claims can be made, and so forth. While both rely on discourse data, they also make use of introspection but in different ways and for different purposes. The status of introspection is perhaps more explicit in discourse analysis, where it is important to clearly signal a separation of speculative elements from findings in the construction of the argument. And finally, a potential meeting point: both can manifest eclectic and interdisciplinary or cross-disciplinary tendencies. At one and the same time corpora serve to disclose aspects of discourse of very different nature while discourse analysts hope to investigate the dynamism of the interplay of different factors influencing the phenomena that they are interested in and the functions that discourses serve for interlocutors. Despite differences, then, the relationship of corpora and discourse analysis is a steady one, in which both parties can complement one another in constructive ways.

7. Literature

- Aarts, Jan/Granger, Sylviane (1998), Tag Sequences in Learner Corpora: A Key to Interlanguage Grammar and Discourse. In: Granger 1998, 132–141.
- Aijmer, Karin (1996), *Conversational Routines in English: Convention and Creativity*. London/New York: Longman.
- Aijmer, Karin/Stenström, Anna-Brita (eds.) (2004), *Discourse Patterns in Spoken and Written Corpora*. (Pragmatics & Beyond New Series 120.) Amsterdam/Philadelphia: Benjamins.
- Altenberg, Bengt (1987), Causal Ordering Strategies in English Conversation. In: Monaghan, J. (ed.), *Grammar in the Construction of Texts*. London: Frances Pinter, 50–64.
- Altenberg, Bengt/Tapper, Marie (1998), The Use of Adverbial Connectors in Advanced Swedish Learners’ Written English. In: Granger 1998, 80–93.

- Andersen, Gisle (2001), *Pragmatic Markers and Sociolinguistic Variation: A Relevance-theoretical Approach to the Language of Adolescents*. Amsterdam/Philadelphia: Benjamins.
- de Beaugrande, Robert (2000), Text Linguistics at the Millennium: Corpus Data and Missing Links. In: *Text* 20(2), 153–195.
- de Beaugrande, Robert (2004), Language, Discourse, and Cognition: Retrospects and Prospects. In: Virtanen 2004b, 17–31.
- Bhatia, Vijay K. (2005), Generic Patterns in Promotional Discourse. In: Halmari/Virtanen 2005, 213–225.
- Biber, Douglas (1988), *Variation across Speech and Writing*. Cambridge etc.: Cambridge University Press.
- Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, Douglas/Csomay, Eniko/Jones, James K./Keck, Casey (2004), Vocabulary-based Discourse Units in University Registers. In: Partington/Morley/Haarman 2004, 23–40.
- Bilmes, Jack (1997), Being Interrupted. In: *Language in Society* 26, 507–531.
- Blanche-Benveniste, Claire (1986), La notion de contexte dans l'analyse syntaxique des productions orales: Exemples des verbes actifs et passifs. In: *Recherches sur le français parlé* 8, 39–57.
- Bondi, Marina (2004), The Discourse Function of Contrastive Connectors in Academic Abstracts. In: Aijmer/Stenström 2004, 139–156.
- Brinton, Laurel J. (1996), *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin/New York: Mouton de Gruyter.
- Brodda, Benny (1991), Doing Corpus Work with PC Beta; or, How to be your Own Computational Linguist. In: Johansson, Stig/Stenström, Anna-Brita (eds.), *English Computer Corpora: Selected Papers and Research Guide*. (Topics in English Linguistics 3.) Berlin/New York: Mouton de Gruyter, 259–282.
- Chafe, Wallace (1982), Integration and Involvement in Speaking, Writing, and Oral Literature. In: Tannen, Deborah (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex, 35–54.
- Culpeper, Jonathan/Kytö, Merja (1999), Modifying Pragmatic Force: Hedges in Early Modern English Dialogues. In: Jucker et al. 1999, 293–312.
- Culpeper, Jonathan/Kytö, Merja (2000), Data in Historical Pragmatics: Spoken Interaction (Re)cast as Writing. In: *Journal of Historical Pragmatics* 1(2), 175–199.
- Dietrich, Rainer (ed.) (2003), *Communication in High Risk Environments. Linguistische Berichte* 12, special issue.
- Dorgeloh, Heidrun (2004) Conjunction in Sentence and Discourse: Sentence-initial *And* and Discourse Structure. In: *Journal of Pragmatics* 36, 1761–1779.
- Dorgeloh, Heidrun (2005), Patterns of Agentivity and Narrativity in Early Science Discourse. In: Skaffari, Janne/Peikola, Matti/Carroll, Ruth/Hiltunen, Risto/Wårvik, Brita (eds.), *Opening Windows on Texts and Discourses of the Past*. Amsterdam/Philadelphia: Benjamins, 83–94.
- Du Bois, John W. (2007), The Stance Triangle. In: Englebretson, Robert (ed.), *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam/Philadelphia: Benjamins, 139–182.
- Enkvist, Nils Erik (1984), Contrastive Linguistics and Text Linguistics. In: Fisiak, J. (ed.), *Contrastive Linguistics, Prospects and Problems*. Berlin: Mouton de Gruyter, 45–67.
- Enkvist, Nils Erik (1987), What has Discourse Linguistics Done to Stylistics? In: Battestine, S. P. X. (ed.), *Georgetown University Round Table on Languages and Linguistics 1986*. Washington, D.C.: Georgetown University Press, 19–36.
- Enkvist, Nils Erik (1991), Discourse Strategies and Discourse Types. In: Ventola, Eija (ed.), *Functional and Systemic Linguistics: Approaches and Uses*. Berlin/New York: Mouton de Gruyter, 3–22.
- Facchinetti, Roberta (2003), Pragmatic and Sociological Constraints on the Functions of *may* in Contemporary British English. In: Facchinetti, Roberta/Krug, Manfred/Palmer, Frank (eds.),

- Modality in Contemporary English.* (Topics in English Linguistics 44.) Berlin/New York: Mouton de Gruyter, 301–327.
- Fairclough, Norman (1992), *Discourse and Social Change*. Cambridge: Polity Press.
- Fernandez, M. M. Jocelyne (1994), *Les particules énonciatives dans la construction du discours*. Paris: Presses Universitaires de France.
- Firth, John Rupert (1968), *Selected Papers 1952–1959*. Edited by F. R. Palmer. London: Longman.
- Fløttum, Kjersti (2005), The Self and the Others: Polyphonic Visibility in Research Articles. In: *International Journal of Applied Linguistics* 15, 29–44.
- Gill, Martin (2008), Authenticity. In: Verschueren/Östman 2005–.
- Givón, Talmy (ed.) (1983), *Topic Continuity in Discourse: A Quantitative Cross-language Study*. (Typological Studies in Language 3.) Amsterdam/Philadelphia: Benjamins.
- Granger, Sylviane (ed.) (1998), *Learner English on Computer*. London/New York: Longman.
- Granger, Sylviane/Tyson, Stephanie (1996), Connector Usage in the English Essay Writing of Native and Non-native EFL Speakers of English. In: *World Englishes* 15, 79–89.
- Hakulinen, Auli (1998), The Use of Finnish *nyt* as a Discourse Particle. In: Jucker, Andreas H./Ziv, Yael (eds.), *Discourse Markers: Descriptions and Theory*. (Pragmatics & Beyond New Series 57.) Amsterdam/Philadelphia: Benjamins, 83–96.
- Halmari, Helena/Virtanen, Tuija (eds.) (2005), *Persuasion across Genres: A Linguistic Approach*. (Pragmatics & Beyond New Series 130.) Amsterdam/Philadelphia: Benjamins.
- Hickey, Leo/Stewart, Miranda (eds.) (2005), *Politeness in Europe*. (Multilingual Matters 127.) Clevedon/Buffalo/Toronto: Multilingual Matters.
- Holmes, Janet/Stubbe, Maria (2003), *Power and Politeness in the Workplace*. London/Harlow: Longman.
- Hunston, Susan (2002), *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Hunston, Susan (2004), Counting the Uncountable: Problems of Identifying Evaluation in a Text and in a Corpus. In: Partington et al. 2004, 157–188.
- Hunston, Susan/Thompson, Geoff (eds.) (2000), *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.
- Jucker, Andreas H./Fritz, Gerd/Lebsanft, Franz (eds.) (1999), *Historical Dialogue Analysis*. (Pragmatics & Beyond New Series 66.) Amsterdam/Philadelphia: Benjamins.
- Kaltenböck, Gunther (2004), Using Non-extraposition in Spoken and Written Texts: A Functional Perspective. In: Aijmer/Stenström 2004, 219–242.
- Kärkkäinen, Elise (2003), *Epistemic Stance in English Conversation: A Description of its Interactional Functions, with a Focus on "I think"*. Amsterdam/Philadelphia: Benjamins.
- Kjellmer, Göran (1994), *A Dictionary of English Collocations*. Oxford: Clarendon Press.
- Kytö, Merja (2000), Robert Keayne's Notebooks: A Verbatim Record of Spoken English in Early Boston? In: Herring, Susan C./Van Reenen, Pieter/Schøsler, Lene (eds.), *Textual Parameters in Older Languages*. Amsterdam/Philadelphia: Benjamins, 273–308.
- Longacre, Robert E. (1996), *The Grammar of Discourse*. 2nd ed. New York/London: Plenum Press.
- Luong, Xuan (ed.) (2003), *La distance intertextuelle. Corpus 2*, special issue.
- Mair, Christian (1990), *Infinitival Complement Clauses in English: A Study of Syntax in Discourse*. Cambridge: Cambridge University Press.
- Mauranen, Anna (2001), Reflexive Academic Talk: Observations from MICASE. In: Simpson/Swales 2001, 165–178.
- Mauranen, Anna (2004), They're a little bit different ... Observations on Hedges in Academic Talk. In: Aijmer/Stenström 2004, 173–197.
- McCarthy, Michael (2002), Good Listenership Made Plain: British and American Non-minimal Response Tokens in Everyday Conversation. In: Reppen et al. 2002, 49–71.
- Miller, Jim/Weinert, Regina (1998), *Spontaneous Spoken Language: Syntax and Discourse*. Oxford: Clarendon Press.
- Nölke, Henning/Fløttum, Kjersti/Norén, Coco (2004), *Scapoline: La théorie scandinave de la polyphonie linguistique*. Paris: Kimé.

- Ochs, Elinor (1979), Transcription as Theory. In: Ochs, Elinor/Schiffrin, Bambi B. (eds.), *Developmental Pragmatics*. New York: Academic Press, 43–72.
- Östman, Jan-Ola (2005), Persuasion as Implicit Anchoring: The Case of Collocations. In: Halmari/Virtanen 2005, 183–212.
- Pander Maat, Henk/Sanders, Ted (2001), Subjectivity in Causal Connectives: An Empirical Study of Language in Use. In: *Cognitive Linguistics* 12(3), 247–273.
- Partington, Alan (1998), *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. (Studies in Corpus Linguistics 2.) Amsterdam/Philadelphia: Benjamins.
- Partington, Alan/Morley, John/Haarman, Louann (eds.) (2004), *Corpora and Discourse*. (Linguistic Insights. Studies in Language and Communication 9.) Bern etc.: Peter Lang.
- Petch-Tyson, Stephanie (1998), Writer/Reader Visibility in EFL Written Discourse. In: Granger 1998, 107–118.
- Reppen, Randi/Fitzmaurice, Susan M./Biber, Douglas (eds.) (2002), *Using Corpora to Explore Linguistic Variation*. (Studies in Corpus Linguistics 9.) Amsterdam/Philadelphia: Benjamins.
- Rock, Frances (2001), Policy and Practice in the Anonymisation of Linguistic Data. In: *International Journal of Corpus Linguistics* 6(1), 1–26.
- Short, Mick/Semino, Elena /Culpeper, Jonathan (1996), Using a Corpus for Stylistics Research: Speech and Thought Representation. In: Thomas/Short 1996, 110–131.
- Simpson, Rita C./Swales, John M. (eds.) (2001), *Corpus Linguistics in North America*. Ann Arbor: The University of Michigan Press.
- Sinclair, John (1991), *Corpus, Concordance, Collocation*. Oxford/New York: Oxford University Press.
- Sinclair, John (2004), *Trust the Text: Language, Corpus and Discourse*. London/New York: Routledge.
- Sorjonen, Marja-Leena (2001), *Responding in Conversation: The Study of Response Particles in Finnish*. (Pragmatics & Beyond New Series 70.) Amsterdam/Philadelphia: Benjamins.
- Stenström, Anna-Brita (1994), *An Introduction to Spoken Interaction*. London/New York: Longman.
- Stenström, Anna-Brita (2004), What is Going on between Speakers. In: Partington/Morley/Haarman 2004, 259–283.
- Stubbs, Michael (1996), *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.
- Svartvik, Jan (1979), Well in Conversation. In: Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (eds.), *Studies in English Linguistics for Randolph Quirk*. London/New York: Longman, 167–177.
- Swales, John M./Malczewski, Bonnie (2001), Discourse Management and New-episode Flags in MICASE. In: Simpson/Swales 2001, 145–164.
- Taavitsainen, Irma (1999), Dialogues in Late Medieval and Early Modern English Medical Writing. In: Jucker et al. 1999, 243–268.
- Thomas, Jenny/Short, Mick (eds.) (1996), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London/New York: Longman.
- Thomas, Jenny/Wilson, Andrew (1996), Methodologies for Studying a Corpus of Doctor-Patient Interaction. In: Thomas/Short 1996, 92–109.
- Tognini Bonelli, Elena (2001), *Corpus Linguistics at Work*. (Studies in Corpus Linguistics 6.) Amsterdam/Philadelphia: John Benjamins.
- Verschueren, Jef (1995), The Pragmatic Perspective. In: Verschueren, Jef/Östman, Jan-Ola/Blommaert, Jan (eds.), *Handbook of Pragmatics: Manual*. Amsterdam/Philadelphia: Benjamins, 1–19.
- Verschueren, Jef/Östman, Jan-Ola (eds.) (2005-), *Handbook of Pragmatics Online*. Amsterdam/Philadelphia: Benjamins. <<http://www.benjamins.com/online/hop>>.
- Vine, Bernadette (2004), *Getting Things Done at Work: The Discourse of Power in Workplace Interaction*. Amsterdam/Philadelphia: Benjamins.
- Virtanen, Tuija (1992), *Discourse Functions of Adverbial Placement in English*. Åbo: Åbo Akademi University Press.

- Virtanen, Tuija (1997), Text Structure. In: Verschueren, Jef/Östman, Jan-Ola/Blommaert, Jan/Bulcaen, Chris (eds.), *Handbook of Pragmatics*. 3rd installment. Amsterdam/Philadelphia: Benjamins. [Reprinted in Verschueren/Östman 2005.]
- Virtanen, Tuija (1998), Direct Questions in Argumentative Student Writing. In: Granger 1998, 94–106.
- Virtanen, Tuija (2004a), Point of Departure: Cognitive Aspects of Sentence-initial Adverbials. In: Virtanen 2004b, 79–97.
- Virtanen, Tuija (ed.) (2004b), *Approaches to Cognition through Text and Discourse*. (Trends in Linguistics. Studies and Monographs 147.) Berlin/New York: Mouton de Gruyter.
- Virtanen, Tuija (2005), “Polls and surveys show”: Public Opinion as a Persuasive Device in Editorial Discourse. In: Halmari/Virtanen 2005, 153–180.
- Wårvik, Brita (1996), Grounding. In: Verschueren, Jef/Östman, Jan-Ola/Blommaert, Jan/Bulcaen, Chris (eds.), *Handbook of Pragmatics*. 2nd installment. Amsterdam/Philadelphia: Benjamins. [Reprinted in Verschueren/Östman 2005.]
- Wårvik, Brita (2003), “When you read or hear this story read”: Issues of Orality and Literacy in Old English Texts. In: Hiltunen, Risto/Skaffari, Janne (eds.), *Discourse Perspectives on English: Medieval to Modern*. (Pragmatics & Beyond New Series 119.) Amsterdam/Philadelphia: Benjamins, 13–55.
- Wichmann, Anne (2004), The Intonation of *please*-requests: A Corpus-based Study. In: *Journal of Pragmatics* 36, 1521–1549.
- Wikborg, Eleanor (1985), Types of Coherence Breaks in University Student Writing. In: Enkvist, Nils Erik (ed.), *Coherence and Composition: A Symposium*. Åbo: Publications of the Research Institute of Åbo Akademi Foundation 101, 93–133.
- van der Wouden, Ton (2002), Particle Research Meets Corpus Linguistics: On the Collocational Behavior of Particles. In: *Belgian Journal of Linguistics* 16(1), 151–174.
- Yates, Simeon J. (2001), Researching Internet Interaction: Sociolinguistics and Corpus Analysis. In: Wetherell, Margaret/Taylor, Stephanie/Yates, Simeon J. (eds.), *Discourse as Data: A Guide for Analysis*. Milton Keynes: The Open University, 93–146.

Tuija Virtanen, Åbo (Finland)

50. Corpus linguistics and stylometry

1. Introduction
2. Lexical discriminators
3. Statistical measures
4. Automatic selection of discriminators
5. Conclusion
6. Acknowledgments
7. Literature

1. Introduction

Computational Stylometry is an attempt to capture the essence of the **style** of a particular author by reference to a variety of quantitative criteria, usually lexical, called **discriminators**. The most common application of computational stylometry has been author identi-

- Virtanen, Tuija (1997), Text Structure. In: Verschueren, Jef/Östman, Jan-Ola/Blommaert, Jan/Bulcaen, Chris (eds.), *Handbook of Pragmatics*. 3rd installment. Amsterdam/Philadelphia: Benjamins. [Reprinted in Verschueren/Östman 2005.]
- Virtanen, Tuija (1998), Direct Questions in Argumentative Student Writing. In: Granger 1998, 94–106.
- Virtanen, Tuija (2004a), Point of Departure: Cognitive Aspects of Sentence-initial Adverbials. In: Virtanen 2004b, 79–97.
- Virtanen, Tuija (ed.) (2004b), *Approaches to Cognition through Text and Discourse*. (Trends in Linguistics. Studies and Monographs 147.) Berlin/New York: Mouton de Gruyter.
- Virtanen, Tuija (2005), “Polls and surveys show”: Public Opinion as a Persuasive Device in Editorial Discourse. In: Halmari/Virtanen 2005, 153–180.
- Wårvik, Brita (1996), Grounding. In: Verschueren, Jef/Östman, Jan-Ola/Blommaert, Jan/Bulcaen, Chris (eds.), *Handbook of Pragmatics*. 2nd installment. Amsterdam/Philadelphia: Benjamins. [Reprinted in Verschueren/Östman 2005.]
- Wårvik, Brita (2003), “When you read or hear this story read”: Issues of Orality and Literacy in Old English Texts. In: Hiltunen, Risto/Skaffari, Janne (eds.), *Discourse Perspectives on English: Medieval to Modern*. (Pragmatics & Beyond New Series 119.) Amsterdam/Philadelphia: Benjamins, 13–55.
- Wichmann, Anne (2004), The Intonation of *please*-requests: A Corpus-based Study. In: *Journal of Pragmatics* 36, 1521–1549.
- Wikborg, Eleanor (1985), Types of Coherence Breaks in University Student Writing. In: Enkvist, Nils Erik (ed.), *Coherence and Composition: A Symposium*. Åbo: Publications of the Research Institute of Åbo Akademi Foundation 101, 93–133.
- van der Wouden, Ton (2002), Particle Research Meets Corpus Linguistics: On the Collocational Behavior of Particles. In: *Belgian Journal of Linguistics* 16(1), 151–174.
- Yates, Simeon J. (2001), Researching Internet Interaction: Sociolinguistics and Corpus Analysis. In: Wetherell, Margaret/Taylor, Stephanie/Yates, Simeon J. (eds.), *Discourse as Data: A Guide for Analysis*. Milton Keynes: The Open University, 93–146.

Tuija Virtanen, Åbo (Finland)

50. Corpus linguistics and stylometry

1. Introduction
2. Lexical discriminators
3. Statistical measures
4. Automatic selection of discriminators
5. Conclusion
6. Acknowledgments
7. Literature

1. Introduction

Computational Stylometry is an attempt to capture the essence of the **style** of a particular author by reference to a variety of quantitative criteria, usually lexical, called **discriminators**. The most common application of computational stylometry has been author identi-

fication in cases of disputed authorship. The underlying assumption in studies of authorship is that although authors may consciously vary their own style, there will always be the subconscious consistent use of certain stylistic features throughout their work (Holmes 1997). However, there is no clear and indisputable evidence that such features exist (McEnery/Oakes 2000). Discriminators used in quantitative stylometric studies must occur frequently in the texts, and represent traits that can be expressed numerically. The identification of features such as Semitisms in the New Testament is a subjective affair, and hence these would not be good discriminators (Morton 1978). Discriminators which have been used in studies of stylometry include:

1. **Word and sentence length.** In perhaps the earliest study, in 1887, Mendenhall showed that the modal word length was 2 in the writings of J. S. Mill, but 3 for *Oliver Twist* (Kenny 1982). These measures tend to work less well for studies of disputed authorship, since they are more under the conscious control of the author. They work better as discriminators of genre or register, as a comparison of the texts used in different newspapers will show.
2. **Vocabulary studies:** The choice and frequency of words, and measures of vocabulary richness such as Yule's K measure (Yule 1944, 57). The overwhelming majority of corpus-based stylometry studies use lexical (single word) discriminators, so we will look at these more closely in section 2.
3. **Fragments of words**, such as bigrams, or pairs of adjacent characters (Kjell 1994). There is a whole literature concerned with discovering automatically which word fragments might act as the best discriminators, and we will explore this in section 4.
4. **Words commencing with an initial vowel** (Hilton/Holmes 1993).
5. **Collocations of words.** (Hoover 2002, 2003) used pairs of words in his studies of disputed authorship, both contiguous sequences of two words, and collocations, which he defined as "any two words that occur repeatedly within a specified distance from each other, counted in words". In some cases he found that word pairs gave better results than single words, but in others not.
6. **Positions of words within sentences.** In Michaelson/Morton's chronology of Isocrates (1976), the frequency of use of *gar* (meaning *for*) as the second or third word showed a negative correlation with time. Milić (1966) describes "unconscious ordering in the prose of Swift".
7. **Syntactic analysis.** Such analyses require the preparation of syntactically annotated corpora. For example, Antosch (1969) showed that a high adjective to verb ratio was found in folk tales, but a much lower ratio was found in scientific texts. Baayen/van Halteren/Tweedie (1996) studied the frequencies of use of phrase structure rewrite rules in annotated corpora to distinguish between crime fiction authors, and Santini (2004) extracted syntactic features for genre classification.
8. **Pause patterns in Shakespeare's verse** (Jackson 2002).

The ideal situation for authorship studies is when there are large amounts of undisputed text, and few contenders for the authorship of the disputed text(s). The **experimental methodology** for determining which of two authors are the more likely to have written a newly-discovered or disputed text is as follows: Build corpora A and B, containing texts undisputedly written by authors A and B respectively, and then build corpus C consisting of works of disputed authorship, but probably written by either A or B. Then select a set of **discriminators** and an appropriate **statistical measure**. When these have been shown

to discriminate effectively between A and B, try them on corpus C, to see whether the works in corpus C more closely resemble those in corpus A or those in corpus B. This method of distinguishing between authors also works for other differences in authorial style such as gender. Corpus A would contain texts known to have been written by female authors, Corpus B would contain texts definitely written by male authors, and a suitable choice of discriminators and a statistical test would show whether corpus C, containing text(s) by an unknown author, has characteristics more typical of male or female authorship.

A number of modern (post 1945) developments have enabled great advances in stylometry. Modern statistical techniques enable us to look for significant differences between the data sets as opposed to chance variation, such as the t-test (Binongo/Smith 1999) and the z-score (Burrows 2002). Modern sampling techniques mean that we have no need to examine the entire works of a given author before making inferences about that author's style. Thirdly, the advent of computers has enabled fast and accurate calculations, and storage of large text corpora. To date, there are no commercial computer packages for stylometry. Many studies have developed analyses of texts by hand, or using simple frequency counting programs written in languages such as Perl (which has been designed especially for text handling), they have extracted the relevant data and processed it using statistical software packages (such as SPSS or MATLAB, see article 36) or manual statistical analysis.

The major difficulty with identifying the style of an individual author is that an individual style can be masked by a number of related issues, such as the following:

1. Heterogeneity of authorship over **time**. A number of authors have been shown to vary in their stylistic traits over time. The earliest such study producing a chronology of texts, by Yardi (1946) showed that certain features of Shakespeare's writing varied as he got older. In this article we will look more closely at how discriminators were chosen to distinguish between the younger and the older Yeats (Forsyth 1999).
2. Authorship and **genre**. Genre differences have been found to be more pronounced than author differences. Thus Baayen/van Halteren/Tweedie (1996) ensured that in order to discriminate between a pair of authors, the comparison texts were all in the genre of crime fiction.
3. Authorship and **gender**. Rayson/Leech/Hodges (1997) showed differences in the vocabulary used by men and women in the spoken component of the British National Corpus, and Koppel/Argamon/Shimon (2002) were able to automatically categorise texts by author gender.
4. Variation **within a single author**. A number of very versatile authors, such as Jane Austen (DeForest/Johnson 2001) and Oliver Goldsmith (Dixon/Mannion 1993), show consistent variations between the characters of their novels. A related problem is that traits of a school of writers (such as the school of Anglo-Irish writers of which Goldsmith was a member) can overshadow any personal tendencies. According to Laan, “[i]deally, a study of the differences in style within the works of an author should precede an attribution or chronology study concerning that same author” (Laan 1995, 271).
5. In the field of Information Retrieval, which is concerned with searching on the internet, the text categorisation literature is more interested in **categorisation by topic** than by writing style. Categorisation by topic is typically based on medium frequency keywords that reflect the content of the document. However, categorisation of texts

by author style uses precisely those features (such as high frequency function words) that are independent of content (Koppel/Argamon/Shimon 2002).

Despite these competing influences, there have been many successful studies of authorship attribution.

2. Lexical discriminators

Qualitative studies of literary style have focused on the **hapax legomena**, the words which appear only once in the entire text. These tend to be obscure, out-of-date or technical terms, or convey delicate shades of meaning, and thus reflect the background and experience of the author. They form the largest group of words in the vocabulary of a text. For example, in a 100,000 word sample of the British National Corpus, about 9500 are hapax legomena, while only about 1500 words come up twice, about 1000 three times, about 600 four times, and about 300 five times. The problem with the hapax legomena as discriminators is that their individual low rate of occurrence makes them difficult to handle statistically. Statistical tests such as the chi-squared test require at least five occurrences of a discriminator in at least one of the corpora. Thus **quantitative studies** of literary style which use lexical discriminators must make use of the words which appear frequently in the texts.

2.1. Vocabulary richness

Many quantitative studies rely on the concept of **vocabulary richness**. A text has low vocabulary richness if the same limited vocabulary is repeated over and over again, while it has high vocabulary richness if new words continually appear. In the following discussion of **measures of vocabulary richness**, we make use of the following notation:

- (a) tokens N = length of text in words
- (b) types V = number of different words in the text
- (c) hapax legomena V_1 = number of words occurring just once in the text
- (d) dislegomena V_2 = number of words occurring exactly twice in the text
- (e) V_i = number of words occurring exactly i times

The **type / token** ratio depends on the length of the text (being generally less for longer texts), but is a useful measure of vocabulary richness when the comparison texts are of equal length. **Honoré's measure R** (Honoré 1979) depends on the hapax legomena:

$$R = 100 \log N / (1 - (V_1 / V))$$

Sichel's measure S (Sichel 1975) depends on the dislegomena, and is relatively constant with respect to N :

$$S = V_2 / V$$

Brunet's measure W (Brunet 1978) is:

$$W = N^{V-a},$$

where α is a constant (usually 0.17). W was found to be relatively unaffected by text length and to be author specific (Brunet 1978). **Yule's characteristic K** depends on words of all frequencies:

$$K = 10,000 * (M - N) / N^2, \text{ where } M = \sum i^2 \cdot Vi.$$

Yule (1944) used his characteristic K to determine whether *De Imitatione Christi* was more likely to have been written by Kempis or Gerson. The results (see Table 50.1) show that the vocabulary richness of *De Imitatione Christi* is much closer to that of the works by Kempis than that of those by Gerson, and hence Kempis is the more likely author.

Tab. 50.1: Yule's Characteristic K for *De Imitatione Christi*

<i>De Imitatione Christi</i>	K = 84.2
Works definitely by Kempis	K = 59.7
Works definitely by Gerson	K = 35.9

Holmes (1992) performed a stylometric analysis of **Mormon Scripture** and related texts, which used five measures of vocabulary richness: Honoré's R, Yule's K, Sichel's S and two associated parameters called α and θ . His results, averaged over all five measures, are shown in Table 50.2, where all the texts apart from the personal writings of Joseph Smith (the movement's founder) and the King James Bible are samples of Mormon Scripture.

These scores were used to **cluster** the texts, producing a pictorial representation called a **dendrogram** (because it looks like a tree) where texts similar in vocabulary richness would appear close together, and those dissimilar in vocabulary richness would appear

Tab. 50.2: Yule's Characteristic K for Mormon scripture and related texts

J. Smith – personal 1 (J1)	K = 57.7	Mormon 3 (M3)	K = 119.2
J. Smith – personal 2 (J2)	K = 82.1	Mormon 4 (M4)	K = 168.9
J. Smith – personal 3 (J3)	K = 78.6	Mormon 5 (M5)	K = 125.5
Nephi 1 (N1)	K = 145.2	Alma 1 (A1)	K = 149.0
Nephi 2 (N2)	K = 155.2	Alma 2 (A2)	K = 150.6
Nephi 3 (N3)	K = 150.5	Doctrine 1 (D1)	K = 126.9
Jacob (JB)	K = 134.3	Doctrine 2 (D2)	K = 91.6
Lehi (LI)	K = 109.4	Doctrine 3 (D3)	K = 98.9
Moroni 1 (R1)	K = 131.5	Isaiah 1 – King James (I1)	K = 81.3
Moroni 2 (R2)	K = 115.7	Isaiah 2 – King James (I2)	K = 114.2
Mormon 1 (M1)	K = 183.8	Isaiah 3 – King James (I3)	K = 90.9
Mormon 2 (M2)	K = 132.7	Book of Abraham (AB)	K = 146.4

far apart. This involved using the raw values to produce a **similarity matrix**, which stored the similarity between each pair of text samples, using the formula

$$1 - ((X_r - X_s) / \text{range})^2$$

For example, using only the data for Yule's characteristic K , if text R is Jacob, and text S is Lehi, $X_r - X_s = 134.3 - 109.4 = 24.9$. The range is the difference between Yule's K characteristic for the text with the richest vocabulary and the text which was most sparse in vocabulary, which is $183.8 - 57.7 = 126.1$. According to the formula, the similarity between Jacob and Lehi is $1 - (24.9 / 126.1)^2 = 0.96$. Two texts identical in vocabulary richness would have a similarity of 1, while the pair of texts most dissimilar in vocabulary richness has a similarity of 0. The similarity scores were averaged over all five measures. From the similarity matrix, the dendrogram is produced step by step as follows. First the two most similar texts are joined together to form a cluster. Then the next most similar pair of texts is joined together, but if the similarity between the newly formed cluster and a single text is greater than the similarity between any two single texts, then the cluster is joined to that single text to form a larger cluster. The process continues, with the most similar pair of clusters or single texts being joined at each stage, until all the texts belong to a single cluster.

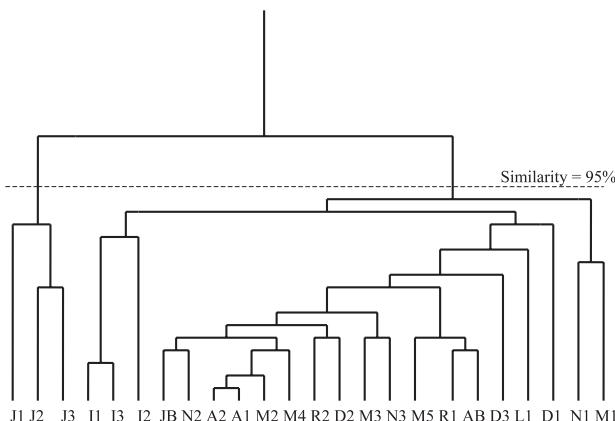


Fig. 50.1: Dendrogram for the total data set

In Holmes' study, the dendrogram (which is reproduced in Figure 50.1) showed that distinct subtrees were found for each of Joseph Smith's personal writings, Isaiah, and the Mormon prophets. However, variation within each Mormon prophet's writings was greater than the variation between prophets. Commenting on this, Holmes stated that he had found no evidence of any multiple authorship in the Book of Mormon, though it was purportedly written by a variety of different prophets at different times. He suggested Joseph Smith's personal writings differed from the Book of Mormon in that he had adopted the style of a "prophetic voice" when writing the Book of Mormon (Holmes 1991). In another study, Pollatschek/Radday (1985) used vocabulary richness to examine the authorship of Genesis. Weaknesses of methods based on vocabulary richness are discussed by Hoover (2004).

2.2. Individual words

Many studies have looked at authors' use of **function** words, which are high-frequency closed-class vocabulary items. A standard list of such words is Taylor's list of ten function words (T10): *but, by, for, no, not, so, that, the, to, with*. Merriam/Matthews (1993) used ratios of these: *no/T10, (of x and)lof, so/T10, (the x and)lthe, with/T10* in a study which showed that an anonymous play, Edward III, was more likely to have been written by Shakespeare than Marlowe. It is believed by many that function words are less under the conscious control of the author compared with rare words. Mosteller/Wallace, in their study of the Federalist Papers (described in more detail in section 3.1.), used 30 individual words which were used much more often by either Hamilton or Madison (the two possible authors of the disputed papers), such as *direction, innovation(s), language, vigor(ous), kind, matter(s), particularly, probability, work(s)*.

DeForest/Johnson (2001) write that the proportion of **Latinate** words to words of Germanic origin in an English text can be used as a stylometric measure. This technique requires the compilation of a large dictionary of words and their origins: DeForest/Johnson classified all 13,809 unique words in Jane Austen's writings. The characters in the novels were found to vary in the proportion of Latinate words they used, with high proportions of Latinate words being indicative of high social class, formality, insincerity and euphemism, self-control (as opposed to emotion), men's speech (since education was the preserve of men in the 18th century) and stateliness as opposed to squalor (for example, contrasting Mansfield Park with the less loved house in Portsmouth). This study is also discussed in article 6.

Collins et al. (2004) used a corpus tagged for **rhetorical language choices**, and found that the writings of Hamilton and Madison in the Federalist Papers differed most in their use of "Think positive" and "Think negative" (used more by Hamilton) and "Past events" (used more by Madison).

3. Statistical measures

Having discussed the choice of discriminators, we must now consider the choice of statistical measure. Standard statistical tests, such as the chi-squared test, the z-score or the t-test, help us to decide whether differences found in the use of discriminators by different authors genuinely show an underlying pattern, or whether they have resulted from relatively small chance variations in the data. Other statistical tests, particularly the multivariate methods such as clustering (Holmes 1992), Correspondence analysis (Mealand 1995) or Principal Components Analysis (PCA) provide pictorial representations, ideally much clearer than the raw data, from which subjective judgements can be made. Related techniques are described in articles 38 and 40. Although not statistical tests as such, we will discuss neural networks and genetic algorithms, used in artificial intelligence, but originally inspired by biology. These models perform the role of statistical tests in stylometry, in that they take in raw data about the frequency of discriminators in a text, and produce as output a decision about the authorship of that text.

3.1. Bayesian probability

Bayesian probability was used by Mosteller/Wallace (1984) to examine a case of disputed authorship in the Federalist Papers. The Federalist Papers were published under the pseudonym “Publius” in 1787–1788 to persuade the people of New York to accept the new American constitution. It is undisputed that Jay wrote 5 of the essays, Hamilton wrote 43, and Madison wrote 14. However, 12 of the essays are disputed. There is some historical evidence that Madison was the author of these twelve essays, but Hamilton, on the night before he was killed in a duel, left a list of the essays and their authors at a friend’s house. On this list, the 12 disputed essays were attributed to Hamilton. Mosteller/Wallace found that the styles of Hamilton and Madison varied in the frequency of use of certain words. For example, *enough* was found in 14 papers by Hamilton, but none by Madison; *whilst* was found in no papers by Hamilton, but in 13 by Madison. 30 such discriminating terms were found. They proceeded as follows (Francis 1966):

If we know the average number of times a word appears in a text of fixed length, we can find the proportion of text sections of that length which have none, one, two, etc. occurrences of that word using the **Poisson distribution**:

$$P_n = \frac{\lambda^n \cdot e^{-\lambda}}{n!}$$

e to the power $-\lambda$ is the $\exp(-\lambda)$ on a calculator. λ is the average number of times the word occurs per section of text, while P_n is the proportion of text sections which have n occurrences of the word. Suppose the average rate of use of *also* is 0.5 words per 2000 for Hamilton and 1.0 words per 2000 for Madison, and suppose, too, that the word appears four times in a disputed paper of length 2000 words.

The probability of *also* occurring exactly four times in text by Hamilton is denoted by P_4 :

$$P_4 = \frac{\lambda^4 \cdot e^{-\lambda}}{4!} = \frac{0.5^4 \cdot e^{-0.5}}{24} = 0.00158$$

The corresponding calculation for Madison is:

$$P_4 = \frac{1 \cdot e^{-1}}{24} = 0.0153$$

Thus it is more likely that Madison wrote the paper, with a likelihood ratio of 0.0153 / 0.00158 giving odds of about 10 to 1. This evidence is combined with other evidence, such as historical evidence giving initial odds of 3 to 1, and other words with high discriminating power, such as *an* which occurs 7 times in the unknown document, thus favouring Madison with odds of about 8 to 3. These three pieces of evidence can be combined using Bayes’ theorem, giving odds

$$\frac{3}{1} \times \frac{10}{1} \times \frac{8}{3} = \frac{80}{1} \text{ in favour of Madison.}$$

Mosteller/Wallace judged all 12 disputed papers to have been written by Madison.

3.2. Univariate analyses

All the common univariate analyses (which get their name because only the variation in the use of one discriminator between authors is considered) such as the t-test are covered in standard statistics text books, such as the one by Woods/Fletcher/Hughes (1986). However, we will include the **z-score** here as an example, with data derived from a study by Burrows (2002). To understand the z-score, we must be familiar with the notions of the mean and the standard deviation. The mean is the average of all values in a data set, found by adding up all of the values and then dividing by the number of values. The standard deviation is a measure which takes into account the distance of every data item from the mean. If all the values in the data set are exactly equal to the mean, then the standard deviation is 0, otherwise, if they vary from each other, standard deviation will be more than 0.

Using Burrows' (2002) data, a sample of 25 Restoration writers shows that the mean occurrence of the word *the* is 4.719%, and the standard deviation is 0.63%. In one of these text samples (actually taken from Milton's *Paradise Lost*), the occurrence of *the* is 4.242%. We can now calculate the measure called the z-score, as follows:

$$z = \frac{x - \bar{x}}{s} = \frac{4.719 - 4.242}{0.63} = 0.757$$

This value can be looked up in a normal distribution table (there is one at the back of most statistics textbooks), to show that we would expect a sample which belongs to the collection of Restoration writers to have a z-score of 0.757 or more about 45% of the time. Only if the resulting z-score would be expected to occur in the collection just 5% of the time or less, would we suspect that the Milton sample did not really belong to this collection. Burrows extends the z-score idea to produce the delta score, designed as a measure capable of distinguishing the most likely candidate from a large group, while traditional studies discriminate only among a small number of possible authors.

3.3. Cumulative sum charts

A. Q. Morton believed that the rate of occurrence of a (stylistic) habit is so consistent for each individual that any distinct variation in the proportion of occurrences of the habit within a sample of sentences is *prima facie* evidence that the sentences are the utterance of more than one person. This is the rationale behind Morton/Michaelson's (1990) controversial cusum (cumulative sum) technique which has been accepted by several courts of law in cases revolving around allegedly forged confessions, such as the successful appeal against a robbery conviction by Tommy McCrossen in 1991. However, in another case where Bob Maynard and Reg Dudley were accused of two London gangland murders, Morton's evidence in their favour was successfully rebutted at the Court of Appeal (Campbell 1992, 1997).

Two plots are drawn on the same graph – one corresponding to the lengths of the sentences in the text under study, and one corresponding to the occurrence of some other

linguistic feature, such as the number of words with initial vowels. The cusum values s_i to be plotted for each sentence i are given by the formula

$$s_i = \sum_{r=1}^i (x_r - \bar{x})$$

where r refers to the individual sentences from the start of the text up to and including sentence i , x_r is either the sentence length or the number of times the chosen linguistic feature is found in sentence r , and \bar{x} is either the average (mean) sentence length for that text, or the average number of times the linguistic feature is found per sentence in that text. On the cusum plot, s_i is plotted on the vertical axis, while i is plotted on the horizontal axis. The theory is that if a work was written by a single author, the ratio of occurrences of the linguistic feature to sentence length will be relatively constant, and thus the two lines drawn on the cusum chart, if suitably scaled on the vertical axis, will follow each other almost exactly. On the other hand, if the early part of the text was written by one author, and the later part of the text by another, the two lines on the cusum chart will start to diverge at the point where the authorship changes. In the example shown in Figure 50.2, from Hilton/Holmes (1993), there is a divergence in the lines near to sentence 25, the point at which a sample of *Northanger Abbey* ends and a sample of *The Great Gatsby* begins.

On two occasions rebuttal evidence against the cusum technique has been prepared for the crown by David Canter (Campbell 1992). Canter (1992) showed that the technique was not reliable, whether one judges by a subjective comparison of the two plots on the cusum chart, or attempts to quantify the correlation between the two lines using Spearman's rank correlation coefficient. Holmes/Tweedie (1995) also give a critique of

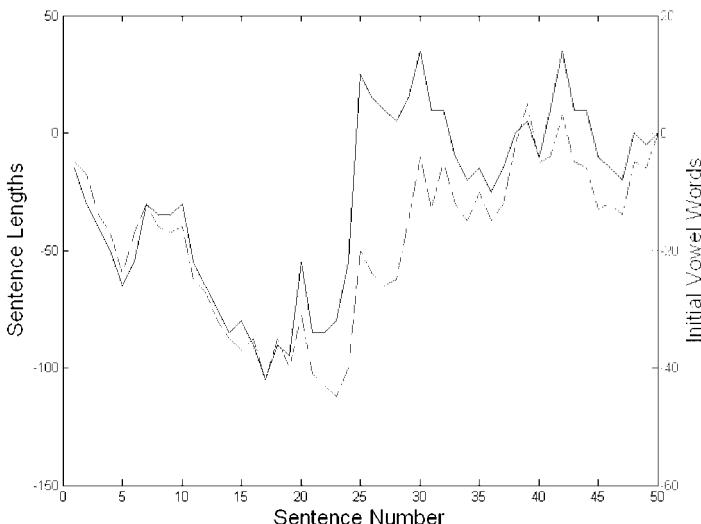


Fig. 50.2: Cusum Plot for *Northanger Abbey* and *The Great Gatsby*. There is a significant discrepancy in the plot near sentence twenty-five, where the samples are concatenated. The solid line is the plot of sentence lengths, while the dotted line is the plot of words with initial vowels

the cusum controversy. To overcome the subjectivity of the technique, Hilton/Holmes proposed the weighted cumulative sums technique, where the two lines of the cusum chart are combined into a single line. They use a version of the t-test to see if there is a significant difference between an earlier and a later portion of the line. Although this puts the technique on surer statistical foundations, Hilton/Holmes found that the weighted cusum technique performed only marginally better than the cusum test, and that neither technique gave consistently reliable results. They concluded that authors are not as consistent in their selection of linguistic features as would be required for cusum techniques to determine authorship correctly.

Barr (1997) found that cusum graphs were useful for examining scale differences between authors. As an example of scale differences, one author might have a typical opening pattern of long sentences, a middle pattern of sentences of mixed length, and a closing pattern of short sentences. These proportions might be retained in other works by the same author, even if these other works are of varying length.

3.4. Multivariate analysis

Multivariate analyses (see also articles 38 and 40) are so-called because they take into account a large number of variables or discriminators. If each discriminator can be thought of as a dimension, multivariate analyses attempt a **reduction in dimensionality**. If several discriminators tend to vary together, i. e. in texts where one is frequently found, the others are often found also, they are combined into a single dimension (called a factor or principal component, according to the type of analysis being performed), thereby achieving this reduction in dimensionality. For example, we could start with four dimensions (though a real study might use many more), being the whole-word discriminators *he*, *her*, *him*, *she*. These words might be distributed across five texts as shown in Table 50.3.

Tab. 50.3: Candidate data set for dimensionality reduction

	He	Her	Him	She
Text 1	10	2	12	0
Text 2	11	1	15	3
Text 3	9	1	10	1
Text 4	0	14	0	14
Text 5	1	12	2	11

Here we see that the discriminators *he* and *him* tend to be found in the same texts as each other, and so do *her* and *she*. Thus we can reduce the original four dimensions to just two (one for *he/him*, and one for *her/she*). Here there are only two dimensions left, but in a typical study, the two most important dimensions (in terms of accounting for the variation in the data) will account for about half the total variation, and the less important dimensions are disregarded. The advantage of cutting down to two dimensions is that they can become the two axes of a graph, on which all the texts can be

plotted. Text 1 contains 22 words in the *he/him* dimension, and 2 in the *her/she* dimension. Thus it can be plotted at point (22,2) on the graph. (In real life, the correspondence between the raw data and the position on the dimension axis is not so exact). Once all the texts have been positioned on the graph, they will hopefully appear in clusters, where all the texts written by one author will be positioned close together, clearly set apart from the texts by other authors.

In this section we will consider a study which used **Principal Components Analysis** (PCA), although Factor Analysis and Correspondence Analysis also follow this dimensionality reduction and two-dimensional plot paradigm. Burrows (2002) found that a problem of PCA was that if one of the texts was added to or removed from the analysis, the whole pattern can alter so that we are no longer able to make strict comparisons between graph and graph. However, “in experienced hands, such methods yield excellent results” (*ibid.*, 269). It is important to perform some sort of cross-validation, to ensure that the main clusters remain, when each single text is removed in turn.

Fifty years after the American Civil war, General George Pickett’s widow, LaSalle Corbell Pickett, published letters purportedly written by her husband, many of them written during his active service in the war. Historians were divided as to their authentic-

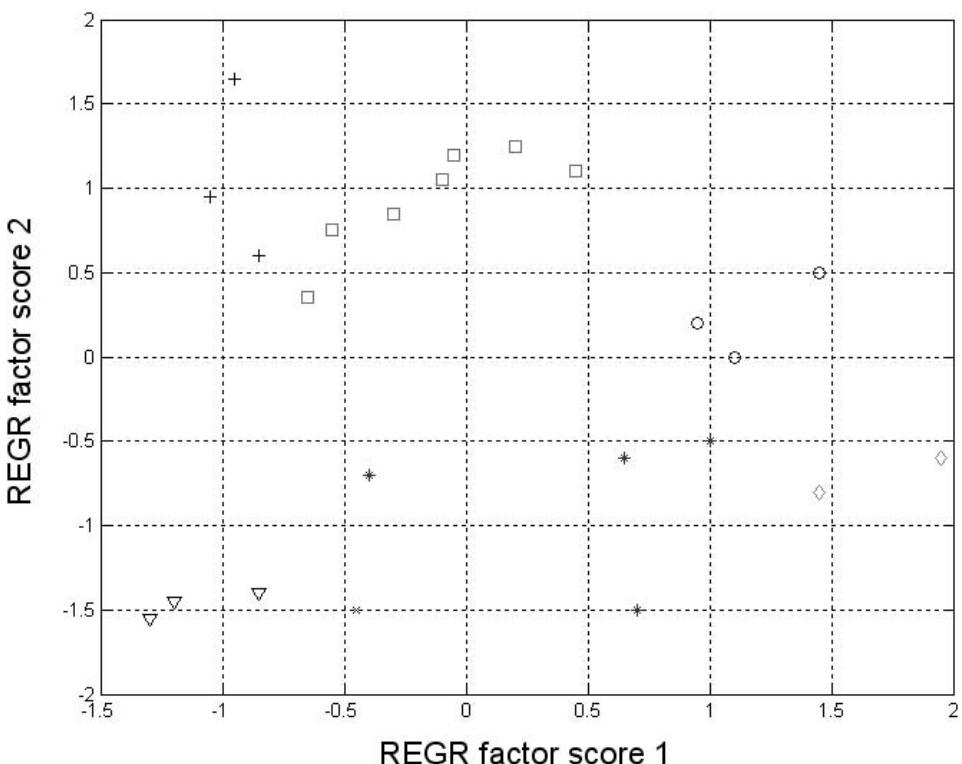


Fig. 50.3: Principal components plot: attribution. + = autobiography, o = LaSalle letters, * = George personal, x = George war, square = “Heart”, diamond = Inman papers, triangle = Harrison.

ity, so Holmes/Gordon/Wilson (2001) used PCA for a stylometric investigation into the Pickett letters. Starting with the 60 most common words in the texts as discriminators, the analysis produced two principal components. Various text samples were plotted on the resulting two-dimensional plot for comparison, as shown in Figure 50.3.

- (a) auto, LaSalle Pickett's autobiography
- (b) ltr, LaSalle Pickett's personal letters
- (c) gp, George Pickett's personal pre-war and post-war letters
- (d) hs, the disputed letters
- (e) gw, George Pickett's war reports
- (f) i, Inman papers, genuine handwritten letters by George Pickett
- (g) har, Walter Harrison's book "Pickett's men" (one theory being that the letters were plagiarised from this book).

The investigation strongly suggests that LaSalle Pickett composed the published letters herself. All the seven sources listed above are internally consistent, forming clusters, and the samples of LaSalle Pickett's autobiography fall suspiciously close to the disputed letters. Tweedie/Holmes/Corns (1998) used PCA to examine the provenance of *De Doctrina Christiana*, traditionally attributed to John Milton. Mealand (1995) carried out a Correspondence Analysis of *Luke* confirming modern theological opinion on the sources of that gospel. Factor analysis has been successfully used for the study of register variation in English (Biber 1995). Clustering (see section 2.1.) is also a form of multivariate analysis.

3.5. Neural networks

A typical neural network architecture, called a multi-layer perceptron, is exemplified by Figure 50.4, taken from Matthews/Merriam (1993). It consists of three layers of nodes: the input layer has five nodes, one for each discriminator used in the study, there is a middle, "hidden" layer, and an output layer of two nodes (one for each possible author of a text). For each text, the input nodes are activated in proportion to how often that discriminator is found in the text. This activation is passed on to the middle layer, according to the strengths of connection (or **weights**) between the nodes of those two layers. In turn, the activation is passed on to the outer layer – one node will be activated more if the network "thinks" that author 1 wrote the text, while the other will be activated more if it seems that author 2 wrote the text. This will only work if the weights are correct. They are initially random, but in a prior "training" session, the weights are gradually updated in response to the frequency of discriminators and the identity of authors of known texts, until the network is right every time. The unknown texts (the "test data") are then presented to the network one by one, and in each case the network gives an "opinion" (author 1 or author 2) as to who wrote that text.

Matthews/Merriam (1993) created a neural network to distinguish Shakespeare and Fletcher. Hoorn et al. (1999) produced networks to distinguish three Dutch poets, Bloem, Slauerhoff and Lucebert. Waugh/Adams/Tweedie (2000) discuss how to minimise the number of nodes in the hidden layer for a stylometric study. Kjell's (1994) study is interesting, because his network has 26×26 input nodes, one for each possible bigram

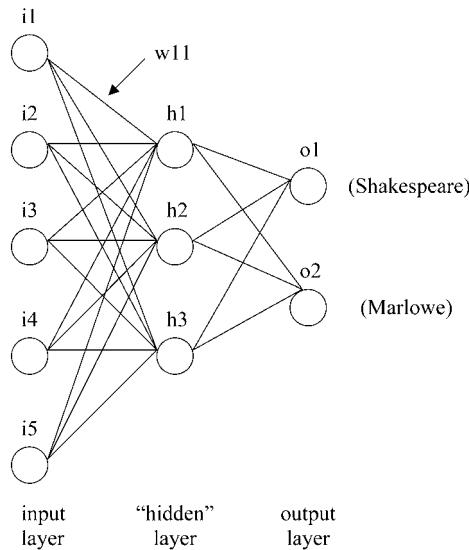


Fig. 50.4: A stylometric multi-layer perceptron

or pair of adjacent characters. This means that there is no need to manually create a set of discriminators on an “author pair” basis. After training, bigrams which have not proved to be useful discriminators have been given zero weights, and thus by process of elimination, the network has “learned” a set of good discriminators.

3.6. Genetic algorithms

Another model based on biological analogy is the genetic algorithm, based on the Darwinian idea of Natural Selection. Forsyth's PC/BEAGLE system, originally devised for weather forecasting, was applied to stylometry in a new study on the Federalist Papers (Holmes/Forsyth 1995). It consisted of a set of **rules**, made up of operators, variables and constants, in the form “IF such-and-such is true, THEN the essay was written by Hamilton”. For example $(KIND < 0.00002) \& (TO < 35.205) \rightarrow \text{Hamilton}$ means that if the word *kind* appears less than 0.00002 times per thousand words, and also the word *to* appears less than 35.205 times per thousand, then the article must have been written by Hamilton. Another example would be $(UPON - BOTH) < WHILST \rightarrow \text{Hamilton}$. The word variables are drawn from Mosteller/Wallace's (1984) 30 marker words. PC/BEAGLE's learning algorithm is given below:

1. Create an initial population of candidate rules at random (all of which must be syntactically valid, but most of which will be semantically meaningless).
2. Evaluate each rule on every training example (texts of known authorship) and compute a fitness score, based on how often the rule predicts the author correctly, with a penalty for rule length to encourage brevity.
3. Rank the rules in descending order of merit and remove the bottom half.

4. Replace discarded rules by crossing a pair of randomly selected survivors to produce “offspring”. This “mating” is achieved by picking out a random sub-expression from each of two surviving rules and tying them together with a randomly chosen connective. Possible descendants of the two sample rules given above could be (*UPON – BOTH*) < 35.205 → *Hamilton*, and (*KIND* < 0.00002) & (*TO* < *WHILST*) → *Hamilton*.
5. Mutate a small number of rules picked at random (excluding the best rule), by changing one component, and apply a tidying procedure to reduce redundancy. If the termination conditions are met (e. g. a new generation of rules has not shown any improvement over the last generation), training is complete. Otherwise, return to stage (2).

Once training is complete, the set of rules can be applied to cases that may not have been seen before, in order to determine their authorship. As with Kjell’s neural network, the genetic algorithm, rather than being restricted to the 30 marker words, is free to choose its own set of discriminators.

4. Automatic selection of discriminators

There are two main advantages of selecting discriminators automatically, namely that to do this manually is time consuming, and also that it results in sets of discriminators that work on one author pair but not necessarily on others. The process is necessarily subjective, and thus each stylometrist might have a “tool-kit” of favourite marker types, leading them to overlook the vast majority of those discriminators that might be used (Holmes/Forsyth 1995). A fourth reason is given by Burrows (2002, 268):

“a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. Strong features, perhaps, are easily recognised and modified by an author and just as easily adopted by disciples and imitators. At all events, a distinctive stylistic signature is usually made up of many tiny strokes.”

Koppel/Argamon/Shimon (2002) explore the possibility of automatically classifying formal written texts by author **gender**. They start with 1081 features, chosen solely for relative topic independence: 405 function words, the 500 commonest part-of-speech trigrams, the 100 commonest bigrams, and all 76 single part of speech categories. Using their “Winnow” learning method, they were able to iteratively discard the features which were not good discriminators. The last features to die off for fiction texts were *a, the, as* for male authors; *she, for, with, not* for female authors. When training on non-fiction they found the last features to disappear were *that, one* for male authors, and *for, with, not, and, in* for female authors. For parts of speech the male indicators were determiners, numbers and modifiers, while the most effective female discriminators were negation, pronouns and some prepositions. Best results for determining the gender of the authors of new texts was obtained when 64 to 128 features were retained, which gave an accuracy of 84% for non-fiction, and 80% for fiction.

As we saw in sections 3.5. and 3.6., neural networks and genetic algorithms are able to learn their own sets of discriminators. In sections 4.1. and 4.2. respectively, we will look at machine learning techniques for selecting whole word and character substring discriminators (for machine learning techniques see also article 39).

4.1. Whole words elimination

Five methods are included in the text classification study by Yang/Pedersen (1997), namely document frequency (DF), information gain (IG), mutual information (MI), a chi-squared statistic (CHI) and term strength (TS). In each case, every word in the texts to be classified is initially a potential discriminator, but the majority of these are discarded by giving each one a numeric score. If, for example, we wanted to reduce the number of discriminators to 100, then only the 100 highest scoring words would be retained. Only the first method, **document frequency**, is unsuitable for stylometry, since it selects the mid-frequency terms that are more useful for detecting the topic rather than the writing style.

Information gain measures the amount of information, measured in bits, obtained for category prediction by knowing the presence or absence of that term in a document. To calculate **mutual information**, consider the two-way contingency table of a word w and author a , where A is the number of times word w is used by the first author, B is the number of times word w is used by the other author, C is the number of times the first author used any term other than w , and D is the number of times the other author uses any word other than w . N is the total number of word tokens in the entire data set. $MI(w,a)$, the mutual information between a word and an author, is estimated using

$$MI(w,a) \approx \log \frac{AN}{(A+C)(A+D)}$$

$MI(w,a) = 0$ if the word and the author are independent of each other, but has a positive value if the use of the word suggests that author, and a negative value if the author consistently avoids the use of that word. The problem with this measure is that MI favours rare words. Using the same contingency table, we can work out the **chi-squared** statistic:

$$X^2(w,a) = \frac{N(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

Chi-squared = 0 if the term and the category are independent. The chi-square statistic is also known not to be reliable for low-frequency terms. Calculation of Yang/Pederson's fifth measure, **term strength**, is more complex.

Binongo/Smith (1999) describe the use of the **t-test** (independent samples, i. e. non-matched pairs) to find words that discriminate between ten blocks of text by Shakespeare and five by Wilkins. The best discriminants (those producing the highest t-scores) were in order: *then*, *the*, *by*, *alan/awhile*, *on*, *no*, *most*, *tol/into*, *there* and *for/forever*. Classes of words not considered were verbs (due to their various inflected forms), nouns and personal pronouns (which are often context-dependent) and rare words. For example, using data provided by Binongo/Smith, the occurrence of the word *the* in five samples of *Cymbeline* (Shakespeare) is 166, 163, 165, 177 and 174. The same word is found in four samples of Wilkins' *The Miseries of Enforced Marriage*, 97, 112, 138 and 112 times. Every sample consisted of 5000 words. The mean and standard deviation are 169 and 6.12 for *Cymbeline*, and 114.75 and 17.04 for *Miseries*. These two standard deviations are combined to produce a common standard deviation, s :

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(5 - 1).169^2 + (4 - 1).114.75^2}{5 + 4 - 2}} = 12.44$$

This value is then used in the calculation of t , as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{169 - 114.75}{\sqrt{\frac{12.44^2}{5} + \frac{12.44^2}{4}}} = 6.501$$

The corresponding calculation for *and* yields a t value of only 2.097, so the occurrence rate of *the* is a better discriminator than the occurrence rate of *and* for *Cymbeline* and *Miseries*.

4.2. Monte-Carlo Feature Finding

Assigning a date to a text is the main task of **stylochronometry**. In this section we will look at Forsyth's (1999) study which compared the writing of the older (post-1915) and the younger Yeats. He aimed to first develop a stylochronometric technique for an author whose chronology is well established, so that this technique could later be used for authors whose dating is less well documented. Yeats himself insisted that his language changed as he grew older, and most readers would agree, but what exactly were these changes? To find out, Forsyth used an algorithm, called **Monte-Carlo Feature Finding (MCFF)**. Initially, all character sequences of eight characters or less, found in at least one of the poems chosen for the training set, are regarded as potential discriminators. Since there are so many of these, a random sample of just 4096 was taken. Each of these substrings were ranked according to their **distinctiveness**, as measured by the **chi-squared test** (see section 4.1.). The training data was divided into two portions: 72 poems representing the younger Yeats, and 70 poems representing the older Yeats. 88 distinctive substrings were identified, and the most distinctive of all are tabulated in Table 50.4.

The columns labelled YY-count and OY-count show how often each substring was found in the younger and older Yeats samples respectively. The counts of all 88 retained substrings were then found in thirteen other poems written between 1891 and 1931, and

Tab. 50.4: Top six discriminators for the younger and older Yeats

Rank	Substring	Chi-squared	YY-count	OY-count
1	“what”	35.1	30	100
2	“can”	34.3	21	82
3	“s, an”	25.4	63	19
4	“whi”	25.4	67	21
5	“with”	22.3	139	74
6	“?”	21.9	30	83

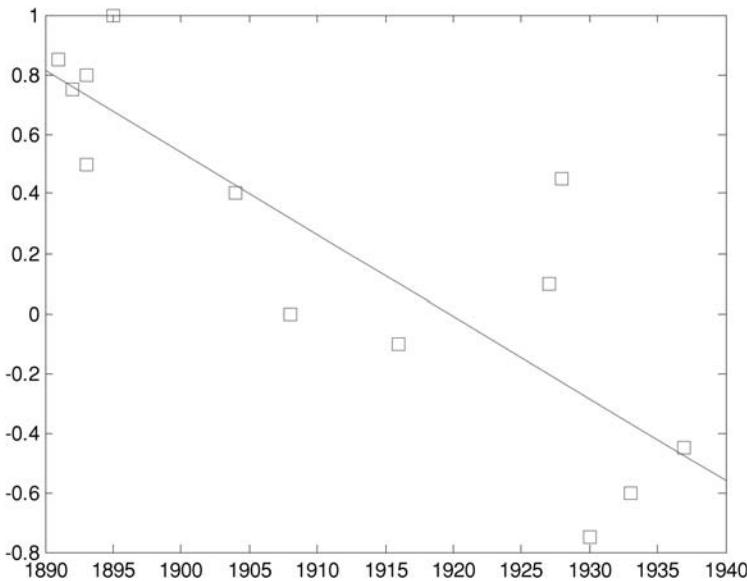


Fig. 50.5: Plot of Youthful Yeatsian Index against date

a plot was drawn of “youthful Yeatsian index” (YYI) against year, as shown in Figure 50.5. A straight line of best fit (produced by the least squares method) is drawn on this plot, from which the date of unseen texts can be read off. The “youthful Yeatsian index” was defined as follows:

$$YYI = (YY - OY) / (YY + OY)$$

To date an unseen text, one should count the number of substrings distinctive of either the younger or older Yates (these are YY and OY respectively), then calculate YYI. Find this value of YYI on the vertical axis, then read across to the line of best fit, then read down to find the estimated date of authorship on the horizontal axis.

5. Conclusion

In this review of computational stylometry we have looked mainly at studies of disputed authorship, but also considered that very similar techniques have been used for studies of genre, gender, and variation within a single author’s writings which occur with time or between characters. The basic technique is to find samples of text known to have been written by each author considered in a particular study. In these texts we must find, whether manually or using machine-learning techniques, features such as whole words or character sequences which act as discriminators, since they occur more frequently in texts of one category than another. The frequencies of these discriminators are then found for unseen texts. Statistical tests, or models based on biological analogy,

are then used to convert the raw data of discriminator counts in each of the texts into a definitive answer as to who was the likeliest author of each unseen text.

In the introduction, a caveat was given that differences in the styles of individual authors can be swamped by differences in genre and time. In fact, many of the techniques described in this article are also used in studies designed specifically to look at genre and diachronic studies. A technique very similar to that of Forsyth (1999) was used by Milton (1998) to compare the English used by native speaker students and learners of English in Hong Kong. Instead of looking for sequences of up to eight characters to identify either the younger or older Yeats, he identified sequences of four words which were typically overused or underused by the learners of English compared with the native speakers. Recently, Baroni/Bernardini (2006) used text categorisation techniques to compare original Italian texts and texts translated into Italian ("translationese") from other languages. The techniques studied in this article are largely language independent, apart from the need for separate lists of function words for each language. Hu/Williamson/McLaughlin (2005) are building a corpus for a diachronic study of Chinese, where the characteristics of Chinese written at widely different times and in different genres will be sought.

6. Acknowledgements

Permission to reproduce Figure 50.1 in this article, which originally appears in the *Journal of the Royal Statistical Society*, was granted by Blackwell Publishing. This figure was Figure 1 of Holmes (1992). Permission to reproduce Figures 50.2 to 50.5 in this article, which originally appeared in the *Journal of Literary and Linguistic Computing*, was kindly granted by the Oxford University Press. These figures were Figure 5 of Hilton and Holmes (1993), Figure 11 of Holmes/Gordon/Wilson (2001), Figure 1 of Merriam/Matthews (1993), and Figure 1 of Forsyth (1999). All figures appear by kind permission of the original authors.

7. Literature

- Antosch, F. (1969), The Diagnosis of Literary Style with the Verb-Adjective Ratio. In: Dolezel, L./Bailey, R. W. (eds.), *Statistics and Style*. New York: American Elsevier.
- Baayen, H. H./van Halteren, H./Tweedie, F. (1996), Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, In: *Literary and Linguistic Computing* 11, 121–132.
- Baroni, M./Bernardini, S. (2006), A New Approach to the Study of Translationese: Machine Learning the Difference between Original and Translated Text. In: *Literary and Linguistic Computing* 21, 259–274.
- Barr, G. K. (1997), The Use of Cumulative Sum Graphs in Literary Scaleometry. In: *Literary and Linguistic Computing* 12(2), 105–111.
- Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
- Binongo, J. N. G./Smith, M. W. A. (1999), The Application of Principal Component Analysis to Stylometry. In: *Literary and Linguistic Computing* 14(4), 445–465.

- Brunet, E. (1978), *Vocabulaire de Jean Giraudoux: Structure et évolution*. Paris: Slatkine.
- Burrows, J. (2002), ‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship. In: *Literary and Linguistic Computing* 17(3), 267–287.
- Campbell, D. (1992), Writing’s on the Wall. In: *The Guardian*, October 8, 1992.
- Campbell, D. (1997), Body of Evidence. In: *The Guardian*, August 7, 1997.
- Canter, D. (1992), An Evaluation of the “Cusum” Stylistic Analysis of Confessions. In: *Expert Evidence* 1(3), 93–99.
- Collins, J./Kaufer, D./Vlachos, P./Butler, B./Ishizaki, S. (2004). Detecting Collaborations in Text. In: *Computers and the Humanities* 38, 15–36.
- DeForest, M./Johnson, E. (2001), The Density of Latinate Words in the Speeches of Jane Austen’s Characters. In: *Literary and Linguistic Computing* 16(4), 389–401.
- Dixon, P./Mannion, D. (1993), Goldsmith’s Periodical Essays – a Statistical Analysis of Eleven Doubtful Cases. In: *Literary and Linguistic Computing* 8(1), 1–19.
- Forsyth, R. S. (1999), Stylochronometry with Substrings, or: A Poet Young and Old. In: *Literary and Linguistic Computing* 14(4), 467–477.
- Francis, I. (1966), An Exposition of a Statistical Approach to the Federalist Dispute. In: Leed, J. (ed.), *The Computer and Literary Style*. Kent OH: Kent State University Press.
- Hilton, M. L./Holmes, D. I. (1993), An Assessment of Cumulative Sum Charts for Authorship Attribution. In: *Literary and Linguistic Computing* 8, 73–80.
- Holmes, D. I. (1991), Vocabulary Richness and the Prophetic Voice. In: *Literary and Linguistic Computing* 6, 259–268.
- Holmes, D. I. (1992), A Stylometric Analysis of Mormon Scripture and Related Texts. In: *Journal of the Royal Statistical Society Series A* 155, 91–120.
- Holmes, D. I. (1997), Stylometry, its Origins, Development and Aspirations. In: Rudman, J./Holmes, D. I./Tweedie, F. J./Baayen, R. H. (chairs), session entitled The State of Authorship Attribution Studies. In: *ACH-ALLC ’97 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computers*. Kingston, Ontario, Canada, June 3–7, 1997. <http://www.cs.queensu.ca/achallc97/papers/s004.html>.
- Holmes, D. I./Forsyth, R. S. (1995), The Federalist Revisited: New Directions in Authorship Attribution. In: *Literary and Linguistic Computing* 10(2), 111–127.
- Holmes, D. I./Gordon, L. J./Wilson, C. (2001), A Widow and her Soldier: Stylometry and the American Civil War. In: *Literary and Linguistic Computing* 16(4), 403–420.
- Holmes, D. I./Tweedie, F. (1995), Forensic Stylometry: A Review of the Cusum Controversy. In: *Revue Informatique et Statistique dans les Sciences Humaines* 31, 19–47.
- Honoré, A. (1979), Some Simple Measures of Richness of Vocabulary. In: *Association for Literary and Linguistic Computing Bulletin* 7(2), 172–177.
- Hoorn, J. F./Frank, S. L./Kowalczyk, W./van der Ham, F. (1999), Neural Network Identification of Poets Using Letter Sequences. In: *Literary and Linguistic Computing* 14(3), 311–338.
- Hoover, D. I. (2002), Frequent Word Sequences and Statistical Stylistics. In: *Literary and Linguistic Computing* 17(2), 157–180.
- Hoover, D. I. (2003), Frequent Collocations and Authorial Style. In: *Literary and Linguistic Computing* 18(3), 261–286.
- Hoover, D. I. (2004), Another Perspective on Vocabulary Richness. In: *Computers and the Humanities* 37(2), 151–178.
- Hu, X./Williamson, N./McLaughlin, J. (2005). Sheffield Corpus of Chinese for Diachronic Linguistic Study. In: *Literary and Linguistic Computing* 20, 281–293.
- Jackson, MacD. P. (2002), Pause Patterns in Shakespeare’s Verse: Canon and Chronology. In: *Literary and Linguistic Computing* 17(1), 37–46.
- Kenny, A. J. P. (1982), *The Computation of Style*. Oxford: Pergamon Press.
- Kjell, B. (1994), Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. In: *Literary and Linguistic Computing* 9, 119–124.

- Koppel M./Argamon, S./Shimon, A. R. (2002), Automatically Categorizing Written Texts by Author Gender. In: *Literary and Linguistic Computing* 17(4), 401–412.
- Laan, N. (1995), Stylometry and Method. The Case of Euripides. In: *Literary and Linguistic Computing* 10(4), 271–278.
- McEnery, A. M./Oakes, M. P. (2000), Authorship Identification and Stylometry. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 545–562.
- Mealand, D. L. (1995), Correspondence Analysis of Luke. In: *Literary and Linguistic Computing* 10, 85–98.
- Merriam, T. V. N./Matthews, R. A. J. (1993), Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. In: *Literary and Linguistic Computing* 9(1), 1–6.
- Michaelson, S./Morton, A. Q. (1976), Things ain't What they Used to be. In: Jones, A./Churchhouse, R. F. (eds.), *The Computer in Literary and Linguistic Studies*. Cardiff: The University of Wales Press, 79–84.
- Milić, L. T. (1966), Unconscious Ordering in the Prose of Swift. In: J. Leed (ed.), *The Computer and Literary Style*. Kent OH: Kent State University Press, 79–106.
- Milton, J. (1998), Exploiting L1 and Interlanguage Corpora in the Design of an Electronic Language Learning and Production Environment. In: Granger, S. (ed.), *Learner English on Computer*. Harlow: Longman, 186–198.
- Morton, A. Q. (1978), *Literary Detection*. East Grinstead: Bowker Publishing.
- Morton, A. Q./Michaelson, S. (1990), *The Qsum Plot*. Technical Report CSR-3-90, University of Edinburgh.
- Mosteller, F./Wallace, D. L. (1984), *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer.
- Pollatschek, M./Radday, Y. T. (1985), Vocabulary Richness and Concentration. In: Radday, Y. T./Shore, H. (eds.), *Genesis – An Authorship Study*. Rome: Biblical Institute, 191–214.
- Rayson, P./Leech, G./Hodges, M. (1997), Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: *International Journal of Corpus Linguistics* 2(1), 133–152.
- Santini, M. (2004). A Shallow Approach to Syntactic Feature Extraction for Genre Classification. In: *7th Annual Research Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK)*. University of Birmingham, 207–214.
- Sichel, H. S. (1975). On a Distribution Law for Word Frequencies. In: *Journal of the American Statistical Association* 70, 542–547.
- Tweedie, F. J./Holmes, D. I./Corns, T. N. (1998), The Provenance of De Doctrina Christiana, Attributed to John Milton: A Statistical Investigation. In: *Literary and Linguistic Computing* 13(2), 77–87.
- Waugh, S./Adams, A./Tweedie, F. (2000), Computational Stylistics Using Artificial Neural Networks. In: *Literary and Linguistic Computing* 15(2), 187–197.
- Woods, A./Fletcher, P./Hughes, A. (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Yang, Y./Pedersen, J. (1997), A Comparative Study on Feature Selection in Text Categorization. In: *International Conference on Machine Learning (ICML-97)*. Nashville, TN, 412–420.
- Yardi, M. R. (1946), A Statistical Approach to the Problem of Chronology in Shakespeare's Plays. In: *Sankhya (Indian Journal of Statistics)* 7(3), 263–268.
- Yule, G. U. (1944), *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Michael P. Oakes, Sunderland (UK)

51. Historical corpus linguistics and evidence of language change

1. Introduction
2. Relevant terminology
3. Background considerations for collecting and evaluating historical evidence
4. Corpus-based evidence for historical linguistic developments in English
5. Conclusion
6. Literature

1. Introduction

Historical corpora have opened new possibilities for what linguists can realistically gather and analyze as evidence of language change, and they have, thereby, enriched not only knowledge of the history of particular languages but also theoretical models of language change. First, the technology behind historical corpora offers linguists ready access to large amounts of data as well as mechanisms for searching the data systematically and reliably. These data capture stages of linguistic development over time, allowing researchers to pursue contrastive or comparative studies of the language at different moments in its history. Second, computerized corpora facilitate the analysis of linguistic relationships and their statistical significance – be those collocational relationships between words or sociolinguistic relationships between linguistic features and extralinguistic factors or functional relationships between frequency and changes in linguistic structure. The examination of these relationships critically contributes to a more fully developed understanding of how linguistic changes progress. Third, as collections of actual written texts, historical corpora have placed the emphasis of study squarely on language use, providing a means for scholars to explore the implications of the underlying assumption in more functional linguistic approaches, that language structure is affected by language use. Fourth, the interest in variability that has shaped historical corpus design and studies has had at least these two significant effects: historical corpora often include some less canonical texts, making these texts part of the readily available evidence for studying a language's development; and any sharp division that may have existed between the study of present-day languages and historical stages of languages has been blurred as modern variationist sociolinguistic findings are projected backwards, historical data are used to test modern theories about variation and change, and evidence of variation and change across periods becomes more and more detailed and integrated.

As stated above, modern computer technology and storage capacity have changed what is possible in the gathering and analysis of historical linguistic evidence both in terms of the sheer quantity of data and in terms of the ability to uncover and statistically analyze (socio)linguistic relationships. A computer can read through one million words of text and list all examples of a particular construction in a matter of seconds – a search that would probably take a human reader months at best. And humans are known to get tired while searching and miss evidence in ways that computers do not – if, of course, the linguistic evidence appears in a form the computer can recognize as the

targeted construction. Computers can also relatively quickly process features such as the co-occurrence of variables and the significance of that co-occurrence. With computers doing more of the searching and “number crunching”, human researchers can focus sustained attention on the analysis and interpretation of comparatively large amounts of systematically compiled data and thereby enrich our understanding of how languages change generally and of how specific languages have changed over their history.

Given the greater availability of historical corpora of the English language and of research based on these corpora, this article focuses primarily on the evidence provided in studies of historical developments in the English language and its implications for the study of language change. Many of the observations here, however, hold true for the corpus-based study of any language’s history. True also of other languages, some of the most exciting and innovative new work in scholarship on the history of the English language is corpus-based. Books and conferences on English historical linguistics now regularly feature corpus-based studies (e.g., Wright 2000; Minkova/Stockwell 2003; Curzan/Emmons 2004; the International Conference on English Historical Linguistics); and books and conferences on English corpus linguistics now regularly feature diachronic studies of the English language (e.g., Biber/Conrad/Reppen 1998; the International Computer Archive of Modern and Medieval English Conference).

2. Relevant terminology

The title of this article raises several complex questions about the definition of terms. For example, what is encompassed by “language change”? Spoken language has traditionally been given priority or assumed as what counts as “language” in linguistic studies, both synchronic and diachronic, unless otherwise specified. But historical corpora contain only written texts as evidence.

2.1. Defining what counts as “evidence” for “language change”

A full discussion of what exactly constitutes “language change” extends beyond the scope of this introductory article. One thing is clear: language change is always more complex than any generalization of the form “Language A changed in X way” suggests. For any Language A, the spoken and written forms (where applicable) change in distinct ways and/or at different speeds; in addition, different registers of the written language and different dialects of the spoken language typically change in distinct ways and/or at different speeds. Broad generalizations about language change are certainly useful in capturing sweeping linguistic developments over (often long periods of) time. That said, closer studies that lead to more specific observations about the implementation of a given change are critical. Of the five problems at the heart of investigating language change, as outlined by Weinreich/Labov/Herzog (1968), the transmission or implementation problem – the study of how change spreads, be that through the lexicon, grammar, or speech community – has perhaps been best served by the innovative work in historical corpus linguistics. Variability is built into historical corpora, allowing researchers to pursue what Rissanen (1998, 400) sets out as the aim of historical linguistic studies:

“integrated accounts combining the discussion of linguistic processes of change with the influence of various extralinguistic factors”.

Evidence of language change not only looks like variation over time but also, perhaps more importantly, looks like variation at a given moment in time. As Weinreich/Labov/Herzog (1968, 188) state in their foundational article on the nature of language change, “Not all variability and heterogeneity in language structure involves change; but all change involves variability and heterogeneity”. Histories of a language often frame linguistic change in terms of endpoints: Feature B dropped out of use, replaced by Feature C; or form D shifted from meaning P to meaning Q. Such descriptions make linguistic changes sound tidy, and in retrospect they can look that way. At any given moment during a linguistic change, however, speakers typically experience the change as variation, with some speakers using one variant and other speakers using other variants or with the same speakers using multiple variants, perhaps in different registers. One of the critical insights of modern variationist sociolinguistics is the concept of orderly heterogeneity (Weinreich/Labov/Herzog 1968, 100) – that variation is often systematically linked to extralinguistic factors such as speakers’ age, gender, race or ethnicity, socioeconomic status, and/or region as well as the context and form of their language use (see also article 6). In this orderly heterogeneity, historical linguists can often find evidence of language change spreading from, for example, generation to generation, region to region, socioeconomic class to class, or register to register (including from spoken to written language or vice versa and from informal to formal registers of spoken or written language or vice versa). The evidence lies in changing distributions or proportions of possible variants within a variant field over time.

As this description makes clear, the examination of frequency and co-occurrence of features lies at the heart of many of these historical studies. For this reason, computerized corpora have been a boon for historical sociolinguistic scholarship. Scholars working with historical corpora have studied the implementation of linguistic changes, not just the endpoints – in other words, they have used corpora to pay close attention to how a given language change spreads through the grammar and/or lexicon of a language, through a community of speakers, and/or through registers of a language over time. Historical corpora and measures of frequency and co-occurrence have also enhanced functional approaches to the study of language change, which highlight the role of frequency in, for example, functional change, shifts in markedness, and grammaticalization. There are limits to how much historical linguists can know from the available written evidence for earlier periods of any language, but historical corpora have encouraged scholars to explore the limits of what is discoverable.

What is the relationship of written evidence to “language change” if the unmarked meaning of *language* in linguistics is “the spoken language”? There are two different ways to answer this question. First, the development of written registers is an integral part of the history of a language, no matter the relationship of the written to the spoken. The written form of a language is central to an understanding of “language in use” in any literate speech community. Second, corpus-based work with early written texts makes certain assumptions about the relationship of spoken and written language. Of the possible assumptions, some seem less likely: for example, the spoken and written forms of a language are completely different entities; or the spoken and written forms of a language are entirely the same. Historical linguists often assume that the written language lags behind the spoken language in terms of evidence of language change,

particularly after standardization. That said, written conventions can affect the development of spoken registers, particularly in highly literate societies, and not all changes in the spoken or written language will ever appear in the other. Historical linguists also tend to assume spoken and written registers, while they may share much in grammar and lexicon, have distinct features at many levels: lexical, grammatical, discursive/stylistic. In medieval times, before widespread literacy and standardization, the written language may have been closer to the “spoken language written down”. Prescriptivism and written standards have often meant more divergence between the written and spoken language. But even in medieval times and certainly since, one cannot talk about the relationship between the written and spoken language without differentiating among registers. The content, purpose, audience, form, etc., of any register affects its linguistic form. Historical linguists typically assume that some less formal registers of written texts (e.g., personal letters) may be closer to the spoken vernacular; and while letters often contain, for example, interactional features (e.g., terms of address) and more oral language features according to the dimensions established by Biber/Finegan (see, e.g., Biber/Finegan 1997), letters also clearly have unique written stylistic conventions. Historical linguists can also exploit historical documents such as court records as well as fictional and dramatic dialogue for further evidence of the spoken language of a period. But in the end, in historical linguistic studies of language before the twentieth century, it is a given that when scholars describe the details of language change, they must work from the careful analysis of different types of written evidence to hypotheses about the spoken language.

2.2. Delineating a language’s “history”

The history of English is, of course, happening every day. The history of English should not and cannot be relegated to the past, to “pre-Shakespeare” or “pre-1900” or any such boundary. Because article 52 treats studies of recent changes specifically, this article focuses primarily on the evidence for earlier linguistic developments, developments witnessed only in written documents as they occurred before modern recording equipment. In the case of English, these linguistic changes range from Old English (449 CE – 1066 CE) through the late nineteenth or early twentieth century. Many of the changes witnessed in these periods, it must be noted, are still underway in Present Day English and/or are directly related to current changes in the language (e.g., the loss of grammatical case marking in earlier periods is related to the ongoing loss of case distinctions in the modern personal and interrogative pronouns). And as noted earlier, the study of variation across any language’s history has productively linked – or broken down an artificial barrier between – more synchronic and diachronic linguistics studies.

In article 52, Christian Mair notes that one of the difficulties of studying recent ongoing changes is that we cannot explain them in light of the known endpoint. On the other side, one of the dangers of studying earlier changes is that knowing the endpoint can result in downplaying or dismissing historical evidence that does not clearly lead to this endpoint. In history of English studies, the common endpoint has been Standard English.

Much has been written about the power of standard language ideology and Standard English ideology in particular (see, for example, Milroy/Milroy 1991; Milroy 1992; Lippi-

Green 1997; Bex/Watts 1999). The central presence of Standard English has shaped the story of English as it has been traditionally told. Features of Standard English have been traced backwards, and historical dialects closest to the developing standard have tended to receive more scholarly attention in comprehensive treatments of the history of English. The emphasis in historical corpus design on maximizing representativeness (within the very real constraints of the texts available – see article 14) or perhaps better phrased, maximizing diversity, including region, genre, and the social status/rank/mobility/gender/age/background of authors, offers the possibility of decentering, at least somewhat, more teleological approaches to the history of English that privilege Standard English over all other historical and present dialects. The emergence of smaller, more specialized historical corpora focused specifically on regional varieties (e. g., Scots) and on specific genres (e. g., correspondence) in particular promise to provide evidence that will help re-envision the scope and diverse variety encompassed by the “history of English” or the history of any given language.

3. Background considerations for collecting and evaluating historical evidence

This section addresses the kinds of evidence historical corpus linguistic studies use, some of the assumptions on which they rest, and the limitations they face.

3.1. A brief survey of corpora for English historical linguistic research

Corpus-based studies of the history of English have effectively exploited both general and more specialized corpora. Article 14 provides a more detailed overview of the corpora available as well as relevant issues of compilation. This section focuses on the kinds of evidence different corpora provide for research on the history of English.

The major breakthrough for corpus-based studies of the history of English came with the completion of the *Helsinki Corpus* in 1991, which allowed scholars to explore systematically changes from Old English through Early Modern English (for details on the corpus, see article 14; Kytö 1991a). The *Helsinki Corpus* remains a backbone of the field – a source for “benchmark” general results about overall historical trends that can be tested and supplemented by studies based on specialized corpora that provide more extensive coverage of a specific genre, of a historical period, of a dialect, etc.

Since 1991, many historical corpora have supplemented the *Helsinki Corpus*. *A Representative Corpus of Historical English Registers* (ARCHER) adopts a similar approach in terms of covering multiple genres, and it extends the historical coverage forward: 1650 – 1990. Other historical corpora have focused on particular genres: for example, correspondence/letters (e. g., the *Corpus of Early English Correspondence*, 1417–1681), newspapers (e. g., *Zurich English Newspapers Corpus*, 1661–1791), medical texts (e. g., the *Corpus of Early English Medical Writing*, 1375–1750), dialogues (e. g., the *Corpus of English Dialogues*, 1560–1760), and pamphlets (e. g., the *Lampeter Corpus*, 1640–1740). Others have included language from only one regional variety, such as the *Helsinki Corpus of Older Scots* (1450–1700). All these corpora have narrowed the historical range of

coverage from the scope of the *Helsinki Corpus*, to anywhere from a few decades to a few centuries. A handful of corpora with grammatical annotation (e.g., the *Penn-Helsinki Corpus of Middle English*, the *York-Helsinki Parsed Corpus of Old English Prose*) have been designed to facilitate morphosyntactic research – see article 14 for more details.

The largest electronic collections of historical texts available right now are not systematically compiled for linguistic research. They are not linguistic corpora in the technical sense. For example, the *Middle English Compendium* offers access to over sixty full-text electronic versions of Middle English works, as well as all of the material in the *Middle English Dictionary*; but it makes no claims to maximizing diversity or representativeness in terms of text types, region, authors, etc., and it continues to expand. *Early English Books Online* (EEBO) contains digital images of almost 125,000 texts from 1473 to 1700. Given the limitations of what is searchable in digital images, the University of Michigan and Oxford, with support of the international library community, have undertaken a project to keyboard and tag 25,000 of these works, with direct links to the digital versions. This massive electronic resource has been designed more with literary scholars, rather than linguists, in mind. However, given their relatively large size, these kinds of electronic text collections still offer a rich and valuable resource to historical linguists gathering evidence of language change, particularly of less frequent linguistic phenomena, and they can complement more traditional corpus-based studies (see Curzan/Palmer 2006). For example, historical linguists sometimes need to determine if particular constructions occur at all in a particular time period or are searching for examples of particular phenomena without being concerned in that phase of the study with frequency, in which case a large unstructured electronic collection will suffice. Researchers can also create their own smaller, stable corpora from these large text collections, targeting a particular genre, region, or author(s), calculating the size of the corpus overall for frequency statistics, and otherwise tailoring the corpus to suit their needs.

3.2. Focus of historical corpus-based research on English

Most corpus-based research on the history of English has focused on morphosyntactic developments. Computerized corpora arguably provide the most dramatic improvement in providing a more adequate amount of data for these kinds of studies. First, they require searching relatively large quantities of text to generate an adequate list of a feature as quotidian as, for example, subordinate clauses with conditional *if* or present perfect constructions with modals. Second, morphosyntactic developments often occur over significant periods of time (i.e., many centuries), which necessitates an adequate amount of data both at any one historical moment and over many centuries' worth of historical moments. Morphosyntactic features are also relatively readily and reliably searchable and of continuing historical and theoretical interest. Studies of lexical items and lexical fields have also been productive. Many corpus-based studies, however, usefully blur an easy distinction between the lexical and grammatical studies; as Bybee/Hopper (2001, 2) note in their introduction: “One especially important claim coming out of corpus studies is that the dividing line between grammar and lexicon, which has been virtually a dogma in linguistics, cannot be sustained”.

More recently, as corpora with written dialogue (from drama and fiction) and transcribed speech (particularly court records, including depositions) have become available, scholars have exploited corpora to examine discourse features and pursue questions in historical pragmatics. As Jacobs/Jucker (1995) point out, scholars interested in historical pragmatics have also come to recognize that written texts, as communicative acts in and of themselves, can be the focus of pragmatic analysis. Corpus-based register studies have taken written registers as their sole focus, examining historical register variation and development, and they have thus enhanced current understandings of the changing relationship of spoken and written language in the history of English.

Obviously missing from this summary is the study of historical phonology. Corpora may not prove to facilitate historical phonological studies particularly effectively for several reasons. To begin with, a more limited data set can often provide enough evidence for particular phonological developments. That said, Labov's uniformitarian principle maintains that historical sound changes would have spread through speech communities as current sound changes do, correlating with extralinguistic factors such as speakers' region, class/rank, age, gender, etc. It requires a relatively large data set to examine the effects of these factors on any language change, including historical phonological change. Second, once spelling becomes fairly well standardized in any language (in English, this begins to occur in the Early Modern English period), most printed texts reveal little about pronunciation variants. And even with documents that feature less standardized orthography (be they medieval texts or later personal letters), computers must be assisted to search effectively for spellings of interest. Wild card characters (e.g., *m*ht*) will gather some variants, but a researcher must intuit all forms of variants to search for. Alternately, corpus designers can lemmatize the corpus. In either case, the corpus would need to include editions that not only do not normalize the text but also preserve even the most idiosyncratic orthographic features, as these spellings can be evidence of pronunciation variants. To date, historical phonology has generally relied on more traditionally compiled evidence.

The rise of corpus-based historical linguistics has come on the heels of, if not in conjunction with, historical sociolinguistics. Suzanne Romaine's book *Sociohistorical Linguistics* (1982) is typically cited as the beginning of the field. Romaine works from the premise that linguistic change is affected by social factors in addition to linguistic constraints. Work in historical sociolinguistics has shaped the design of many historical corpora: the *Helsinki Corpus*, for example, tags every text for the author's age, gender, and rank (where known), as well as for the text's regional origins (where known) and genre. With these tags, studies can examine the effects of these sociolinguistic factors on linguistic change. The impetus for a corpus like the *Corpus of Early English Correspondence* (CEEC) was fundamentally sociolinguistic: letters are presumed to provide one of the more informal registers of the written language, one that is perhaps closer to the vernacular, and letters provide evidence of genuine communication between people of different genders, ages, and social ranks – all of which facilitates the systematic examination of social factors in language change. Just over twenty years after Romaine's book, Terttu Nevalainen and Helena Raumolin-Brunberg published the introductory text *Historical Sociolinguistics* (2003), which relies primarily on corpus-based studies for evidence of historical processes of variation and change.

The scope of the many types of corpus-based studies described above, in addition to encompassing more functional approaches to language study, extends over a broad defi-

nition of sociolinguistics, from variationist studies of the relationship of age, gender, region, and class to dialect features and change to studies in historical pragmatics to studies of register shifts. What they all share is an interest in the relationship of linguistic variation and change in the spoken and written manifestations of a language in use in given speech communities.

3.3. Methodological issues

The methodological issues involved in historical linguistic studies that exploit corpora often parallel those involved in compiling historical corpora (see article 14), with a few additional concerns relevant to scholars analyzing the data rather than those making it available.

Most importantly, a historical corpus can only be as thorough as the available texts. For the history of English, there are significant gaps in the evidence available. In general, the further back in time one goes, the less evidence has survived. The entire corpus of Old English texts is only about 3.5 million running words, and most of these texts are in formal genres: religious texts, historical chronicles, literature, etc. The dearth of texts that might capture more colloquial English continues through the Middle English period, and the CEEC is so valuable because it compiles many of the available correspondence documents, beginning near the end of the Middle English period through the nineteenth century. In the history of English, there is also the gaping gap spanning the eleventh and twelfth centuries from which there are almost no surviving documents of any kind written in English.

Historical corpora also face the challenge of limited information about relevant socio-linguistic factors for early texts. Many if not most medieval texts are anonymous. *The Linguistic Atlas of Late Mediaeval English* (McIntosh et al. 1987) did pioneering work in localizing Middle English texts, but the provenance of medieval texts is not always certain. With later texts, for which scholars can often find biographical information about authors, sociolinguistic classification has been a central issue. For example, Nevalainen/Raumolin-Brunberg (2003) have argued for rank and/or social and regional mobility as more appropriate classifications for the Early Modern English period than the modern sociolinguistic category “class”.

The available evidence in early corpora then poses the challenge of significant spelling variation. For example, in Old English, the third-person plural pronoun can be spelled *hie* and *hi* (to name just two variants of many), both of which can also be spelling variants of oblique forms of the feminine singular third-person pronoun. The auxiliary verb *might* has at least twelve different spellings in Middle English. As a result, historical corpus linguists must design searches carefully so that computers can find all the relevant evidence, if possible; or corpus builders must build corpora that appropriately address the issue of spelling variation. Texts can be normalized or taken from normalized editions; however, normalized texts introduce the real possibility of losing critical evidence to be found in the variation. Corpora can also be lemmatized so that a search for a given headword pulls all variants spellings. Corpus designers can rely on historical dictionaries to provide many spelling variants, but even they may miss some orthographic idiosyncrasies in any given text. And in the end, linguists must sort through the com-

puter-generated results to determine which hits are relevant. Computers may not get tired like humans do, but they also do not possess humans' intuition or analytic ability to handle the variability of many written medieval languages.

4. Corpus-based evidence for historical linguistic developments in English

Through a survey of selected historical studies of the English language, this section highlights some of what corpus-based evidence has revealed about the history of English and about language change more generally. Within the space constraints, this survey can capture only a small part of the range of studies available and must treat most studies in a highly summary fashion. Each subsection features, in addition to a broader survey, one more detailed case study to provide a sense of scholars' approaches to specific questions, their new discoveries about English, and the implications of these discoveries for theories of language change.

4.1. Tracing morphosyntactic change

As mentioned above, historical corpora have been particularly effective for research on morphosyntactic developments in English. Linguists continue to learn more about the progression of the loss of inflectional endings in English and the overall shift toward a more analytic syntax; the grammaticalization of particular forms; and developments in functional word classes such as the pronouns and auxiliary verbs. Corpus-based evidence has provided an ever more detailed picture of the progression of the structural changes that have distinguished English from other Germanic languages.

Corpus-based studies, for example, have described in new detail the chronology and both the structural and sociolinguistic factors involved in the restructuring of parts of the English pronoun system. Specifically, studies have supplied new evidence on the replacement of subject *ye* by *you*, which seems to have been a relatively rapid process, with a steep S-curve of change from 1520–1600 (Nevalainen/Raumolin-Brunberg 2003). Nevala (2002) uses private letters between family members and friends from the CEEC to examine the seventeenth-century increase in writers' use of *thou* – both in terms of the number of writers employing the pronoun and the diversity of the recipients (including women to their husbands) – before the form came to appear almost exclusively in Biblical and poetic contexts in the eighteenth century; Walker (2007) examines the dramatic changes in use of these second-person pronouns in English dialogues from 1560–1760. The eventual replacement of *mine/thine* by *my/thy* follows different patterns of change between 1500 and 1620 depending on whether the possessive pronoun precedes a consonant, a vowel, /h/, or the word *own* (Nevalainen/Raumolin-Brunberg 2003). Other studies have focused on the rise of *-one* and *-body* indefinite pronouns (e. g., Raumolin-Brunberg/Kahlas-Tarkka 1997); the diffusion of *its* in the seventeenth century to replace *thereof* and the more archaic *his* (Nevalainen/Raumolin-Brunberg 1994); the marginalization of *the which* by *which* by the middle of the sixteenth century (Nevalainen 1996); and the rise of the prop-word *one* (Raumolin-Brunberg/Nurmi 1997).

Historical changes in verb forms have been the focus of much scholarly attention. On the replacement of third-person singular present tense ending *-th* by *-s*, studies such as Stein (1987), Kytö (1993), and Nevalainen/Raumolin-Brunberg (2003) have located the dramatic curve of change from the mid-sixteenth to mid-seventeenth century and have examined the role of factors such as register and the gender and social rank of the writer in the progression of the change. The modern form seems to spread from the northern parts of England, as does the replacement of *be* by *are* in the third-person plural (Nevalainen 1996). English modals have undergone significant changes, from the encroachment of *can* into the territory of *may* (cf. Kytö 1991b; Facchinetto 2003), the increasing use of *must* in both deontic and epistemic uses (cf. Nurmi 2003), the competition of *shall* and *will* (cf. Kytö 1991b; Rissanen 2000), just to name a few. The rise and fall of different roles of periphrastic *do* have been charted (cf. Rissanen 1991; Nurmi 1996, 1999); and as Raumolin-Brunberg/Nurmi (1997) demonstrate, more data (perhaps especially from more vernacular sources) can critically disrupt any neat patterns in the development of periphrastic *do*. Working with material from the CEEC, Raumolin-Brunberg/Nurmi show that the decline of affirmative *do* in the seventeenth century looks more like a rise-fall-rise, which may be in part due to competition with the incoming progressive forms (Nurmi 1996).

Progressive forms have themselves been the subject of detailed study. For example, Fitzmaurice (2004) draws on the *Network of Eighteenth-century English Texts* to extend her own previous scholarship on what she terms the experiential and/or subjective progressive construction and consider the effects of peer group on the usage of such constructions. In a response to this study, Smitterberg (2004) concludes that experiential or subjective readings correlate with genres in which writers can express themselves in a speech-like manner. Smitterberg (2005) provides an extended study of the development of the progressive in the nineteenth century which demonstrates, among many other things, the critical importance of genre/register to studying the integration of the progressive. At the theoretical level, Smitterberg's analysis of the corpus data leads to a productive differentiation of integration and grammaticalization, both of which are integrally tied to frequency.

Other corpus-based studies address developments at the level of the clause. For example, Meurman-Solin (2002) charts the increase in subordinators of posteriority and anteriority in Scots, 1450–1700.

Importantly, corpus data can often surprise, challenging linguists' intuitions and larger historical trends. For example, given the overall shift from a more synthetic to a more analytic syntax in the history of English, one would expect the periphrastic forms of comparative and superlative adjectives (e.g., *more gentle*, *most stupid*) to be overtaking the inflected forms (e.g., *gentler*, *stupidest*). However, Kytö/Romaine (2000) demonstrate that the inflected forms have been reasserting themselves since the Early Modern English period – with British English leading American English – and inflected forms constitute the majority in current forms of both varieties. The use of corpora also allows Kytö/Romaine (1997) to study systematically the adjectives that tend to favor uninflected forms (e.g., *-ful*, *-ous*) and those that favor inflected forms (e.g., *-y/-ly*, *-le*).

4.2. Tracing lexical change, semantic change, and grammaticalization

Corpus-based historical studies of the lexicon have focused both on single words and on lexical fields. For example, Heikkinen/Tissari (2002) examine the semantics of the Old

English noun *bliss* alongside the semantics of the adjective *happy* in Early Modern and Modern English. Whereas *bliss* appears almost entirely in religious contexts in Old English, *happy* seems to undergo secularization in its meaning from early modern times to the present, with the social and metaphysical domains being replaced by more personal and material ones. Curzan (2003) describes shifts within the lexical fields of words for men, women, boys, and girls focusing particularly on conjoined phrases such as ‘man and woman’. In his study of the development of “bad language” in English from 1586 to the present, McEnery (2006) combines more traditional uses of corpora – the study relies on corpora to track swear words and other taboo terms over the centuries – and highly innovative ones: the study also exploits a specialized corpus of commentary on swearing, and through the examination of collocational patterns and frequency of particular terms in these texts, McEnery provides insight into the evolution of the discourse of purity, which has historically taken bad language as one target. As this study demonstrates, historical corpora can enhance the understanding both of internal developments in the language and of the social context in which these developments occur.

As noted above, corpus studies have importantly blurred the dividing line between lexicon and grammar, and of particular theoretical interest are the studies of lexical items that seem to have undergone grammaticalization. Corpus-based approaches to grammaticalization, given the centrality of frequency, collocational patterns, and the interaction of lexicon and grammar to this process, have been strikingly productive. Recently, there have been book-length treatments, based on corpus-based studies, of constructions undergoing grammaticalization such as emerging modals like *gotta*, *gonna*, and *wanna* (Krug 2000) and complex prepositions like *in view of* (Hoffmann 2005); and the nine articles collected in Lindquist/Mair (2004) all focus on corpus approaches to grammaticalization in English. Individual studies over the past ten years have focused on, for example, the pronominalization of *one* (Rissanen 1997); the rise and then decline of *there-* compounds as they come to be replaced by prepositional phrases (Österman 1997); the grammaticalization of self-reflexive pronouns (Peitsara 1997) and of *methinks* as an expression of evidentiality (Palander-Collin 1996); the grammaticalization of a written phrase such as *videlicet* in court texts (Moore 2004); and the restriction of *wit* to a specialized appositive linking use in *to wit* (Koivisto-Alanko/Rissanen 2002). In the case of the semantic changes that *wit* undergoes, searches of the Helsinki Corpus dramatically demonstrate the interaction of the rise of *know* in Middle English and the rapid decrease of *wit* in the sense of ‘know’ in the sixteenth and seventeenth centuries (see Tables 51.1 and 51.2).

Corpus-based studies such as this, that juxtapose information about semantic change with data about frequency and distribution, can capture the interactional dynamics of

Tab. 51.1: *Wit* and *know* in the Middle English sub-sections of the *Helsinki Corpus* (affirmative simplex forms only). Reproduced with permission from Koivisto-Alanko/Rissanen (2002, 18)

	<i>wit</i>	(/10,000 words)	<i>know</i>	(/10,000 words)
ME1 (1150–1250)	129	(11.4)	12	(1.1)
ME2 (1250–1350)	96	(9.8)	47	(4.8)
ME3 (1350–1420)	244	(13.2)	278	(15.1)
ME4 (1420–1500)	150	(7.0)	265	(12.4)

Tab. 51.2: *Wit* and *know* in the Early Modern English sub-sections of the *Helsinki Corpus*. Reproduced with permission from Koivisto-Alanko/Rissanen (2002, 18)

	<i>wit</i>	(/10,000 words)	<i>know</i>	(/10,000 words)
E1 (1500–1570)	39	(2.1)	333	(17.5)
E2 (1570–1640)	10	(0.7)	451	(23.7)
E3 (1640–1710)	7	(0.4)	344	(20.0)

shifts in the lexicon in ways that even the best historical dictionary such as the *Oxford English Dictionary* cannot. They also demonstrate the importance of examining chunks of language longer than the individual word, as the frequency of co-occurrence of linguistic forms can be as critical to language change as the frequency of any given form.

4.3. Tracing change at the level of discourse

In recent years, historical pragmatic studies have begun to employ corpus-based methodologies, complementing the historical work on the development of written registers. Given the limited information available in corpora of written texts about spoken features, scholars must often create or restrict studies to more specialized (sub)corpora. Moore (2003) employs a corpus of early modern slander depositions to examine written strategies for introducing reported direct speech. Jucker (2002) provides a historical look at five discourse markers, drawing on data from plays, fiction, and trial records in the *Helsinki Corpus*, describing the process of pragmatalization as well as the relative frequencies of each discourse marker. The data for the period 1500–1710 show a steady increase in the frequency of *well* and *why*, a dramatic increase in the use of *o/oh* and *pray/prithee*, and a steady decrease in the use of *marry*.

Palander-Collin (1999) examines male and female discourse styles in the CEEC and provides evidence to support the assumption in much modern language and gender scholarship of identifiable stylistic differences between genders. Palander-Collin's data indicate that in these historical letters, women tend to focus on the affective functions of an interaction and use linguistic devices that stress solidarity more often than men.

Raumolin-Brunberg (1996) uses data about forms of address in the CEEC to argue that the growth of literacy and privacy seems to coincide with increasing use of positive politeness strategies, in this case forms of address in letters such as the extended use of *sir*, *madam*, *Mrs.* and *Mr.* through various ranks of addressees. Looking also at social roles and author-addressee relationships, Palander-Collin (2002) examines how these factors play into one author's use of first-person pronouns and verbs of different types (activity, mental, communicative) and shows how much systematic variation there is within one speaker's patterns of discourse.

The historical development of written registers has been a particularly rich area of study, pioneered by Douglas Biber and Edward Finegan. The ARCHER corpus has been the source of many such studies. For example, Biber/Finegan (1997) describe the increasing differentiation both in terms of style and intended audience of fiction and news texts from medical, science, and legal prose, from the seventeenth century through the present. As fiction and news have become more popular registers, they have reversed

an early trend toward more “literate” styles and have adopted many more “oral” characteristics. Medical, science, and legal texts, which have become highly specialized registers, have followed a steady development toward more “literate” styles with little to no narrative, for example. Atkinson (1996, 1999) provides additional evidence of the shift from author-centered rhetoric to highly abstract forms of language in scientific discourse. Recent studies based on the *Corpus of Early English Medical Writings* have described specific language changes within the medical register, often in conjunction with the history of science, to analyze shifts in scientific thought (cf. Taavitsainen et al. 2002). The evidence in these studies effectively demonstrates that the history of written English is the history of registers and that these registers show shifting relationships to the more oral style that characterizes at least less formal registers of spoken language. Historical corpus linguistics encourages language historians to consider all this variation, both spoken and written, as part of the history of a language – not as “a language” in abstraction but as a language as it is used.

4.4. Testing variationist sociolinguistic principles

Many of the morphosyntactic studies described above examine sociolinguistic factors involved in the particular linguistic change, such as the social rank and gender of the writer. Studies such as Nurmi (1996) and many of those covered in Nevalainen/Rau-molin-Brunberg (2003) confirm that social background seems to be a relevant factor in the spread of language change in earlier historical periods, as it is in the present. As discussed above, questions remain about how best to describe social classifications as well as at what points in the diffusion of a change (e.g., incipient stages, robust periods of change) social status plays a relevant role.

Historical studies of gender as a factor have both confirmed and challenged modern findings. Two principles about the role of women in language change have guided much modern sociolinguistic research: (1) Women tend to favor incoming prestige forms more than men in change from above given their higher sensitivity to standards; (2) Women tend to be innovators in change from below. Of the morphosyntactic developments summarized above, Nevalainen/Rau-molin-Brunberg (2003) list the following as changes led by women: the rise of *you*, *my/thy*, and *its*; the use of *-s* instead of *-th*; the use of *do* in negative statements; and the use of *which*. All this historical evidence seems to support the hypothesis that women lead in changes from below. However, other results challenge the historical extension of the principle that women favor prestige forms in change from above. As Nevalainen (2000) demonstrates, in the early sixteenth century, when multiple negation was gradually being replaced by single negation in Standard English, men promoted the change (see Figure 51.1). Nevalainen hypothesizes that this development represents a change from above, promoted by, for example, legal language (cf. Rissanen 2000); given women’s limited educational and professional opportunities in Tudor and Stewart England, they were not positioned to lead in changes that originated or spread primarily from this sphere.

General theoretical principles of language change can only be stronger for being tested against both modern and historical evidence.

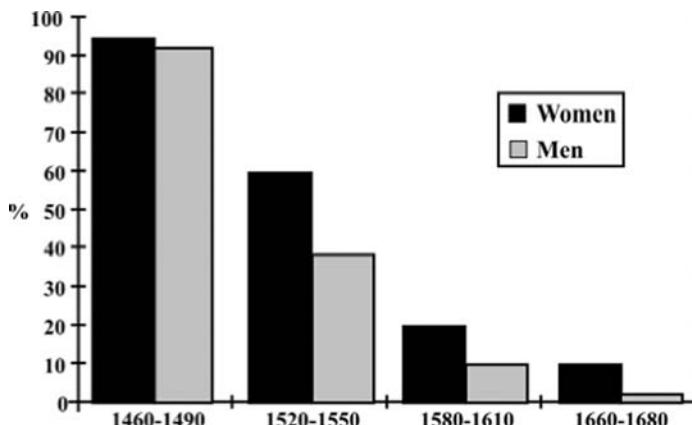


Fig. 51.1: The frequency of multiple negation according to gender in four periods: 1460–1490, 1520–1550, 1580–1610, and 1660–1680 (*Corpus of Early English Correspondence* 1998 and supplement, means of period totals). Reproduced with permission from Nevalainen (2000)

4.5. Tracing the process of standardization

Well-designed historical corpora allow scholars to explore beyond Standard English, both through the inclusion of nonstandard dialects in more comprehensive corpora and through the creation of corpora focused specifically on regional varieties beyond standard British and American English. Historical corpora also hold new insights on the rise of Standard English: the sources of particular forms and the driving social forces influencing change. And as Meurman-Solin (2000) points out, studies of the history of Scots point to additional factors in the standardization of varieties beyond British English, such as Scotland's political and sociocultural independence.

Some of the central work on the history of Standard English was carried out by John Hurt Fisher (see his collected essays in Fisher 1996) and Malcolm Richardson (1980), establishing Chancery English as one of the main sources of written Standard English. Recent corpus work has offered productive follow-up studies. For example, Heikkonen (1996) provides a case study of the Signet letters (1417–1422) – a subcorpus of the CEEC – to establish that Signet English was a model for Standard English. Rissanen (2000) turns to early statutory texts to examine their influence on the written standard; he argues that while legal texts may have provided the model for single negation rather than multiple negation and introduced the phrase *provided that* into other genres, other written genres continued to favor *shall* over *will* – an indication that the standard did not always follow change from above from formal written texts, such as legal documents.

Taavitsainen (2000) compares spellings in the *Early English Corpus of Medical Writing* to Chancery forms and to Samuels's description of the Central Midlands variety and discovers that the Central Midlands variety pervades the corpus. She comes to the provocative conclusion that Wycliffite writings may have taken scientific texts as their written model, rather than vice versa. Studies such as these demonstrate the power of corpus-based, data-driven studies to revise traditional understandings of central developments in the history of English.

5. Conclusion

At the language-specific level, current knowledge about the history of particular languages such as English is significantly richer for the evidence described and analyzed in historical corpus-based studies. From morphosyntactic developments to changes in register to the development of standard varieties, scholars have enhanced current understandings of the chronology of linguistic developments, the structural details, and the sociolinguistic factors involved. At a theoretical level, historical corpora have allowed scholars to test linguistic principles about language change and come to new understandings about the role of different kinds of factors (e.g., functional, structural, extralinguistic) in language change. Historical corpora have also challenged the field to reconsider the relationship of the written language to the history of “the language”, as scholars explore the many ways to examine, interpret, and exploit historical written evidence.

With all this technological richness, it is critical not to lose sight of the importance of complementary studies: the combination of more quantitative corpus-based studies with close, more qualitative examinations of full texts (vs. the extracts in many corpora); and careful work with other historical resources as a way to provide sociolinguistic context for changes in language use. Keene (2000), for example, persuasively argues for how much historical linguists can learn from medieval evidence about commerce and migration patterns – information that must inform the analysis of linguistic data from the period. And historical linguists should preserve contact with original manuscripts, which can never be fully replaced by even good editions, particularly for the qualitative analysis of the implementation of a language change that can complement large quantitative studies. Writers, from medieval scribes to more modern letter writers, can send linguistic insights from the past through word spacing and alignment, idiosyncratic and ambiguous spellings, glosses, corrected text, and the other human touches that exist in hand-written documents.

Historical corpora offer exciting new possibilities for how linguists can examine and interpret large quantities of evidence from systematic collections of written texts – as well as smaller quantities of evidence from more targeted, yet still electronic and fully searchable collections of written texts. Full original manuscripts offer invaluable qualitative evidence of linguistic forms in context, which can provide an even more detailed picture of the implementation of a language change. Historical records offer important evidence about contemporaneous social, cultural, and economic contexts that must inform the analysis of all linguistic evidence for language change. Corpora provide even more tools, rather than replace tools, in the historical linguist’s toolbox.

6. Literature

- Atkinson, Dwight (1996), *The Philosophical Transactions of the Royal Society of London, 1675–1975: A Sociohistorical Discourse Analysis*. In: *Language in Society* 25(3), 333–371.
- Atkinson, Dwight (1999), *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Mahwah, NJ/London: Lawrence Erlbaum.
- Bex, Tony/Watts, Richard J. (1999), *Standard English: The Widening Debate*. London/New York: Routledge.

- Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge/New York: Cambridge University Press.
- Biber, Douglas/Finegan, Edward (1997), Diachronic Relations among Speech-based and Written Registers in English. In: Nevalainen, T./Kahlas-Tarkka, L. (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 253–275.
- Bybee, Joan/Hopper, Paul (2001), *Frequency and the Emergence of Linguistic Structure*. Amsterdam/Philadelphia: John Benjamins.
- Curzan, Anne (2003), *Gender Shifts in the History of English*. Cambridge: Cambridge University Press.
- Curzan, Anne/Emmons, Kimberly (eds.) (2004), *Studies in the History of the English Language II: Unfolding Conversations*. Berlin/New York: Mouton de Gruyter.
- Curzan, Anne/Palmer, Chris C. (2006), The Importance of Historical Corpora, Reliability, and Reading. In: Facchinetto, R./Rissanen, M. (eds.), *Corpus-based Studies in Diachronic English*. Bern: Peter Lang, 17–34.
- Facchinetto, Roberta (2003), The Modal ‘Can’ in EModE Translations of *De Consolatione Philosophiae*. Paper presented at the 24th International ICAME conference. Guernsey, April 2003.
- Fisher, John Hurt (1996), *The Emergence of Standard English*. Lexington: University of Kentucky Press.
- Fitzmaurice, Susan (2004), The Meanings and Uses of the Progressive Construction in an Early Eighteenth-century English Network. In: Curzan, A./Emmons, K. (eds.), *Studies in the History of the English Language II: Unfolding Conversations*. Berlin/New York: Mouton de Gruyter, 131–173.
- Heikkilä, Kanerva/Tissari, Heli (2002), *Gefeo and Geblissa or Happy Birthday! On Old English Bliss and Modern English Happy*. In: Raumolin-Brunberg, H./Nevala, M./Nurmí, A./Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 59–76.
- Heikkonen, Kirsi (1996), Regional Variation in Standardization: A Case Study of Henry V’s Signet Office. In: Nevalainen, T./Raumolin-Brunberg, H. (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam/Atlanta: Rodopi, 111–127.
- Hoffmann, Sebastian (2005), *Grammaticalization and English Complex Prepositions: A Corpus-based Study*. London/New York: Routledge.
- Jacobs, Andreas/Jucker, Andreas H. (1995), The Historical Perspective in Pragmatics. In: Jucker, A. H. (ed.), *Historical Pragmatics: Pragmatic Developments in the History of English*. Amsterdam/Philadelphia: John Benjamins, 3–33.
- Jucker, Andreas H. (2002), Discourse Markers in Early Modern English. In: Watts, R./Trudgill, P. (eds.), *Alternative Histories of English*. London/New York: Routledge, 210–230.
- Keene, Derek (2000), Metropolitan Values: Migration, Mobility, and Cultural Norms, London 1100–1700. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 93–114.
- Koivisto-Alanko, Päivi/Rissanen, Matti (2002), *We Give You to Wit: Semantics and Grammaticalisation of the Verb Wit in the History of English*. In: Raumolin-Brunberg, H./Nevala, M./Nurmí, A./Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 13–32.
- Krug, Manfred (2000), *Emerging English Modals: A Corpus-based Study of Grammaticalization*. Berlin/New York: Mouton de Gruyter.
- Kytö, Merja (compiler) (1991a), *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Helsinki: Department of English, University of Helsinki.
- Kytö, Merja (1991b), *Variation and Diachrony, with Early American English in Focus: Studies on CAN/MAY and SHALL/WILL*. Frankfurt: Lang.

- Kytö, Merja (1993), Third-person present singular verb inflection in Early British and American English. In: *Language Variation and Change* 5, 113–139.
- Kytö, Merja/Romaine, Suzanne (1997), Competing Forms of Adjective Comparison in Modern English: What Could be More Quicker and Easier and More Effective? In: Nevalainen, T./Kahlas-Tarkka, L. (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 329–352.
- Kytö, Merja/Romaine, Suzanne (2000), Adjective Comparison and Standardization Processes in American and British English from 1620 to the Present. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 171–194.
- Lindquist, Hans/Mair, Christian (eds.) (2004), *Corpus Approaches to Grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins.
- Lippi-Green, Rosina (1997), *English with an Accent: Language, Ideology, and Discrimination in the United States*. London/New York: Routledge.
- McEnery, Tony (2006), *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London/New York: Routledge.
- McIntosh, Angus/Samuels, M. L./Benskin, Michael/Laing, Margaret/Williamson, Keith (1987), *Linguistic Atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.
- Meurman-Solin, Anneli (2000), Change from Above or from Below? Mapping the Loci of Linguistic Change in the History of Scottish English. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 155–170.
- Meurman-Solin, Anneli (2002), Simple and Complex Grammars: The Case of Temporal Subordinators in the History of Scots. In: Raumolin-Brunberg, H./Nevala, M./Nurmi, A./Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 187–210.
- Milroy, James (1992), *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
- Milroy, James/Milroy, Lesley (1991), *Authority in Language: Investigating Language Prescription and Standardisation*. 2nd ed. London/New York: Routledge.
- Minkova, Donka/Stockwell, Robert (eds.) (2003), *Studies in the History of the English Language: A Millennial Perspective*. Berlin/New York: Mouton de Gruyter.
- Moore, Colette (2003), Reporting Direct Speech in Early Modern Slander Depositions. In: Minkova, D./Stockwell, R. (eds.), *Studies in the History of the English Language: A Millennial Perspective*. Berlin/New York: Mouton de Gruyter, 399–416.
- Moore, Colette (2004), Representing Speech in Early English. Ph.D. dissertation, University of Michigan, Ann Arbor, Michigan.
- Nevala, Minna (2002), *Youre moder send a letter to the: Pronouns of Address in Private Correspondence from Late Middle to Late Modern English*. In: Raumolin-Brunberg, H./Nevala, M./Nurmi, A./Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 135–159.
- Nevalainen, Terttu (1996), Social Stratification. In: Nevalainen, Terttu/Raumolin-Brunberg, Helena (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam/Atlanta: Rodopi, 57–76.
- Nevalainen, Terttu (2000), Gender Differences in the Evolution of Standard English: Evidence from the *Corpus of Early English Correspondence*. In: *Journal of English Linguistics* 28(1), 38–59.
- Nevalainen, Terttu (2002), Women's Writing as Evidence for Linguistic Continuity and Change in Early Modern English. In: Watts, R./Trudgill, P. (eds.), *Alternative Histories of English*. London/New York: Routledge, 191–209.
- Nevalainen, Terttu/Raumolin-Brunberg, Helena (1994), *Its Strength and the Beauty of It: The Standardization of the Third Person Neuter Possessive in Early Modern English*. In: Stein, D./Tieken-Boon van Ostade, I. (eds.), *Towards a Standard English, 1600–1800*. Berlin/New York: Mouton de Gruyter, 171–216.

- Nevalainen, Terttu/Raumolin-Brunberg, Helena (2003), *Historical Sociolinguistics*. London: Longman.
- Nurmi, Arja (1996), Periphrastic Do and BE + ING: Interconnected Developments? In: Nevalainen, T./Raumolin-Brunberg, H. (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam/Atlanta: Rodopi, 151–165.
- Nurmi, Arja (1999), *A Social History of Periphrastic Do*. (Mémoires de la Société Néophilologique 56.) Helsinki: Société Néophilologique.
- Nurmi, Arja (2003), The Role of Gender in the Use of MUST in Early Modern English. In: Granger, S./Petch-Tyson, S. (eds.), *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. (Language and Computers: Studies in Practical Linguistics 48.) Amsterdam/Atlanta: Rodopi, 111–120.
- Österman, Aune (1997), There compounds in the History of English. In: Rissanen, M./Kytö, M./Heikkonen, K. (eds.), *Grammaticalization at Work: Studies of Long-term Developments in English*. Berlin/New York: Mouton de Gruyter, 191–276.
- Palander-Collin, Minna (1996), The Rise and Fall of METHINKS. In: Nevalainen, T./Raumolin-Brunberg, H. (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam/Atlanta: Rodopi, 131–149.
- Palander-Collin, Minna (1999), *Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English*. (Mémoires de la Société Néophilologique 55.) Helsinki: Société Néophilologique.
- Palander-Collin, Minna (2002), Tracing Patterns of Interaction in Historical Data. In: Raumolin-Brunberg, H./Nevala, M./Nurmi, A./Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 117–134.
- Peitsara, Kirsti (1997), The Development of Reflexive Strategies in English. In: Rissanen, M./Kytö, M./Heikkonen, K. (eds.), *Grammaticalization at Work: Studies of Long-term Developments in English*. Berlin/New York: Mouton de Gruyter, 277–370.
- Raumolin-Brunberg, Helena (1996), Forms of Address in Early English Correspondence. In: Nevalainen, T./Raumolin-Brunberg, H. (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam/Atlanta: Rodopi, 167–181.
- Raumolin-Brunberg, Helena/Kahlas-Tarkka, Leena (1997), Indefinite Pronouns with Singular Human Reference. In: Rissanen, M./Kytö, M./Heikkonen, K. (eds.), *Grammaticalization at Work: Studies of Long-term Developments in English*. Berlin/New York: Mouton de Gruyter, 17–85.
- Raumolin-Brunberg, Helena/Nurmi, Arja (1997), Dummies on the Move: Prop-ONE and Affirmative DO in the 17th Century. In: Nevalainen, T./Kahlas-Tarkka, L. (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 395–417.
- Richardson, Malcolm (1980), Henry V, the English Chancery and Chancery English. In: *Speculum* 55(4), 726–750.
- Rissanen, Matti (1991), Spoken Language and the History of Do-periphrasis. In: Kastovsky, D. (ed.), *Historical English Syntax*. (Topics in English Language 2.) Berlin/New York: Mouton de Gruyter, 321–342.
- Rissanen, Matti (1997), The Pronominalization of One. In: Rissanen, M./Kytö, M./Heikkonen, K. (eds.), *Grammaticalization at Work: Studies of Long-term Developments in English*. Berlin/New York: Mouton de Gruyter, 87–143.
- Rissanen, Matti (1998), Towards an Integrated View of the Development of English: Notes on Causal Linking. In: Fisiak, J./Krygier, M. (eds.), *Advances in English Historical Linguistics*. (Trends in Linguistics 112.) Berlin/New York: Mouton de Gruyter, 389–406.
- Rissanen, Matti (2000), Standardisation and the Language of Early Statutes. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 117–130.
- Romaine, Suzanne (1982), *Sociohistorical Linguistics*. Cambridge: Cambridge University Press.

- Samuels, M. L. (1989/1963), Some Applications of Middle English Dialectology. In: Laing, M. (ed.), *Middle English Dialectology: Essays on Some Principles and Problems*. Aberdeen: Aberdeen University Press, 64–80.
- Smitherberg, Erik (2004), Investigating the Expressive Progressive: On Susan Fitzmaurice's "The Meanings and Uses of the Progressive Construction in an Early Eighteenth-century English Network". In: Curzan, A./Emmons, K. (eds.), *Studies in the History of the English Language II: Unfolding Conversations*. Berlin/New York: Mouton de Gruyter, 175–182.
- Smitherberg, Erik (2005), *The Progressive in 19th-century English: A Process of Integration*. (Language and Computers: Studies in Practical Linguistics 54.) Amsterdam/New York: Rodopi.
- Stein, Dieter (1987), At the Crossroads of Philology, Linguistics, and Semiotics: Notes on the Replacement of *th* by *s* in the Third Person Singular in English. In: *English Studies* 5, 406–415.
- Taavitsainen, Irma (2000), Scientific Language and Spelling Standardisation 1375–1550. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 131–154.
- Taavitsainen, Irma/Pahta, Päivi/Leskinen, Noora/Ratia, Maura/Suhr, Carla (2002), Analysing Scientific Thought-styles: What Can Linguistic Research Reveal about the History of Science? In: Raumolin-Brunberg, H./Nevala, M./Nurmi, A./Rissanen, M. (eds.), *VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 251–280.
- Walker, Terry (2007), *Thou and You in Early Modern English Dialogues*. (Pragmatics & Beyond New Series 158.) Amsterdam/Philadelphia: John Benjamins.
- Weinreich, Uriel/Labov, William/Herzog, Marvin I. (1968), Empirical Foundations for a Theory of Language Change. In: Lehmann, W. P./Malkiel, Y. (eds.), *Perspectives on Historical Linguistics*. (Current Issues in Linguistic Theory 24.) Amsterdam/Philadelphia: John Benjamins, 95–195.
- Wright, Laura (2000), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press.

Anne Curzan, Ann Arbor, MI (USA)

52. Corpora and the study of recent change in language

1. Introduction
2. "Real-time" and "apparent-time" approaches in the study of recent and ongoing change
3. Some representative corpus-based work on recent change in English grammar and vocabulary
4. Conclusion and outlook: A plea for methodological pluralism
5. Acknowledgement
6. Literature

1. Introduction

As far as basic research design is concerned, corpus-based work on recent change in language follows the same principles as all historical corpus linguistics. However, the fact that much common ground is shared does not mean that there are not also impor-

- Samuels, M. L. (1989/1963), Some Applications of Middle English Dialectology. In: Laing, M. (ed.), *Middle English Dialectology: Essays on Some Principles and Problems*. Aberdeen: Aberdeen University Press, 64–80.
- Smitherberg, Erik (2004), Investigating the Expressive Progressive: On Susan Fitzmaurice's "The Meanings and Uses of the Progressive Construction in an Early Eighteenth-century English Network". In: Curzan, A./Emmons, K. (eds.), *Studies in the History of the English Language II: Unfolding Conversations*. Berlin/New York: Mouton de Gruyter, 175–182.
- Smitherberg, Erik (2005), *The Progressive in 19th-century English: A Process of Integration*. (Language and Computers: Studies in Practical Linguistics 54.) Amsterdam/New York: Rodopi.
- Stein, Dieter (1987), At the Crossroads of Philology, Linguistics, and Semiotics: Notes on the Replacement of *th* by *s* in the Third Person Singular in English. In: *English Studies* 5, 406–415.
- Taavitsainen, Irma (2000), Scientific Language and Spelling Standardisation 1375–1550. In: Wright, L. (ed.), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press, 131–154.
- Taavitsainen, Irma/Pahta, Päivi/Leskinen, Noora/Ratia, Maura/Suhr, Carla (2002), Analysing Scientific Thought-styles: What Can Linguistic Research Reveal about the History of Science? In: Raumolin-Brunberg, H./Nevala, M./Nurmi, A./Rissanen, M. (eds.), *VARIENG Studies on English for Terttu Nevalainen*. Helsinki: Société Néophilologique, 251–280.
- Walker, Terry (2007), *Thou and You in Early Modern English Dialogues*. (Pragmatics & Beyond New Series 158.) Amsterdam/Philadelphia: John Benjamins.
- Weinreich, Uriel/Labov, William/Herzog, Marvin I. (1968), Empirical Foundations for a Theory of Language Change. In: Lehmann, W. P./Malkiel, Y. (eds.), *Perspectives on Historical Linguistics*. (Current Issues in Linguistic Theory 24.) Amsterdam/Philadelphia: John Benjamins, 95–195.
- Wright, Laura (2000), *The Development of Standard English, 1300–1800*. Cambridge: Cambridge University Press.

Anne Curzan, Ann Arbor, MI (USA)

52. Corpora and the study of recent change in language

1. Introduction
2. "Real-time" and "apparent-time" approaches in the study of recent and ongoing change
3. Some representative corpus-based work on recent change in English grammar and vocabulary
4. Conclusion and outlook: A plea for methodological pluralism
5. Acknowledgement
6. Literature

1. Introduction

As far as basic research design is concerned, corpus-based work on recent change in language follows the same principles as all historical corpus linguistics. However, the fact that much common ground is shared does not mean that there are not also impor-

tant differences on points of detail and emphasis. First, more so than historical linguists working on remoter stages in the history of languages, students of recent and ongoing change will find themselves working in the intersection of two scholarly traditions: historical linguistics, which – whatever the theoretical foundation of a particular approach may be – is usually informed by a consistently diachronic and usually long-term perspective, and sociolinguistic variationism, with its twin emphasis on recent and short-term developments and on the integrated description of synchronic variation and diachronic change. Secondly, the amount and quality of the data is different. For most languages there is a paucity of evidence in the study of historically remote stages (cf. article 14), whereas the amount and diversity of material increases as one approaches the present. In the case of English and several other major standardised languages, the past two decades have seen an explosion in the amount of corpora and digitised textual data available – to the point that it is sometimes difficult to identify the really valuable and useful data in a mass of available ones (see article 20). With regard to the quality of the data, the greatest advantage for the student of recent change is probably direct access to spoken data, which – after all – are the site of origin of almost all non-prestige innovations in language. For older stages of the language developments in the spoken language can only be reconstructed indirectly, from written evidence. Written data generally still dominate in corpora of present-day language, but some spoken data are usually available, though not necessarily in the quality that would be desirable for the fine-grained study of phonetic change (the quality of transcriptions or the lack of access to the original recordings being the main limiting factors).

If the greater amount and diversity of available data, in particular the opportunity to include spoken language, put the student of recent change at an advantage in comparison to those who focus on historically remote periods, there is one factor which makes such investigations problematical and their results provisional in ways which work on older stages of the language is not. Trying to identify and document ongoing change is always an attempt at “hitting a moving target” (Bauer 2002, 55): we cannot explain what is going on now in the light of the presumable end-point of a change, because (1) we cannot be sure about this end-point and (2) even if we have plausible assumptions, we have to keep in mind that an observed trend might be halted or reversed, with the result that a budding diachronic development may revert to become part of the always greater background “noise” of synchronic (regional, social or stylistic) variation.

Ultimately, differences in the amount and quality of the data may lead to slightly different methodological stances among students of remote and recent change in language. Corpus-linguistic purism, as reflected, for instance, in the careful construction of relatively small corpora, has generally been the dominant trend in the study of older stages in the history of languages. Consider, for example, the Helsinki Corpus, which offers a balanced and generically varied collection of Old, Middle and Early Modern English texts prepared to high philological standards but amounting to a total of c. 1,6 million words only (which is meritorious in its field but small by the standards of present-day synchronic corpora). To some extent, such philological purism may serve as a model also for work on more recent periods, but it definitely cannot be the last word in the study of all important ongoing processes of change. For many investigations, textual databases and digital archives will have to be consulted in addition to standard linguistic reference corpora, although such resources were not originally compiled for the purposes

of linguistic analysis and are therefore deficient in many ways if compared to true linguistic corpora. Even that messiest of all corpora, the World Wide Web, is a potentially useful resource for the study of current changes. Its many deficiencies are obvious, but sometimes these are not enough to offset the one powerful advantage which makes the Web indispensable as a source of data. It is an ever-accumulating self-updating monitor corpus documenting current usage with almost no time delay. (For a comprehensive survey of the state of the art in web-based corpus-work see Lüdeling/Evert/Baroni (2007) and article 18.)

As the present article will demonstrate, the corpus-based approach has a lot to offer for the study of recent and ongoing change. However, it must be admitted that it is not equally well suited to the investigation of all types of linguistic change. Given the present state of most corpora of spoken language, the potential for research into ongoing changes in pronunciation is still limited. Depending on the size and quality of corpora available for the language in question, it is, however, usually possible to study a wide range of changes in the grammar and the lexicon, either in “real-time” or “apparent-time” research designs (or a combination of the two). An additional advantage of the corpus-based approach is that it makes it possible to systematically explore the interrelations between the spread of structural innovations (which usually proceeds at differential rates in different text-types or genres) and changes of historically evolving traditions of speaking and writing (which of course ultimately reflect social and cultural changes in the community).

Owing to the author’s particular expertise, exemplification in this article will be provided from the recent history of the English language, and from English language corpora. (For some illustrative examples of the use of corpora in the study of recent changes in languages other than English see Belica 1996 or the “Digitale[s] Wörterbuch der deutschen Sprache des 20. Jahrhunderts [DWDS]” website (<http://www.dwds.de>) for German, Asmussen 2006 for Danish, Falk 2004 for Spanish, or Blanche-Benveniste 2001 for French.) The methodological assumptions and research procedures, however, are not language-specific, and the limiting factor in their application to other languages is the availability of corpora and other digital resources. For an overview of available resources for English and other languages, see article 20.

2. “Real-time” and “apparent-time” approaches in the study of recent and ongoing change

2.1. Introduction: The pitfalls of anecdotal observation

Claims about changes in Old, Middle and Early Modern English are usually backed up by systematic empirical research. In view of this, it is unfortunate to see that the main source of “evidence” in the literature on recent and ongoing change is very often anecdotal and impressionistic observation. This inevitably results in a distorted picture of linguistic developments. The spread of salient new uses is exaggerated, while the less salient persistence of older forms is not noted or, worse still, a diachronic trend is read into a situation which merely shows variable or fluctuating usage. This will be illustrated in the present article with a well-known case of variable prepositional usage in contempo-

rary English: the use of *from*, *to* and *than* after the adjective *different*. Common assumptions voiced in reference works and the linguistic literature (cf. below) will be compared with the results of systematic corpus-based inquiry (cf. 2.2.1. below).

From almost the beginning of the twentieth century, there has been a tradition of comment which regards *from* as correct and *to* and *than* as problematical innovations, with the added complication that *to* is dominantly British and *than* dominantly US usage. In his *American Language*, H. L. Mencken quotes a letter to the editor of the *New York Herald* written in September 1922:

“Within a few years the abominable phrase *different than* has spread through the country like a pestilence. In my own Indiana, where the wells of English undefiled are jealously guarded, the infection has awakened general alarm.”

(Meredith Nicholson, quoted in Mencken 1963, 570)

By the end of the 20th century, if we are to trust the comments in the literature, the traditional *different from* is under threat in all parts of the English-speaking world. According to Trudgill/Hannah's widely used standard reference work on varieties of English, *different than* is now the normal form in American English: “The comparative adjective *different* is usually followed by *from* (and sometimes *to*) in EngEng, while in USEng it is more usually followed by *than*” (2002, 74). Jenkins goes further, claiming that *different from* has disappeared from American English altogether: “The comparative adjective ‘*different*’ is followed by ‘*than*’ in USEng and by ‘*from*’ (or more recently, ‘*to*’) in EngEng” (2003, 75). As will be seen (cf. 2.2.1.), these claims have a very tenuous basis in linguistic fact when tested against corpora, which provides proof for the necessity of corpus-based work also on the recent history of the language.

2.2. Documenting recent and ongoing change systematically in corpora

2.2.1. Documentation in real time

As sociolinguistics has taught us, language change can be studied in “real time,” by comparing the state of the language at (at least) two different points in time, or in “apparent time,” by extrapolating diachronic developments from synchronic variation. Other things being equal, the real-time approach would seem to be preferable as the more direct one, which is why it will be treated first.

The ideal type of a real-time study is a sociolinguistic community survey repeated after an appropriate interval. Writing on phonetic change, Labov suggests that confirmation of a suspected linguistic change in real time is obtained “if it is demonstrated in the near future that the trend detected has moved further in the same direction. ‘Recent past’ and ‘near future’ must mean a span of time large enough to allow for significant changes but small enough to rule out the possibility of reversals and retrograde movements: we might say from a minimum of a half generation to a maximum of two” (1981, 177). For lexical change, the minimum span of observation may be shorter, whereas for grammatical change it will almost certainly be longer. Not surprisingly, real-time research is not very popular in sociolinguistics because of the massive logistical and organisational difficulties involved in staging a repeat of a community survey. In a recent

survey, William Labov (1994, 85–98) lists only four projects which qualify for the status of a genuine follow-up. Two of them are concerned with English-speaking communities, namely a repeat of Labov's own 1966 New York City department store survey (by Fowler, apparently unpublished, but see Labov 1994, 86–94 for a detailed summary of the results) and Trudgill (1988), which follows up his own (1974) study of language use in Norwich. The other two real-time studies reported on by Labov concern a French dialect of Switzerland documented at two successive points in time during the first third of the twentieth century (see Labov 1994, 85–86, for a summary of the results) and Spanish in Panama City in 1969 and 1983 (see Labov 1994, 94–97, for a summary of results of work by Henrietta Cedergren). Since Labov surveyed the field, there has been one more follow-up study, re-visiting Martha's Vineyard, the site of one of Labov's pioneering studies of change in progress (Josey 2004).

“Real-time” research on corpora would appear to be easier by comparison. The corpus-linguistic equivalent of the real-time sociolinguistic survey is the construction of matching corpora representing the state of “the language” or some specified variety at different times. The use of parallel or matching corpora has a long and distinguished tradition going back to the pre-computational era in English historical linguistics. It was the method used by all those traditional philologists who based their observations on analyses of successive translations of the Bible into English – as Otto Jespersen did, for example, when he illustrated the increase in the frequency of progressive forms since the Middle English period in this way (see Jespersen 1909–1949, IV: 177). What the advent of digital language processing has brought about is thus not so much an invention of an entirely new method as a widening of the scope of an existing one. It is now no longer just a small number of sacred or otherwise privileged texts which are translated at successive points of time or concordanced, but an increasingly broad range of registers and styles. In addition, the computationally assisted retrieval of forms from digitised corpora makes it possible to access the data faster and, in many cases, to tackle problems which great philologists such as Visser and Jespersen shied away from as being too time-consuming. Writing about the variation of gerundial and infinitival complements after the verb *begin* a little more than thirty years ago, Visser deplored the following apparent dilemma:

Today *begin* + form in -ing is used with striking frequency alongside of *begin* + infinitive. Which of the two alternatives predominates cannot be ascertained because of the lack of statistical data. (Visser 1970–1978, III: 1888)

Today, at least for a language such as English, with its rich panchronic corpus-linguistic working environment, it is easy to fill in this gap in our language-historical knowledge. The extent to which the gerund has gained ground as a complement of the verb *begin* in the course of the past century has been documented (Mair 2002, 115–121).

In the field of English historical linguistics, work on recent changes was inspired by previous corpus-based diachronic research on older stages of the language and work on synchronic corpora. The early phase of computer-assisted diachronic research on the history of English was inaugurated by the publication of the Helsinki Corpus (Old, Middle and Early Modern English). Coverage was extended to the present in the subsequent ARCHER (= “A Corpus of Historical English Registers”) project, which documents British and American English, sampled according to genre, at 50-year intervals from c. 1650 to the present. The chief limitation of ARCHER for the study of very

recent and ongoing changes is the small size of its 20th-century components (< 0,5 million words). In a research project conducted by the present author, the Brown and LOB corpora, one-million word corpora which document 15 different genres of written texts in American and British English in 1961, were complemented with matching databases representing the state of the two varieties in 1992 and 1991 respectively. These corpora are generally known under the abbreviations “Frown” (for “Freiburg update of the Brown corpus”) and “FLOB” (“Freiburg update of the LOB corpus”). A visual representation of the relations among the four corpora is provided in Figure 52.1:

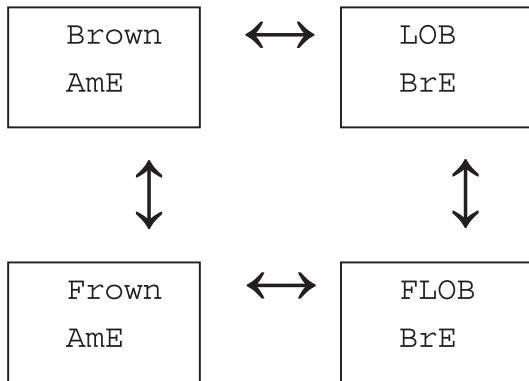


Fig. 52.1: Four matching one-million-word corpora of written English

The arrows show that this quartet of corpora makes it possible not only to study developments in each of the two varieties in real time, but also to investigate the question of how these short-term diachronic developments are related to synchronic (i. e. regional or stylistic) variation at any one time. The corpus-based real-time approach has recently been extended to cover spoken English by researchers based at the Survey of English Usage (University College London), who have created a diachronic corpus of spoken British English by matching fitting samples of the London-Lund-Corpus (1959–1988) and the British component of the International Corpus of English (1990–1993). For information on the “parsed and searchable diachronic corpus of present-day spoken English” (DCPSE), see <<http://www.ucl.ac.uk/english-usage/diachronic.index.htm>> and articles 20 and 30.

For a simple illustration of how the four corpora can be used, consider Table 52.1, with figures for the frequency of the preposition *on* and its archaic variant *upon*:

Tab. 52.1: Proportion of *on/upon* in four corpora*

	1961	1991/92
British English (LOB/FLOB)	6913/407	7123/243
American English (Brown/Frown)	6719/493	6900/196

* Significances: LOB: FLOB p < 0.001, Brown: Frown p < 0.001; LOB: Brown p < 0.01, FLOB: Frown p > 0.05

The major significant development is a parallel decline in the frequency of *upon* in both varieties. Whereas in LOB c. 5.6 per cent of the combined total of occurrences was realized “conservatively”, as *upon*, this share sank to c. 3.3 per cent in FLOB. The corresponding figures for Brown and Frown (c. 6.8 and c. 2.8 per cent respectively) further show that the decline has proceeded somewhat more rapidly in American English (from a frequency slightly higher than the British in 1961 to one lower in 1992). Regional contrasts, by comparison, were less significant statistically (and probably not salient psychologically) in 1961, and have weakened further since. In terms of choice between the variants, *upon* had a share of around 7 per cent of all relevant forms in Brown, which decreased to less than three per cent in Frown. The short-term late-20th-century trend reflected in the four corpora smoothly continues a long-term development, which can be documented on the basis of data from the OED quotation base. To illustrate usage c. 1700, c. 1800, and c. 1900, the frequencies of *on* and *upon* were obtained from three different ten-year selections of the quotation base, as shown in Table 52.2:

Tab. 52.2: Proportion of *on/upon* in three samples from the OED quotation base

OED 1696–1705	1824/1000
OED 1796–1805	2525/744
OED 1896–1905	5999/902

There is a linear increase in the proportion of *upon*, as one goes back in time – from c. 13 per cent in 1900, to c. 22 per cent in 1800, and c. 28 per cent in 1700. All in all, the figures presented in Tables 52.1 and 52.2 show that *upon* has been obsolescent for some time. The results of this small-scale experiment also support a methodological point made above – namely that in the study of recent and ongoing change it is useful to combine traditional corpora (the Brown-quartet in the present case) with other digitised language databases (in this case the quotation base of the OED, a digitised dictionary). Note that the results tabulated so far have been mainly quantitative. Qualitative and/or explanatory follow-up studies, though beyond the scope of the present survey, would form an essential second stage in the work (see Rohdenburg 2002 for some promising directions to explore in accounting for the changing distribution of *upon* and *on*).

For some non-trivial exemplification of the usefulness of the corpus-based approach, let us now turn to the case of *different*, for which, as was pointed out in 2.1., impressionistic observations suggest rapid diachronic change and considerable regional divergence between varieties of English. Table 52.3 presents the figures from the four one-million-word 20th-century reference corpora mentioned above, which suggest that the impressionistic analysis needs to be qualified:

Tab. 52.3: Prepositions following *different* in four corpora (cf. Mair 2006, 26)

	Brown (US 1961)	LOB (Britain 1961)	Frown (US 1992)	FLOB (Britain 1991)
<i>different from</i>	29	20	32	39
<i>different to</i>	–	1	–	3
<i>different than</i>	1	1	1	–

Note that only instances were counted in which the adjective *different* occurred next to the preposition. This led to some (insignificant) under-collection, because forms such as “a different theory from/to/than my own”, with the adjective separated from the preposition, are possible (for counts of LOB and Brown which aimed for complete coverage but do not produce entirely compatible results, apparently due to slight differences in the handling of borderline cases, see Hundt 1998a, 106 and Kennedy 1998, 195).

One thing is obvious. Contrary to claims, American and British English are rather similar in their use of prepositions after *different*, and little seems to be happening diachronically. Impressionistic observation is supported to the extent that *to* is restricted to British English. A closer look at the three instances of *different than* reveals the influence of syntactic environment on the choice of the preposition. Only the Frown example has *than* interchangeable with the other two (“the second Olympics is different than Seoul” – A 19 01); the LOB case has *than* in front of another preposition (“different than in the first part” – D 04 88), while in Brown it introduces a clause (“no evidence that anything was different than it had been” – L 10 670). As regards long-term developments, the stable and overwhelming dominance of *different from* can easily be established from an analysis of the OED quotation base. A search for occurrences of *different* and *from/to/than* in direct contact yielded 650 cases of *different from*, 40 of *different to*, and a mere 9 of *different than*. The massive diachronic changes noted in impressionistic observation thus seem to have been mainly in the eye of the beholder.

This general impression, that *different from* dominates the alternatives, is confirmed by an analysis of data from the World Wide Web, whose results deserve discussion as they might help allay widely held reservations about the usefulness of this very messy mega-corpus. Returns from regionally stratified searches of the Web are surprisingly robust and show that the picture suggested by the four corpora analysed above – dominance of *different from* in all varieties of English, and a minor contrast in the preference for *than* and *to* in American- and British-influenced varieties respectively – is correct. As much more material is analysed, we obtain solid regional trends of preference for *than* and *to* in American- and British-derived varieties respectively.

As can be seen, it is difficult to identify a suitable domain for American English on the Web. The “.us” domain is not much used. The “.edu” domain, which contains enough material, is used mainly by US institutions of higher learning but includes some others. The “.gov” domain (US government) is biased for text type and register. It thus goes without saying that in interpreting these figures, allowances have to be made, not only with regard to the targeting of varieties of English through the Google advanced-search facility but also for overcollection, especially common for instances of *different to*. It seems, however, that all the many potential sources of error seem to cancel each other out, for we get the American profile known from the smaller corpora in the American Web domains, and the expected “North American” profile in the Canadian material (“.ca”). Other interesting findings contained in Table 52.4 are that – at least on the present criterion – Irish English (“.ie”) patterns with British English, as do all the Southern hemisphere ex-colonial Englishes (cf. “.au”, “.nz” and “.za” for Australia, New Zealand and South Africa).

The only type of material which lends some support to the impressionistic observations reported in 2.1. above is spoken corpora. The direct conversations from the British component of the International Corpus of English (ICE) (c. 185,000 words) show an even spread of *different from* (4 instances) and *different to* (5), with *than* being absent.

Tab. 52.4: Prepositions following *different* in regionally stratified Web material*

Domain	Region	<i>from</i>	%	<i>than</i>	%	<i>to</i>	%
.us	USA	194.000	68 %	85.200	30 %	6.060	2 %
		1.450.000	79 %	343.000	19 %	33.100	2 %
		787.000	83 %	152.000	16 %	6.050	1 %
		11.000	76 %	3.180	23 %	235	1 %
		2.442.000	79 %	612.000	20 %	45.445	1 %
.ca	Canada	253.000	76 %	68.700	21 %	11.200	3 %
.uk	United Kingdom	469.000	71 %	33.000	5 %	157.000	24 %
.au	Australia	171.000	60 %	14.800	5 %	98.800	35 %
.nz	New Zealand	45.400	67 %	4.290	6 %	17.700	26 %
.za	South Africa	28.700	66 %	2.910	7 %	11.600	27 %
.ie	India	25.600	65 %	2.330	6 %	11.600	29 %
Total		8.160.000	71 %	2.500.000	22 %	825.000	7 %

* Google, 30 May 2004

In the much larger “spoken-demographic” texts of the British National Corpus (4 million plus words), there are 21 instances of *different from*, 46 of *different to*, and 4 of *different than* (of which 3 show *than* being used to introduce clauses). The Santa Barbara Corpus of spoken American English, which contains unplanned dialogue, is too small to yield results (1 instance of *different from*), while the Corpus of Spoken Professional American English, which mainly documents press briefings and faculty meetings, has 91 instances of *different from* slightly outnumbering the 82 cases of *different than* (and none of *different to*). The Longman Corpus of Spoken American English (c. 5 million words) has 97 instances of *from*, 64 of *than*, and 6 of *to*.

To sum up, a corpus-based study of prepositional usage with *different* has led to important corrections of assumptions about ongoing change based on impressionistic observation. Observation unaided by corpora has tended to under-rate the dominance and persistence of *different from*, especially in writing. Conversely, the spread of the two regionally salient informal alternatives, *different than* for American English and *different to* for British English, has been over-rated, vastly so in writing, and to some extent also in informal speech. Again, as in the case of *upon*, the dominantly quantitative data can provide the basis for various types of qualitative and explanatory follow-up studies (cf., e.g., Mair 2006, 18–21, 25–28, or Rohdenburg 2002).

2.2.2. Documentation in apparent time

Whenever a real-time approach is not possible, either because a sociolinguistic survey cannot be repeated or a matching corpus is not available or cannot be compiled, extrapolation of diachronic trends in “apparent time” is the method of choice in the study of change in progress. The basis of this method is the fact that most linguistic changes start with younger speakers, lower-class speakers, and in spoken and informal language and

then spread into formal and written registers and educated middle-class usage. If in a synchronic sample a form is favoured by the younger informants or more frequent in speech or informal writing, it is plausible to regard it as an innovation in the early stages of its spread through the community.

The method is convenient, but certainly not without its pitfalls. It assumes that older speakers add little to their grammar and phonology after adolescence. This is a plausible working hypothesis (underpinning, for example, much formalist and generativist work on diachronic change – cf., e.g., Lightfoot 1999), which, however, has never been proved conclusively. For the study of lexical change, the method is unsuitable from the start, as speakers modify their vocabularies throughout most of their lives and neologisms can be coined by young and old members of the community. Two additional problems need to be solved. One is to find a way to identify “prestige” innovations, which do not follow the usual trend of spreading from “below” but diffuse from educated into general use, from formal into informal language, from writing into speech and – presumably – from old to young. This is not very difficult in most cases as such usages are in the category of what Labov (1972, 140f.) calls “linguistic markers” which the community is aware of and uses consciously (unlike the “linguistic indicators”, which tend to operate below the level of conscious awareness). A much trickier problem is the phenomenon of age-grading, which usually manifests itself in a temporary adolescent affinity to nonstandard or innovative usages which disappears in an individual’s later life. This means that in such cases generations of teenage linguistic rebellion will not lead to a lasting change in community norms.

Apparent-time analyses of ongoing change are usually very easy to undertake on the basis of corpora. The only condition is that the corpus texts have been produced at roughly the same time and that the corpus contains texts from more than one speaker, text-type or genre. Apparent-time studies in a wider sense can also be based on other digitised textual sources, for example electronic newspaper archives from the same year but different regions, or even on regionally or stylistically stratified selections of web texts. As for the interpretation of the results, all the general cautions on the use of the method apply.

A good illustration of the potential and limitations of corpus-based apparent-time analyses of ongoing change is provided by Rayson/Leech/Hodges (1997), who analysed lexical frequencies in the spoken-demographic component of the British National Corpus, more than four million words of transcribed spontaneous speech produced by a sociologically representative sample of British speakers. Table 52.5 shows the ten most typical words in the speech of the over-35s and under-35s, and middle-class and working-class speakers, with typicality being defined not in terms of absolute frequency but as statistical over-representation in the sample in question.

In the speech of the under-35s, four out of ten words are obvious and trivial cases of age-grading: *mum*, *mummy*, *dad*, and *daddy*. The words are not nonstandard but they are clearly of the type which children and adolescents living with their families have special occasion to use. The over-representation of two common swear words, *fucking* and *shit*, is partly due to age grading, as well. It certainly does not show “new” words spreading in the community. If anything, it might be interpreted not as a sign of lexical change but of a change in sociolinguistic norms of propriety governing what is acceptable speech. The presence in the list of *like* and *goes* might be due to their uses as discourse particle and speech-reporting verb respectively (“he’s like fifteen years old like”

Tab. 52.5: Lexical items most characteristic of four groups of speakers in a corpus of spoken British English (adapted from: Rayson/Leech/Hodges 1997)

10 words most characteristic of	over-35s	under-35s	middle-class speakers	working-class speakers
1	Yes	mum	yes	he
2	Well	fucking	really	says
3	Mm	my	okay	said
4	Er	mummy	are	fucking
5	They	like	actually	ain't
6	Said	na*	just	yeah
7	Says	goes	good	its
8	Were	shit	you	them
9	The	dad	erm	aye
10	Of	daddy	right	she

* As in *gonna* or *wanna*, which for the purposes of the CLAWS tagger, are counted as two words

or “and then she goes: no way”). However, to verify this assumption on the basis of BNC data would involve an extremely time-consuming qualitative analysis of tens of thousands of attestations. The most typical features in the language of the older speakers likewise point to phenomena which are unlikely to be part of diachronic change. The over-representation of *the* and *of* suggests the presence of more complex noun-phrases; discourse markers such as *well*, *mm* and the hesitation phenomenon *er* indicate a different conversational atmosphere. Even the question of whether the over-representation of *yes* should be interpreted as a sign of the obsolescence of this form and its impending replacement by *yeah* cannot be answered straightforwardly. So what we are left with as a plausible reflection of genuine lexico-grammatical change is the morpheme **na*, the second element in the contracted forms *gonna* and *wanna*. The BNC codes speakers for six age groups, and the frequency of *wanna* (measured in occurrences per million) gives a near perfect gradient in apparent time, cf. Table 52.6.

Tab. 52.6: Frequency of *wanna* in the BNC per age group (x/1,000,000 words)

Age	wanna	want to
0–14	1178	1210
15–24	700	605
25–34	496	669
35–44	368	780
45–59	330	684
60+	159	679

These figures provide clear evidence for the spread of *wanna* in contemporary British English. The apparent-time distribution is perfect for *wanna*, while some questions remain for the uncontracted form *want to*, which shows no sign of decreasing in frequency even among those groups which contract freely. What becomes apparent is most likely a frequency profile associated with ongoing grammaticalisation: the frequency of *want* as a whole is increasing, the contracted form *wanna* is becoming proportionately more common among younger speakers, and the surprisingly high frequency of the uncontracted form among the youngest group in this case might be due to age-grading (the very young speakers being unable or unwilling to resort to usual politeness strategies in formulating their desires).

In theory, apparent-time interpretations could also be based on lexical differences based on speakers' social class (the two right-hand columns in Table 52.5). However, results tend to be even less conclusive than in the analysis directly based on speaker-age. Middle-class speech seems to be characterised by discourse markers such as *actually*, *okay* or *right*, none of which is old-fashioned or obsolescent. In the spoken BNC, for example, the use of *actually* peaks at a frequency of 1,309 instances per million words in the 15–24 age group and hovers inconclusively between the values of 538 and 838 in the five others (0–14, 25–34, 35–44, 45–59, 60+). The working-class sample suggests that *yeah* might be spreading at the expense of *yes*, which – as will be remembered – has its strongest base among the over-35s, but more research is needed for a conclusive answer.

In sum, the findings from Rayson/Leech/Hodges 1997 and follow-on studies of *wanna* undertaken here show that it is possible to carry out apparent-time analyses of ongoing change on the basis of corpora which, as does the BNC, cover a variety of spoken and written genres. However, as the greater part of age-based and social variation observed reflects stable stylistic and social contrasts or age-grading, “apparent time” is a risky method to use on its own. Its main use is to serve as a complement to corpus-based real-time studies, and to provide clarification in those cases in which there are strong independent grounds to assume that a change is under way.

3. Some representative corpus-based work on recent change in English grammar and vocabulary

How to define “recent” change in language is open to some debate. In the methodological considerations presented above a fairly generous definition was adopted which occasionally allowed us to have a look back even into the 18th century. In the present section, which intends to offer an orientational survey of relevant corpus-based research, the term will be defined more narrowly, as relating to changes which started or gained significant momentum in the course of the 20th century.

A good starting point is Bauer (1994), a monograph which deals with a large number of suspected ongoing phonetic, lexical and grammatical changes in contemporary English, and usually provides some corpus documentation, usually based on small and occasionally on self-compiled databases. Bauer's survey should be complemented by volume IV of the *Cambridge History of the English Language* (Romaine, ed. 1998), which covers the period from 1776 to 1997. The chapters on changes in the lexicon (Algeo 1998) and

syntax (Denison 1998), while not dominantly corpus-linguistic, contain some documentation from corpora and related digital databases, such as the electronic version of the OED.

A large number of specialised studies have been carried out on the basis of the Brown quartet of corpora described in 2.2.1. by members of two research teams based at the Universities of Lancaster (UK) and Freiburg (Germany). Given the size of these corpora, the focus has been on mid- to high-frequency grammatical rather than lexical phenomena. Leech (2003) and Smith (2003b) present findings on an apparent decrease in the frequency of many modal verbs which – to the extent that it is not accounted for by a corresponding rise in the frequency of semi-modals – points to changing stylistic and textual norms in contemporary English writing. Smith (2002) and (2003a) explore 20th-century developments in the frequency and function of the English progressive. Mair (2002) surveys 20th-century changes in the use of three variable non-finite complement structures, namely *prevent* + NP + V-ing (as against *prevent* + NP + *from* + V-ing), *to*-vs. bare infinitive with *help*, and gerund or infinitive with *begin* and *start*. Hundt (1998b) reports on the spread of the mandative subjunctive into written British English as documented in the four corpora, while Hundt/Mair (1999) investigate the genre-specific differential speed of many linguistic changes in a comparison of the journalistic and academic writing samples contained in corpora. Leech (2004) and Mair/Leech (2006) present summaries of the main results obtained in various analyses of the four corpora so far. In addition to enumerating individual lexico-grammatical changes, both papers attempt integrated descriptions of grammatical, stylistic and social changes – for example by showing that several apparently disconnected structural changes observed in the material can be traced back to a common underlying socio-stylistic cause, namely the “colloquialisation” (a term originally introduced in Mair 1998) of the norms of written usage in the course of the second half of the 20th century. Mair et al. (2002) reports on the early results of work on the comparison of the tagged corpora. Among the more interesting results presented in this paper is the finding that, notwithstanding the recent trend towards the informal and colloquial in written English, information density as reflected in various types of nominal compounds has increased rather than decreased in the period of observation. A provisional summary of the group’s work on the Brown family of corpora is provided by Leech/Smith/Hundt/Mair 2008. Notable further work on recent grammatical change in English has been produced by Rohdenburg and his group, who have tended to use different, or additional, databases and to work against a larger diachronic horizon extending from the Early Modern English period to the present. Thus, Rohdenburg (2003) has used OED and BNC data to trace developments in the use of interrogative clause linkers, whereas Schlueter (2005) has explored the phonology/syntax interface to study recent developments in the field of participial adjectives (*drunken* vs. *drunk*, *lighted* vs. *lit*, etc.).

As for the documentation of ongoing lexical change in English, most dictionary publishers now routinely use large corpora in order to detect neologisms and to arrive at informed decisions on what items to include in their regular updates. The work of the compilers of the OED Online presents an exemplary benchmark in this field (<http://dictionary.oed.com>). Copious quarterly updates present a combination of routine revisions of the entries for selected spans of the alphabet and “out-of-sequence” entries, which are in the main comprised of recent neologisms whose inclusion the editors have felt to be particularly urgent. In linguistic morphology, corpora are currently much used

in the assessment of the productivity of word-formation patterns, both established and innovative ones (cf. Baayen/Renouf 1996 or Fischer 1998 for two exemplary studies using electronic newspaper text, or – on German – Lemnitzer/Ule's ‘Wortwarte’ at <<http://www.sfs.uni-tuebingen.de/~lothar/nw/index.html>>).

4. Conclusion and outlook: A plea for methodological pluralism

In his widely used introduction to the field, Chambers has argued that the advances made by sociolinguists in the study of change in progress constitute “the most striking accomplishment of contemporary linguistics” (Chambers 1995, 147). Put in such exclusive terms, this statement may be controversial, but even those who rank other breakthroughs in linguistic research as equally or more striking will usually recognise the sociolinguistic achievement in this field. What the present contribution has shown is that sociolinguistic approaches to ongoing change can usefully be complemented by corpus-linguistic methods. Corpora and related digital-text resources provide important data for the study of recent and ongoing change in language, and corpus-linguistic methods of investigation are one of several essential ingredients in a much needed interdisciplinary approach to this complex subject.

The most immediate use of corpora is as a corrective to impressionistic and anecdotal observations on suspected changes. As has been shown, such observations tend to emphasise what is (or appears to the observer to be) new, unusual or bizarre usage, and to misjudge the speed of developments by compressing into a single lifetime what takes hundreds of years to unfold in reality. Also, impressionistic observation is usually fixated on the mere tip of the iceberg of ongoing change, namely those phenomena which for some reason or other have aroused the concern of prescriptivists or the linguistically aware public. Corpora help to balance such views by emphasizing continuity in linguistic development. They show that changes outside the lexicon are usually much slower than suspected and that they are embedded in synchronic regional and stylistic variation. Most of all, however, they allow us to focus on the groundswell of linguistic change, i. e. those developments in the core grammar which proceed below the threshold of conscious awareness and hence usually escape anecdotal observers.

Sociolinguists have been the pioneers in the study of ongoing change, and corpuslinguists must learn from them. However, they can also make an independent contribution because the strengths and weaknesses of the two approaches complement each other. Sociolinguistic work has focused on phonetic changes, especially in nonstandard varieties, whereas corpuslinguists are best equipped to deal with lexical and grammatical change in the written/standard language.

The ordinary use of corpus data is to correct or refine current assumptions on ongoing change. In some cases, however, corpus analysis will turn out to be a genuine discovery procedure, making it possible – through a systematic comparison of frequencies in matching corpora – to identify changes which have gone unnoticed – either because they have proceeded below the threshold of speakers' conscious awareness and/or have escaped prescriptive censure. Corpus-based empiricism, however, will lead to nothing more than the accumulation of under-analysed and frequently pointless statistics unless the interpretation of the results is carried out in an appropriate theoretical framework.

This is well illustrated by an example presented above. For the proper understanding of the spread of *wanna* in contemporary British English it is not enough to record increases of frequency in corpora. This is a necessary first step, but the full interpretation requires a theory of grammaticalisation to handle the formal and structural changes involved, and a proper sociolinguistic model to assess the extent to which the spread of the innovation is speeded up or slowed down by prestige and stigma. In turn, it should be obvious that the further refinement of existing theoretical models of language change stands to benefit enormously from the results of such theoretically aware and empirically grounded corpus-based diachronic studies – regardless of whether they focus on remote or recent change.

The interdisciplinary and open spirit in which the study of ongoing change should be approached is well put by Rickford et al., who in a study on one ongoing syntactic change in contemporary American English recommend “exploration on the boundaries of sociolinguistic variation, corpus linguistics, historical linguistics, and syntax” (Rickford et al. 1995, 128) as the appropriate method in getting a grip on change in progress.

5. Acknowledgement

The author would like to thank Britta Mondorf for valuable comments on a previous version of the present contribution.

6. Literature

- Algeo, John (1998), Vocabulary. In: Romaine, Suzanne (ed.), *The Cambridge History of the English Language*, Vol. IV: 1776–1997. Cambridge: Cambridge University Press, 57–91.
- Asmussen, Jørg (2006), Towards a Methodology for Corpus-based Studies of Linguistic Change: Contrastive Observations and their Possible Diachronic Interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish. In: Wilson, Andrew/Archer, Dawn/Rayson, Paul (eds.), *Corpus Linguistics around the World*. Amsterdam: Rodopi, 33–48.
- Baayen, R. Harald/Renouf, Antoinette (1996), Chronicling *The Times*: Productive Lexical Innovations in an English Newspaper. In: *Language* 72, 69–96.
- Bauer, Laurie (1994), *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Longman.
- Bauer, Laurie (2002), Hitting a Moving Target. In: *English Today: The International Review of the English Language* 18, 55–59.
- Belica, Cyril (1996), The Analysis of Temporal Changes in Corpora. In: *International Journal of Corpus Linguistics* 1, 61–73.
- Blanche-Benveniste, Claire (2001), Le français au XXIe siècle: Quelques observations sur le grammairien. In: *Le français moderne* 58, 3–15.
- Chambers, Jack K. (1995), *Sociolinguistic Theory*. Oxford: Blackwell.
- Denison, David (1998), Syntax. In: Romaine, Suzanne (ed.), *The Cambridge History of the English Language*, Vol. IV: 1776–1997. Cambridge: Cambridge University Press, 92–329.
- Falk, Johan (2004), ‘La felicidad se trivializa’: Estudio sobre el uso de *estar feliz* basado en las bases de datos CREA y CORDE. In: *Studier i Modern Språkvetenskap* 13, 9–10, 49–71.
- Fischer, Roswitha (1998), *Lexical Change in Present-day English: A Corpus-based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Narr.

- Hundt, Marianne (1998a), *New Zealand English Grammar – Fact or Fiction? A Corpus-based Study in Morphosyntactic Variation*. Amsterdam: Benjamins.
- Hundt, Marianne (1998b), *It is important that this study (should) be based on the analysis of parallel corpora: On the Use of the Mandative Subjunctive in Four Major Varieties of English*. In: Lindquist, Hans/Klintborg, Staffan/Levin, Magnus/Estling, Maria (eds.), *The Major Varieties of English: Papers from MAVEN 97, Växjö 20–22 November 1997*. Växjö: Växjö University Press, 159–173.
- Hundt, Marianne/Mair, Christian (1999), ‘Agile’ and ‘Uptight’ Genres: The Corpus-based Approach to Language-change in Progress. In: *International Journal of Corpus Linguistics* 4, 221–242.
- Jenkins, Jennifer (2003), *World Englishes: A Resource Book for Students*. London: Routledge.
- Jespersen, Otto (1909–1949), *A Modern English Grammar on Historical Principles*. Copenhagen: Munksgaard.
- Josey, Meredith Pugh (2004), A Sociolinguistic Study of Variation and Change on the Island of Martha’s Vineyard. PhD dissertation, New York University.
- Kennedy, Graeme (1998), *An Introduction to Corpus Linguistics*. London: Longman.
- Labov, William (1966), *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.
- Labov, William (1972), *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.
- Labov, William (1981), What can be Learned about Change in Progress from Synchronic Description? In: Sankoff, David/Cedergren, Henrietta (eds.), *Variation Omnibus*. Edmonton, Alberta: Linguistic Research, 177–201.
- Labov, William (1994), *Principles of Linguistic Change*. Vol. I: *Internal Factors*. Oxford: Blackwell.
- Leech, Geoffrey (2003), Modality on the Move: The English Modal Auxiliaries 1961–1992. In: Facchinetto, Roberta/Krug, Manfred/Palmer, Frank (eds.), *Modality in Contemporary English*. Berlin/New York: Mouton de Gruyter, 223–240.
- Leech, Geoffrey (2004), Recent Grammatical Change in English: Data, Description, Theory. In: Altenberg, Bengt/Aijmer, Karin (eds.), *Advances in Corpus Linguistics: Proceedings of the 23rd ICAME Conference, Gothenburg, 2002*. Amsterdam: Rodopi, 61–81.
- Leech, Geoffrey/Smith, Nicholas/Hundt, Marianne/Mair, Christian (2008), *Changes in Contemporary English: a Corpus-Based Study*. Cambridge: Cambridge University Press.
- Lightfoot, David (1999), *The Development of Language: Acquisition, Change, and Evolution*. Malden, MA: Blackwell.
- Lüdeling, Anke/Evert, Stefan/Baroni, Marco (2007), Using Web Data for Linguistic Purposes. In: Hundt, Marianne/Nesselhauf, Nadja/Biewer, Carolin (eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 7–24.
- Mair, Christian (1998), Corpora and the Study of the Major Varieties of English: Issues and Results. In: Lindquist, Hans/Klintborg, Staffan/Levin, Magnus/Estling, Maria (eds.), *The Major Varieties of English: Papers from MAVEN 97, Växjö 20–22 November 1997*. Växjö: Växjö University, 139–157.
- Mair, Christian (2002), – Three Changing Patterns of Verb Complementation in Late Modern English: A Real-time Study Based on Matching Text Corpora. In: *English Language and Linguistics* 6, 105–131.
- Mair, Christian (2006), *Twentieth-century English: History, Variation, Standardization*. Cambridge: Cambridge University Press.
- Mair, Christian/Hundt, Marianne/Smith, Nicholas/Leech, Geoffrey (2002), Short-term Diachronic Shifts in Part-of-Speech Frequencies: A Comparison of the Tagged LOB and F-LOB Corpora. In: *International Journal of Corpus Linguistics* 7, 245–264.
- Mair, Christian/Leech, Geoffrey (2006), Current Changes in English Syntax. In: Aarts, Bas/McMahon, April (eds.), *The Handbook of English Linguistics*. Oxford: Blackwell, 318–342.
- Mencken, Henry Louis (1963), *The American Language*. 4th ed. Revised by Raven I. McDavid. New York: Knopf.

- Rayson, Paul/Leech, Geoffrey/Hodges, Mary (1997), Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: *International Journal of Corpus Linguistics* 2(1), 133–152.
- Rickford, John R./Mendoza-Denton, Norma/Wasow, Thomas A./Espinoza, Juli (1995), Syntactic Variation and Change in Progress: Loss of the Verbal Coda in Topic Restricting *as far as* Constructions. In: *Language* 71, 102–31.
- Rohdenburg, Günter (2002), Processing Complexity and the Variable Use of Prepositions in English. In: Cuyckens, Hubert/Radden, Günter (eds.), *Perspectives on Prepositions*. Tübingen: Niemeyer, 79–100.
- Rohdenburg, Günter (2003), Cognitive Complexity and *Horror Aequi* as Factors Determining the Use of Interrogative Clause Linkers in English. In: Rohdenburg, Günter/Mondorf, Britta (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 205–249.
- Romaine, Suzanne (ed.) (1998), *The Cambridge History of the English Language*, Vol. IV: 1776–1997. Cambridge: Cambridge University Press.
- Schlüter, Julia (2005), *Rhythmic Grammar: The Influence of Rhythm on Grammatical Variation and Change in English*. Berlin/New York: Mouton de Gruyter.
- Smith, Nicholas (2002), Ever Moving on? The Progressive in Recent British English. In: Peters, Pam/Collins, Peter/Smith, Adam (eds.), *New Frontiers of Corpus Research: Papers from the Twenty-First International Conference on English Language Research on Computerized Corpora, Sydney 2000*. Amsterdam: Rodopi, 317–330.
- Smith, Nicholas (2003a), A Quirky Progressive? A Corpus-based Exploration of the *will + be + -ing* Construction in Recent and Present-day British English. In: Archer, Dawn/Rayson, Paul/Wilson, Andrew/McEnery, Tony (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*. (Technical Papers 16.) Lancaster University: UCREL, 714–723.
- Smith, Nicholas (2003b), Changes in the Modals and Semi-modals of Strong Obligation and Epistemic Necessity in Recent British English. In: Facchinetto, Roberta/Krug, Manfred/Palmer, Frank (eds.), *Modality in Contemporary English*. Berlin/New York: Mouton de Gruyter, 241–256.
- Trudgill, Peter (1974), *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trudgill, Peter (1988), Norwich Revisited: Recent Linguistic Changes in an English Urban Dialect. In: *English World-Wide* 9, 33–49.
- Trudgill, Peter/Hannah, Jean (2002), *International English: A Guide to Varieties of Standard English*. 4th ed. London: Arnold.
- Visser, Frederikus Th. (1970–1978), *An Historical Syntax of the English Language*. 3 vols. Leiden: Brill.
- Webster's Dictionary of English Usage* (1989). Springfield MA: Merriam-Webster.

Christian Mair, Freiburg (Germany)

53. Corpus linguistics and dialectology

1. Introduction
2. Aim and scope of traditional dialectology
3. Traditional materials in dialectology
4. Examples of accessible dialect corpora
5. Analytical objectives in corpus-based dialectology
6. Some issues in dialect corpus analysis
7. Conclusion
8. Literature

1. Introduction

In contrast to sociolinguistics, dialectology and corpus linguistics have been rather uneasy bedfellows until relatively recently. In this article, we will focus on the use of modern corpora in the field of traditional dialectology – corpora, that is, which constitute principled, possibly computerized, and broadly representative collections of naturalistic spoken (and sometimes written) dialect material. This sets apart corpus-based dialectology from other approaches in empirical dialectology, which may be based on, e. g., dialect atlases and/or questionnaire data. The subject matter of traditional dialectology (as opposed to urban dialectology) is the distribution of linguistic features where, crucially, the primary parameter of variation is geographical distance and proximity between sampling locales. This means that this article will exclude from discussion corpora documenting variation between standard dialects (e. g. British English vs. American English), and we will also remain agnostic about corpus work on youth dialects and other sociolects, where variation is not stratified geographically but rather sociologically (but on this topic, cf. article 6 on corpora in sociolinguistics).

2. Aim and scope of traditional dialectology

The study of (non-standard) dialects has a venerable tradition going back at least to the nineteenth century (for a short overview, cf. Chambers/Trudgill 1998; for developments more specifically related to English, cf. Ihäläinen 1994). Interest in rural varieties of a language that were (or appeared to be) far removed from the standard language were kindled by the Neogrammarian dictum of the ‘exceptionlessness of sound change’ (Ausnahmslosigkeit der Lautgesetze), a dictum dialect data could support (or, indeed, refute) – many theoretically interesting intermediate stages of proposed changes could actually be documented with the help of dialect data. Especially isolated, rural dialects were seen as unspoilt or uncorrupted varieties, representing diachronically removed stages of the language in a purer form. This more theoretical interest in (at least certain) non-standard varieties seems to have been superseded by an interest in the dialect *per se*, often by dialect speakers themselves – witness the establishment of many regional dialect societies e. g. in England at the end of the nineteenth century, the most prominent

of which is perhaps the Yorkshire Dialect Society, founded in 1897 by the scholar Joseph Wright, himself a Yorkshireman and author of the monumental *English Dialect Dictionary* (1898–1905). (The society exists to this day, see <http://www.ydsociety.org.uk/>.) As, naturally, dialects are as little static as any other living linguistic system, observable tendencies of change were perceived as threatening the ‘purity’ or authenticity of the ‘real’ dialect, leading to efforts to preserve or at least record their ‘original’ state. (Actual sociocultural changes on a scale unknown so far, like increasing industrialization, urban migration, improved education etc. must have played an important role in shaping this perception as well, but cannot be explored further within the scope of this article. For a short overview, cf. Beal 2004.) It is little wonder, then, that traditional dialectologists saw the speech of non-mobile older rural males as the only ‘true’ representative of the dialect in question, since we know from modern sociolinguistic studies that features like reduced mobility, higher age, non-urbanity, or male gender typically contribute to the speaker employing a more conservative variety linguistically. No doubt due to the predilection with sound change, phonological investigations were at the forefront of dialectological interest, closely followed by perhaps one of the most notable features of regional variation, namely differences in lexis. This focus on collecting and classifying lexical items has no doubt contributed to the image of dialectologists ‘collecting butterflies’, i. e. engaging in a time-consuming activity which is however perceived as ultimately irrelevant to the wider community of linguists.

In this article, we will try to show that modern dialectology has moved far beyond these more traditional concerns, not only in the choice of subjects, materials and methods, but also in its relevance to theory-building and indeed to wider linguistic concerns. Not least the method of employing corpus-based materials has contributed to the resurrection and rehabilitation of this discipline.

3. Traditional materials in dialectology

Traditional dialect studies have relied mostly on questionnaires to elicit lexical material and dialect phonology. The German Georg Wenker was one of the pioneers of the postal questionnaire: in 1876 and 1877 he sent out a questionnaire to schoolmasters all over Germany and had them translate his 40 sample sentences into the local dialect. To this day, the so-called “Wenker sentences” are used in research, and have recently been made available electronically at the University of Marburg, Germany (in the form of the so-called DiWA, the Digital Wenker Atlas, <http://www.diwa.info/>), over a century after they were first collected. Towards the end of the nineteenth century, trained fieldworkers were sent out all over the country to record the answers to typical questionnaire questions in some kind of phonetic script; the pioneer of this form of data collection was the Swiss Jules Gilliéron, who sent out his fieldworker Edmond Edmont to cycle all over France. Edmont conducted around 700 interviews between 1886 and 1900. This type of nationwide dialectological fieldwork, especially from the first half of the twentieth century, typically resulted in linguistic atlases, and the French project (Gilliéron 1902–1910) became the benchmark for much subsequent work in Switzerland, Italy or Spain. Much material has also been published in the form of dialect dictionaries since the end of the nineteenth century (for an early specimen for British English dialects, cf. Wright 1898–

1905 mentioned above; for a more modern example involving U.S.-American dialects, cf. the DARE project: Hall (1985–2009 [scheduled]), some volumes of which are still in the process of being published). None of the early atlas projects recorded larger stretches of discourse. This changed with the post-war nation-wide survey of England (The Survey of English Dialects, or SED; all informants' responses are published as Orton/Halliday 1962–1964; Orton/Wakelin 1967–1968; Orton/Barry 1969–1971; Orton/Tilling 1969–1971), as well as the various North American projects, many conceived before the second world war, but conducting field work mostly afterwards (for an overview cf. <http://us.english.uga.edu/>). Especially towards the second half of the century, tape recorders (or their precursors) were increasingly used to aid fieldworkers' jobs in the large atlas projects. Many of the tapes have survived, and they sometimes contain slightly longer stretches of informants' speech. Recently, some of this material from the SED has been transcribed and made accessible (the SED recordings, cf. Klemola/Jones 1999). These recordings probably constitute the oldest dialectological material that we possess today that can be called an (electronic) corpus of non-standard speech. Dialectological work in the 1970s and 1980s used recording technology as a matter of fact, but the recorded material as a rule remained with the researchers. Much work in recording dialects stemmed from university theses (for English dialects, especially from the centres of dialectology, Sheffield and Leeds, e.g. Shorrocks 1980, published as Shorrocks 1999, but recently also Newcastle, cf. below; interesting work has also been done at the University of Helsinki in Finland, for a recent example, cf. Vasko 2005); however, the material itself has not generally been made public.

One exception is a corpus that concentrates on one dialect area, the Newcastle Electronic Corpus of Tyneside English NECTE, cf. <http://www.ncl.ac.uk/necte/>. NECTE combines older dialect material, collected for the Tyneside Linguistic Survey in the late 1960s and 1970s, with recordings from the more modern project on Phonological Variation and Change in Contemporary English (PVC) (Milroy/Milroy/Docherty 1997), collected between 1991 and 1994 in the same area. This corpus has recently been published, making possible some diachronic studies for this dialect area.

Another resource worth mentioning that is just being finished is the Helsinki Corpus of British English Dialects (HD), which aims to gather together the materials collected for the purpose of university theses mentioned above (cf. http://www.eng.helsinki.fi/varieng/team3/1_3_4_2_hd.htm). Due to lack of manpower and general resources, it has taken about 30 years to complete, meaning that materials date from the 1970s and 1980s. The Helsinki dialect corpus (HD) contains material from Lancashire, Cambridgeshire, Suffolk, Essex, the Isle of Ely, Devon and Somerset, so that some regional comparisons are now becoming possible. The corpus contains almost one million words.

Recently, a new resource has been collected, the Freiburg English Dialect corpus (FRED) (Kortmann et al. 2005; Anderwald/Wagner 2007, cf. also <http://www.anglistik.uni-freiburg.de/institut/lskortmann/FRED/>). FRED contains almost 2.5 million words taken from Oral History Projects, conducted in the 1970s and 1980s, orthographically transcribed, from six large dialect areas in the British Isles. At least for English, this for the first time makes possible regional comparisons across Great Britain. A sizable sub-corpus of FRED will be published on the third ICAME-CD (<http://icame.uib.no/corpora.html>).

As far as accessibility is concerned, the situation looks slightly better for German spoken language, because many researchers have deposited their material at the Institut

für Deutsche Sprache (Institute for German Language) at Mannheim, Germany (<http://www.ids-mannheim.de/>), and corpora and their transcriptions are open to the public in the Datenbank Gesprochenes Deutsch (DGD, or database of spoken German). While this collection, too, contains situational or text-type specific discourse, it is also a good source for material from Lower German, Middle German, and Upper German dialects collected since 1950. Austrian dialect researchers have also utilized the Austrian Phonogram recordings, held at the Austrian Audiovisual Research Archive (cf. <http://www.pha.oeaw.ac.at/>), samples of which are available commercially (Schüller 2003). Materials for most other languages seem to sadly lag behind the English (and, to a degree, German) vanguard, even where documenting the standard languages are concerned, not to mention non-standard varieties.

4. Examples of accessible dialect corpora

In the few corpora that are used in dialectology today we can distinguish several categories of materials that have been collected. In this section, we will discuss the use of traditional dialectological interviews, the use of more sociolinguistic materials, data from oral history projects, and participant interviews.

Dialectological “interviews”: in contrast to recordings of spontaneous speech, as employed in many sociolinguistic studies, dialectological corpora often consist of the recordings of more or less extended questionnaire sessions, as discussed above. The recording here was originally envisaged to aid fieldworkers’ memories, but might constitute relevant material for phonetic and lexical investigations. A good example of this typical dialectological material are the SED recordings (however, commercial publication, originally intended, has not been resolved yet); for some first studies on this material cf. Klemola (2002, 2003).

Some material contains traditional sociolinguistic interviews: here, the borderline to sociolinguistics becomes fuzzy. If one defines variationist sociolinguistic studies, as Chambers and Trudgill do, as “urban dialectology” (e. g. Chambers/Trudgill 1998), only the rural – urban axis serves to distinguish the two fields, a somewhat artificial distinction. (However, as pointed out above, the choice of informants in dialectology is still mainly guided by the older traditional dialect speaker, so-called NORMs, whereas urban dialectology aims at socially stratified samples of speakers). An example containing this kind of dialectological-sociolinguistic material would be the Northern Ireland Transcribed Corpus of Speech (NITCS), cf. Kirk (1990).

Oral History Projects: as mentioned above, some dialect projects have used information collected for different purposes, e. g. Oral History Projects (cf. FRED mentioned above, and also Huber (2003), who plans a corpus based on South Wales oral history material, encompassing ca. 3 million words for that dialect area alone). An advantage here is that the participants’ attention was genuinely on matters unrelated to language, thus avoiding the Observer’s Paradox to some degree (cf. Labov 1972). On the other hand, interviewers will have been interested in specific content only, not necessarily paying particular attention to reducing the formal distance between interviewer and interviewee, or trying to elicit particularly informal styles, which might be desirable for the investigation of non-standard features.

Participant interviews/conversations: some dialectological and sociolinguistic work has consisted of dialect speakers, sometimes trained as fieldworkers, recording the speech of other dialect speakers. Thus for example, the PVC project encouraged dyadic interaction without the presence of the investigators (Milroy/Milroy/Docherty 1997).

Finally, recordings of regionally restricted, spontaneous conversations as in the Corpus of London Teenage Speech (COLT) are as amenable to dialectological investigations as to sociolinguistic ones, and studies taking these materials as a basis probably straddle the border between dialectology and sociolinguistics (Andersen 2001; Stenström/Andersen/Hasund 2002).

5. Analytical objectives in corpus-based dialectology

The analysis of dialect corpora – much as the analysis of other corpus data – can serve a variety of linguistic objectives. In what follows, we will illustrate this point by discussing a number of case studies seeking to explore dialect corpora for the primary aim of (i) functional-typological analysis, (ii) variationist analysis and probabilistic model-building, (iii) historical inquiry, and (iv) formalist/generative theory-building (or theory-rejection). In point of fact, the bulk of corpus-based studies in dialectology have addressed more than one objective at the same time, albeit with differing emphases; a survey of the literature reveals that most contemporary corpus-based dialectology does not stop at mere dialect description but offers some added theoretical and interpretational value. There are only some exceptions to this tendency: consider Jones (1985) for a corpus-based description of Tyneside English and Beal/Corrigan (2006) for a corpus-based descriptive follow-up on Tyneside negation specifically, or – in the realm of German dialectology – Patocka (1997) for a discussion of word order phenomena in Bavarian dialects spoken in Austria, drawing on a corpus covering assorted recordings in the Phonogram archives at the Austrian Academy of Sciences (mentioned above). Still, dialect description – i. e. mapping the range and extent of variation, on the basis of naturalistic corpus data, of one or more dialect features in one or more dialect areas – necessarily precedes any interpretation of naturalistic dialect data in the light of a particular theoretical framework.

5.1. Functional-typological analysis

Functional-typological approaches to dialect data endeavor to marry functional typology to dialectological investigation. This means, in a nutshell, that the observable patterns of language-internal variation are analyzed and interpreted in terms of the same empirical and interpretational apparatus which is familiar from the typological study of large-scale cross-linguistic variation (cf. Kortmann 2004 for a collection of papers in this spirit).

Thus, Anderwald (2002, 2003) explores dialectal material in the spoken-demographic section of the *British National Corpus* (in which some texts are tagged for dialect area) to demonstrate that non-standard negation patterns are actually predicted by general cognitive and typological principles: the pervasiveness or invariant *don't* and *ain't* (e. g.

he don'tain't like me) in non-standard English, for example, is argued to bring non-standard English in line with cross-linguistic markedness criteria. In a quite similar vein, Herrmann (2005) presents evidence drawn from FRED that while standard English does not conform with Keenan/Comrie's (1977) Noun Phrase Accessibility Hierarchy, English dialects do in that they allow subject gapping (e. g. *the man __ lived there*) in addition to object gapping (e. g. *the man __ I saw*). As it turns out, the Accessibility Hierarchy is also involved in dialectal pronoun systems and (pronominal) gender marking: Wagner (2004) discusses pronoun usage in traditional dialects in the Southwest of England as well as in Newfoundland, drawing on FRED for the Southwest of England and on a corpus of transcripts from the *Memorial University of Newfoundland Folklore and Language Archive* (MUNFLA). In the traditional dialects Wagner investigates, count nouns have historically been referred to by 'gendered' pronouns (*he, she*) whereas mass nouns have received *it*. This is a gender assignment system which is increasingly being crowded out by the system of Standard English. Crucially, Wagner shows that the way in which the traditional system is breaking down follows a path originally suggested by Ihälainen (for instance, 1991): Standard English forms start out from less accessible positions in the Accessibility Hierarchy and diffuse through the hierarchy to more accessible positions. This may point to a more general, possibly cross-linguistically valid, mechanism of language change. Also in regard to pronominal usage, Geyer (2003) draws on a relatively small corpus (covering about 40 minutes of transcribed speech) of Hetzlerisch Franconian, a dialect of German spoken in the village of Hetzel near Nuremberg, to document the inventory of phoric pronouns in that dialect (cf. colloquial German *der Mann kam in die Bäckerei, und der hatte einen Hut auf*). Geyer claims that phoric pronouns – and thus the interface between syntax and information structure – in Hetzlerisch Franconian ought to be viewed against the backdrop of crosslinguistic, typological variation.

Corpus-based dialectology can thus be embedded in a theoretical framework that seeks to predict, and explain, dialectal variation by well-known cross-linguistic parameters of variation.

5.2. Variationism and probabilistic model-building

Corpus studies in this line of dialectological inquiry explore corpus data of traditional dialects with quantitative techniques similar to those of urban dialectology and sociolinguistics (cf. article 6). This means that the main theoretical interest is in the (probabilistic) constraints that govern the choice between two variant forms, or between a dialectal variant and a standard variant. Thus, Hernández (2002) is concerned with untriggered *self-forms* (as in *for somebody like myself* instead of *for somebody like me*) in non-standard English. Investigating the *Northern Ireland Transcribed Corpus of Speech* (NITCS) and the spoken section of the *British National Corpus* as well as questionnaire data, Hernández establishes that the phenomenon is not, as has been previously claimed in the literature, a feature that is typical merely of (dialectal) Irish English and that the choice between untriggered *self-forms* and their standard alternatives is governed by a hierarchy of language-internal and language-external constraints.

As a genuine study in probabilistic grammar, Pietsch (2005) addresses verbal agreement in (traditional) northern dialects of England. Since Middle English times, Northern

English dialects have been displaying the so-called Northern Subject Rule: invariant verbal *-s* occurs anywhere (e. g. *birds sings*) except when the verb is directly adjacent to a simple personal pronoun (e. g. **I sings*). This system used to be categorical in traditional dialects but is now highly variable and competes with Standard English verbal concord. On the basis of dialect atlas material from the SED, on the one hand, and of corpus data – FRED and the NITCS – on the other hand, Pietsch aims to uncover the factors governing the occurrence or non-occurrence of verbal *-s* in these dialects. In addition to qualitative analysis, Pietsch marshals multivariate analysis methods (more specifically, Variable Rule Analysis) to investigate the probabilistic, intralinguistic constraints that govern the inherent variability of the variable in corpus data. It turns out that verbal concord in Northern varieties in the British Isles is a hybrid system that is best interpreted in terms of usage-based models of grammatical competence, such as cognitive grammar.

Exploiting the pervasive variation in dialect data, among other data sources, as a research site for probabilistic and psycholinguistic model building, Szemrecsanyi (2006) explores morphosyntactic persistence, i. e. the tendency of speakers to re-use linguistic material that they have used or heard before. The phenomenon, which is partly psycholinguistic and partly discourse-functional in nature, plays particular methodological havoc with those corpus-based approaches that rely on naturalistic data where dialect speakers potentially echo standard variants used by their interlocutors. By way of multivariate analyses of a number of corpora of English, among them the dialect corpus FRED, with regard to a number of well-known alternations in the grammar of English, Szemrecsanyi presents evidence that persistence is indeed a major probabilistic constraint on the way (dialect) speakers make linguistic choices. Needless to say, as a psycholinguistic constraint persistence is not specific to English, or English dialects: for example, in Romance dialects where plural expression is variable – such as Brazilian Portuguese (cf. Scherer/Naro 1991) and Puerto Rican Spanish (cf. Poplack 1980), corpus data exhibit similar parallelism effects, and Travis (2007) draws on corpus material to demonstrate that null subjects in two dialects of New World Spanish, Colombian Spanish and New Mexican Spanish, are more likely when another null subject expression was recent.

With a similar interest in recency effects and the dialogic interdependence between interviewer and interviewee utterances, Hollmann/Siewierska (2006) offer a methodological outline of how accommodation theory and, therefore, the concept of sociolinguistic salience can be brought to bear on dialect corpus data (more specifically, in a corpus of oral history transcripts in the Lancashire dialect). Straightforwardly enough, Hollmann/Siewierska suggest, first, to determine the first variable form in the text, and then to calculate whether the likelihood of the interviewee using the dialectal variant increases significantly towards the end of the text; if it does, it is safe to assume that the interviewee accommodates to the interviewer.

Making extensive use of methods and explanatory patterns along the lines of modern variationist sociolinguistics, Sali Tagliamonte and her co-workers, in a series of recent studies, have sought to explore corpus data sampling traditional dialects – with a particular focus on relic areas – in a number of locales in Great Britain and the Americas (note, though, that the corpora subject to analysis are generally not available to the wider research community). This line of research, which rigorously combines the methodological machinery of urban dialectology with traditional dialect data, has investigated the following phenomena in English dialect data:

- *was/were* variation (e. g. *You were hungry but he were thirsty*) (Tagliamonte 1998; Tagliamonte/Smith 2000);
- the habitual past (e. g. *he would always dance* vs. *he used to always dance*) (Tagliamonte/Lawrence 2000);
- *come/came* variation (e. g. *when I come home that day* vs. *when I came home that day*) (Tagliamonte 2001);
- NEG/AUX contraction (e. g. *he isn't going* vs. *he's not going*) (Tagliamonte/Smith 2002);
- markers of stative/possessive meaning (e. g. *He has/has got/got a car*) (Tagliamonte 2003);
- zero complementation (e. g. *he shows that/Ø he can do it*) (Tagliamonte/Smith 2005);
- variation between relative markers (e. g. *the man that/Ø/as/who ... I saw*) (Tagliamonte/Smith/Lawrence 2005);
- (t,d) deletion (e. g. *I was told* vs. *I was tol/Ø*) (Tagliamonte/Temple 2005).

Characteristically, quantitative findings – especially factor weights in Variable Rule analysis and the resulting constraint rankings – are often interpreted in terms of historical or comparative research questions, where similar constraint rankings in different locales are taken to indicate genetic relatedness (cf. Tagliamonte/Temple 2005, 84). In addition, Tagliamonte and associates also occasionally interpret findings in terms of grammaticalization theory or in regard to how and why incoming forms may (or may not) diffuse through the community (for instance, Tagliamonte/Smith/Lawrence 2005). Sometimes, the variable portfolio includes external variables such as speaker age, which allows for tracking changes in apparent time (for instance, Tagliamonte 1998). In Tagliamonte/Temple (2005), (t,d) deletion is discussed against the backdrop of phonological theory.

In sum, dialect corpora can serve as a rich resource for establishing and benchmarking probabilistic grammars across different geographic locales. The patterns that this approach yields can be interpreted in terms of genetic relatedness, or along the lines of more general processing-related constraints that leave their mark on languages and dialects.

5.3. Historical linguistics

Given dialectology's traditional orientation towards historical linguistics and diachronic explanation, it should surprise no one that much corpus-based dialectology is still interested, to varying degrees, in the diachronic evolution and synchronic areal diffusion of linguistic forms. Exactly along these lines, Klemola (1996) investigates non-standard affirmative periphrastic *do* (e. g. *we did always go to school*) in traditional dialects in the Southwest of England, drawing on an oral history corpus and a corpus of SED field-worker material. Klemola shows that unstressed *do* periphrases do not actually differ from simple present tense forms but that past tense *did* often carries habitual aspect. Klemola then takes these corpus findings as a starting point to discuss the historical development of *do*-support in English. Jones/Tagliamonte (2004) further add to our knowledge about the history of periphrastic *do* by considering the constraints operating on preverbal *did* in two corpora, one sampling Samaná English (a variety of English

spoken on the Northeastern peninsula of the Dominican Republic, a community which was originally settled by African American ex-slaves), and the other sampling Somerset English, the traditional dialect which is spoken in England's Southwest. Jones/Tagliamonte establish that the internal constraints governing periphrastic *do* variation work in a curiously similar way in the two dialect corpora and that, moreover, the constraint ranking is exactly the same. Jones/Tagliamonte thus suggest that even relic forms, such as preverbal *did*, follow "diachronic patterns in systematic linguistic conditioning" (2004: 119), and that preverbal *did* "continues to maintain a complex set of constraints [...] that can be traced in the history of English" (2004: 119).

In much the same historical-evolutionary spirit, Pusch (2001) follows up on a historical-evolutionary research question (cf. Pusch 2000 for a partial summary in English), exploring enunciative particles (such as preverbal *que*) in varieties of Gascony Occitan on the basis of the *Corpus Occitano-Gascon*. Through a quantitative and qualitative analysis of the synchronic distribution of the phenomenon in corpus data and by additionally presenting evidence from other languages, Pusch seeks to illuminate the grammaticalization processes which the family of enunciative particles has been undergoing. Pusch argues that the genesis of the phenomenon is best explained in terms of functional and communicative pressures.

Dialect corpora have also been used to shed light on the socio-historical genesis of dialects and language varieties: with an overarching interest in the linguistic consequences of industrialization as a social phenomenon, Grosse et al. (1987), for example, present a historical corpus sampling a variety of dialect materials (such as personal letters, postcards, etc., dating primarily from the second half of the nineteenth century) in order to explore the evolution of Ruhrdeutsch (Ruhr area German).

In sum, what all these case studies have in common is that they draw on naturalistic dialect data to document the genesis, evolution, and/or the resulting synchronic layering of non-standard – or former non-standard – linguistic forms. Thus, corpus-based dialectology is employed to pursue *per se* historical research questions. It seems worth pointing out in this connection that such dialectological inquiry may also serve to trace the consequences of colonial transplantation: Elsig/Poplack (2006), for example, rely on corpora of Quebec French (the *Récits du français québécois d'autrefois* and the *Ottawa-Hull French Corpus*) to document the history of question formation in Québec dialects of French vis-à-vis French French.

5.4. Formalist/generative theory building

If dialectology and corpus linguistics can be said to be strange bedfellows, formalist approaches and corpus linguistics are a methodically even more outlandish pairing. Thanks to the recent theoretical interest in syntactic microvariation (cf. some of the papers in Barbiers/Cornips 2002 and Cornips/Corrigan 2005), however, the past few years have seen a number of formalist linguists exploiting naturalistic dialect corpus data as a source for authentic examples. To name but a few representative examples: Vangsnes (2005, 221) searches the *Oslo Corpus of Tagged Norwegian Texts* for certain *wh*-pronouns; Westergaard (2003) studies a corpus sampling child and adult data of the Tromsø dialect to demonstrate that this dialect exhibits both V2 and V3 word order; Ledgeway (2005)

draws on a corpus of Southern Italian dialect texts to shed light on the dual complementizer system of those dialects (this particular study even features some quantitative analyses). Other formal analysts have relied on dialect corpora as a means to check on the reliability of dialect atlas data (cf., for instance, Cornips 2002 on variation between infinitival complementizers in Dutch dialects).

As for formalist theory rejection (an exercise which certainly comes more easily to most corpus linguists), of course, corpus data have been used to demonstrate *ex negativo* that formalist accounts of dialect phenomena are insufficient. For example, Pietsch (2005) can be seen as an extended empirical argument that formal approaches to the Northern Subject Rule cannot explain the observable range of variability in corpus data, and Anderwald (2008) uses non-standard past tense forms to argue for a usage-based model of language processing.

6. Some issues in dialect corpus analysis

In a number of ways, corpus-based dialectology is subject to methodological and technical challenges that arise out of the particular nature of the data studied. For instance, mapping dialect data to digitized text is not a trivial task.

In the absence of general guidelines many idiosyncratic solutions exist (but cf. the sensible guidelines in Tagliamonte 2006). A purely phonetic transcription, while feasible in principle, is prohibitively labor-intensive to generate. In addition, a narrow phonetic transcription would make impossible the automatic retrieval of most phenomena – something that corpus linguistics is, after all, designed to do. A corpus search for a particular form would become extremely tedious, if not impossible, because each potential phonological or phonetic variant form would have to be considered separately, many of which may not be known to the researcher without extensive supplemental information. Indeed, for many studies this kind of detail is not only unnecessary, but might be a direct hindrance. Not surprisingly, then, most dialect corpora do not contain a phonetic or phonemic transcription of the data (but cf. the questionnaire-based Dutch SAND project, which works with several tiers, one of which is a phonetic transcription, cf. Barbiers et al. 2005). Researchers are thus usually left with the other bad alternative: creating some kind of orthographic representation. The general problem is that this means adapting the standard orthography to some considerable degree. It is clear that the accepted codified orthography of any standard language has been optimized for the standard, although this optimal fit may be some past stage, as the largely fossilized orthographies e.g. of French or English testify (also cf. article 30 on this issue).

Differences in lexis may constitute a first problem, as for purely dialectal words a consistent spelling may never have been devised. Researchers involved in transcribing dialect data have therefore also always paid particular attention to possible dialect literature, where lexical items may be documented (if not consistently). A general guideline has been to exclude phonetic and phonological information in the dialect representation, but to include morphological information. (However, this distinction might require some in-depth knowledge of the phenomenon under investigation.)

Finally, only those phenomena should be included that are genuine dialect features, rather than very general features of allegro speech on the one hand, or simple “eye

dialect” features on the other. Eye dialect features are meant to somehow convey, by means of orthography, the special quality and subjective phonological “feel” of speech (consider, in the case of English, <Whatcher thinkin?> for *What are you thinking?*). Thus, normal processes of spoken language (not necessarily dialectal) are rendered orthographically, something which does not contribute to linguistic heuristics, but only serves to degrade and stigmatize the speaker in question (on this last topic, cf. especially Preston 2000).

Ideally, all decisions made during the transcription process should be documented in a transcription protocol and published with the material at hand, although this is rarely the case. It is fair to say, then, that current corpus technology – as employed in corpus-based dialectology – is not especially well geared to deal with dialect data.

Another concern in corpus-based dialectology is sampling. For somewhat pragmatic reasons (that is, to obtain speech characterized by a sufficiently large number of dialect features) and to ensure comparability to older materials, many English dialect corpora, as pointed out above, heavily rely on sampling non-mobile old rural males as the most traditional, broadest dialect speakers. Yet, this design choice plays havoc with a number of sociolinguistic research questions that would require a balanced, representative, stratified database – which, however, would be less likely to yield a sufficient number of dialect features.

This leads us to a final, more general issue. All the limitations and troubles that are inherent to any corpus approach to language are, in a number of ways, even more clearly apparent in corpus-based dialectology. Thus, for instance, many phenomena theoretically interesting to dialectologists are excruciatingly rare in corpus data. Note that the bulk of corpus-based dialectology research has sought to investigate dialect data from a morphosyntactic perspective. It is well-known that the study of morphological or syntactic phenomena necessitates much larger databases of naturalistic data than the study of, e. g., phonology or lexis. A concrete example may illustrate this problem: double modal constructions are known to recur in a number of non-standard varieties of English, for example in Scottish English (cf. Brown 1991). Yet, the FRED corpus, after all containing 2.5 million words from all major dialect areas in Great Britain including Scotland and the North, yields but one (!) clear example of a double modal construction (*it might should tell you on the tickets how much luggage you're allowed to take*, FRED MLN_005), and one fairly questionable example (*because you said a Italian would could stab you in the back*, FRED DEV_003). With the pragmatic contexts that license double modals being so rare, then, the conclusion is that a thorough analysis of double modal constructions would probably require a dialect corpus spanning several hundreds of millions of words – a size which is, needless to say, illusionary. Similar frequency issues (which are by no means particular to corpus-based dialectology; cf. article 37) bedevil the corpus-based study of other interesting but rare dialect features as well: Hollmann/Siewierska (2006), for instance, note that material from no less than five corpora (among them, FRED and the BNC) is still insufficient to conclusively study the contexts of ditransitive verb complementation (of the type *He gave it me* or *He gave me it*) in the Lancashire dialect of English; in the same vein, Tagliamonte (1998) is unable to quantitatively investigate collective subjects as a determinant of *was/were* variation in York English since in spite of the considerable total size of her corpus, collective nouns are not sufficiently attested.

7. Conclusion

Corpus-based dialectology can certainly constitute a very important first step in the study of micro-variation (intra-, rather than inter-speaker variation). Hollmann/Siewierska (2006) argue that corpus-based dialectology must be complemented by other methods such as elicitation tasks and attitude questionnaires. In the case of double modals, for example, it is often claimed that elicitation tasks and questionnaires tapping speakers' intuitions are needed to determine the exact syntactic, not to mention pragmatic circumstances determining the possible structure, combinations as well as the use of these constructions (cf. Montgomery 1998 for an overview). To date, however, follow-up studies employing a different methodology after a corpus investigation are rarely carried out. Nevertheless, it should be pointed out that despite methodological drawbacks and theoretical problems, the use of corpora in dialectology is still an under-developed and under-researched area that merits much more attention. We are only just beginning to tap this rich resource of natural language data, which might enable us to increase our knowledge of the constraints and parameters of linguistic variation considerably. Especially the use of sufficiently large corpora of dialect speech might allow us to consider intra-linguistic variation in a much less haphazard and idiosyncratic way than has been done so far, and might thus constitute one of the most promising avenues for further research.

8. Literature

- Andersen, G. (2001), *Pragmatic Markers and Sociolinguistic Variation*. Amsterdam/Philadelphia: Benjamins.
- Anderwald, L. (2002), *Negation in Non-standard British English: Gaps, Regularizations, Asymmetries*. (Studies in Germanic Linguistics.) London/New York: Routledge.
- Anderwald, L. (2003), Non-standard English and Typological Principles: The Case of Negation. In: Rohdenburg, G./Mondorf, B. (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 507–530.
- Anderwald, L. (2008), *The Morphology of English Dialects: Verb-formation in Non-standard English*. Cambridge: Cambridge University Press.
- Anderwald, L./Wagner, S. (2007), FRED – the Freiburg English Dialect Corpus. In: Beal, J./Corrigan, K. P./Moisl, H. (eds.), *Creating and Digitizing Language Corpora*, Vol. 1: *Synchronic Databases*. London: Palgrave Macmillan, 35–53.
- Barbiers, S./Bennis, H./De Vogelaer, G./Devos, M./van der Ham, M./Haslinger, I./van Koppen, M./van Craenenbroeck, J./van den Heede, V. (eds.) (2005), *Syntactic Atlas of the Dutch Dialects (Sand)*. Amsterdam: Amsterdam University Press.
- Barbiers, S./Cornips, L. (2002), *Syntactic Microvariation*. Electronic publication of the Meertens Instituut, available at: <http://www.meertens.knaw.nl/books/synmic/>.
- Beal, J. C. (2004), *English in Modern Times, 1700–1945*. London: Arnold.
- Beal, J. C./Corrigan, K. P. (2005), *No, Nay, Never: Negation in Tyneside English*. In: Iyeiri, Y. (ed.), *Aspects of English Negation*. Amsterdam: Benjamins, 139–157.
- Brown, K. (1991), Double Modals in Hawick Scots. *Dialects of English: Studies in Grammatical Variation*. In: Trudgill, P./Chambers, J. (eds.), *Dialects of English: Studies in Grammatical Variation*. London/New York: Longman, 74–103.
- Chambers, J. K./Trudgill, P. (1998), *Dialectology*. Cambridge: Cambridge University Press.

- Cornips, L. (2002), Variation between the Infinitival Complementizers *om/voor* in Spontaneous Speech Data Compared to Elicitation Data. In: Barbiers/Cornips 2002, 75–96.
- Cornips, L./Corrigan, K. P. (eds.) (2005), *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam/Philadelphia: Benjamins.
- Elsig, M./Poplack, S. (2006), Transplanted Dialects and Language Change: Question Formation in Québec. In: *Penn Working Papers in Linguistics* 12(2), 77–90.
- Geyer, K. (2003), *Hetzlerisch: Dokumentation spontansprachlicher Texte und grammatische Analyse der phorischen Pronomina im ostfränkischen Dialekt des Dorfes Hetzles*. München: Lincom Europa.
- Gilliéron, J. (1902–1910), *Atlas linguistique de la France*. Paris: Champion.
- Grosse, S./Grimberg, M./Hölscher, T./Karweick, J./Kuntz, H. (1987), Sprachwandel und Sprachwachstum im Ruhrgebiet des 19. Jahrhunderts unter dem Einfluß der Industrialisierung. In: *Zeitschrift für Dialektologie und Linguistik* 54(2), 202–221.
- Hall, J. H. (ed.) (1985–2009 [scheduled]), *The Dictionary of American Regional English*. Cambridge, MA: Harvard University Press.
- Hernández, N. (2002), A Context Hierarchy of Untriggered Self-forms in English. In: *Zeitschrift für Anglistik und Amerikanistik* 50(3), 269–284.
- Herrmann, T. (2005), Relative Clauses in English Dialects of the British Isles. In: Kortmann et al. 2005, 21–124.
- Hollmann, W./Siewierska, A. (2006), Corpora and (the Need for) Other Methods in a Study of Lancashire Dialect. In: *Zeitschrift für Anglistik und Amerikanistik* 54(1), 21–34.
- Huber, M. (2003), The Corpus of English in South-East Wales and its Synchronic and Diachronic Implications. In: Tristram, H. L. C. (ed.), *The Celtic Englishes III*. Heidelberg: Winter, 182–200.
- Ihalainen, O. (1991), On Grammatical Diffusion in Somerset Folk Speech. In: Trudgill, P./Chambers, J. (eds.), *Dialects of English: Studies in Grammatical Variation*. London/New York: Longman, 104–119.
- Ihalainen, O. (1994), The Dialects of England since 1776. In: Burchfield, R. (ed.), *Cambridge History of the English Language*. Vol. 5: *English in Britain and Overseas: Origins and Development*. Cambridge: Cambridge University Press, 197–274.
- Jones, M./Tagliamonte, S. (2004), From Somerset to Samaná: Preverbal *did* in the Voyage of English. In: *Language Variation and Change* 16, 93–126.
- Jones, V. (1985), Tyneside Syntax: A Presentation of Some Data from the *Tyneside Linguistic Survey*. In: Viereck, W. (ed.), *Focus on England and Wales*. Amsterdam: Benjamins, 163–178.
- Keenan, E./Comrie, B. (1977), Noun Phrase Accessibility and Universal Grammar. In: *Linguistic Inquiry* 8(1), 63–99.
- Kirk, J. M. (1990), *Northern Ireland Transcribed Corpus of Speech*. Colchester: Economic and Social Research Council Data Archive, University of Essex.
- Klemola, J. (1996), Non-standard Periphrastic *do*: A Study in Variation and Change. Unpublished PhD thesis. Essex: University of Essex, Department of Language and Linguistics.
- Klemola, J. (2002), Continuity and Change in Dialect Morphosyntax. In: Kastovsky, D./Kaltenböck, G./Reichl, S. (eds.), *Anglistentag 2001 Wien*. Trier: Wissenschaftlicher Verlag Trier, 47–56.
- Klemola, J. (2003), Personal Pronouns in the Traditional Dialects of the South West of England. In: Tristram, H. L. C. (ed.), *The Celtic Englishes III*. Heidelberg: Winter, 260–275.
- Klemola, J./Jones, M. J. (1999), The Leeds Corpus of English Dialects-project. In: *Leeds Studies in English* 30, 17–30.
- Kortmann, B. (ed.) (2004), *Dialectology Meets Typology: Dialect Grammar from a Cross-linguistic Perspective*. Berlin/New York: Mouton de Gruyter.
- Kortmann, B./Herrmann, T./Pietsch, L./Wagner, S. (2005), *A Comparative Grammar of English Dialects*. Berlin/New York: Mouton de Gruyter.
- Labov, W. (1972), *Sociolinguistic Patterns*. Philadelphia: Philadelphia University Press.

- Ledgeway, A. (2005), Moving through the Left Periphery: The Dual Complementiser System in the Dialects of Southern Italy. In: *Transactions of the Philological Society* 103(3), 339–396.
- Milroy, L./Milroy, J./Docherty, G. (1997), *Phonological Variation and Change in Contemporary Spoken British English: Final Report to the UK Economic and Social Research Council, grant No. R000234892*. Department of Speech, University of Newcastle-upon-Tyne.
- Montgomery, M. B. (1998), Multiple Modals in LAGS and LAMSAS. In: Montgomery, M. B./Nunnally, T. (eds.), *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*. Tuscaloosa/London: University of Alabama Press, 90–122.
- Orton, H./Barry, M. V. (eds.) (1969–1971), *Survey of English Dialects. The West Midland Counties*. Leeds: Arnold.
- Orton, H./Halliday, W. J. (eds.) (1962–1964), *Survey of English Dialects. The Six Northern Counties and the Isle of Man*. Leeds: Arnold.
- Orton, H./Tilling, P. M. (eds.) (1969–1971), *Survey of English Dialects. The East Midland Counties and East Anglia*. Leeds: Arnold.
- Orton, H./Wakelin, M. F. (eds.) (1967–1968), *Survey of English Dialects. The Southern Counties*. Leeds: Arnold.
- Patocka, F. (1997), *Satzgliedstellung in den bairischen Dialekten Österreichs*. Frankfurt am Main: Peter Lang.
- Pietsch, L. (2005), *Variable Grammars: Verbal Agreement in Northern Dialects of English*. Tübingen: Niemeyer.
- Poplack, S. (1980), The Notion of the Plural in Puerto Rican Spanish: Competing Constraints on (s) Deletion. In: Labov, W. (ed.), *Locating Language in Time and Space*. New York: Academic Press, 55–67.
- Preston, D. (2000), ‘Mowr and Mowr Bayud Spellin’: Confessions of a Sociolinguist. In: *Journal of Sociolinguistics* 4, 614–621.
- Pusch, C. D. (2000), The Attitudinal Meaning of Preverbal Markers in Gascon: Insights from the Analysis of Literary and Spoken Language Data. In: Andersen, G./Fretheim, T. (eds.), *Pragmatic Markers and Propositional Attitude*. Amsterdam: Benjamins, 189–206.
- Pusch, C. D. (2001), *Morphosyntax, Informationsstruktur und Pragmatik: Präverbale Marker im gaskognischen Okzitanisch und in anderen Sprachen*. Tübingen: Gunter Narr.
- Scherre, M. M. P./Naro, A. J. (1991), Marking in Discourse: “Birds of a Feather”. In: *Language Variation and Change* 3(1), 23–32.
- Schüller, D. (ed.) (2003), *Tondokumente aus dem Phonogrammarchiv der Österreichischen Akademie der Wissenschaften: ‘Dazähl’n’*, [CD]. Wien: Veröffentlichungen der Österreichischen Akademie der Wissenschaften.
- Shorrocks, G. (1980), *A Grammar of the Dialect of Farnworth and District*. Sheffield: University of Sheffield.
- Shorrocks, G. (1999), *A Grammar of the Dialect of the Bolton Area*. Frankfurt am Main etc.: Peter Lang.
- Stenström, A.-B./Andersen, G./Hasund, I. K. (2002), *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings*. Amsterdam/Philadelphia: John Benjamins.
- Szmrecsanyi, B. (2006), *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin/New York: Mouton de Gruyter.
- Tagliamonte, S. (1998), *Was/were Variation across the Generations: View from the City of York*. In: *Language Variation and Change* 10(2), 153–191.
- Tagliamonte, S. (2001), *Come/came Variation in English Dialects*. In: *American Speech* 76(1), 42–61.
- Tagliamonte, S. (2003), ‘Every Place Has a Different Toll’: Determinants of Grammatical Variation in a Cross-variety Perspective. In: Rohdenburg, G./Mondorf, B. (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 531–554.
- Tagliamonte, S. (2006), *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

- Tagliamonte, S./Lawrence, H. (2000), 'I used to Dance, but I don't Dance Now': The Habitual Past in English. In: *Journal of English Linguistics* 28, 324–353.
- Tagliamonte, S./Smith, J. (2000), Old was, New Ecology: Viewing English through the Sociolinguistic Filter. In: Poplack, S. (ed.), *The English History of African American English*. Malden, MA: Blackwell, 141–171.
- Tagliamonte, S./Smith, J. (2002), 'Either it isn't or it's not': NEG/AUX Contraction in British Dialects. In: *English World-Wide* 23(2), 251–281.
- Tagliamonte, S./Smith, J. (2005), 'No Momentary Fancy!' The Zero 'Complementizer' in English Dialects. In: *English Language and Linguistics* 9(2), 289–309.
- Tagliamonte, S./Smith, J./Lawrence, H. (2005), 'No Taming the Vernacular! Insights from the Relatives in Northern Britain. In: *Language Variation and Change* 17, 75–112.
- Tagliamonte, S./Temple, R. (2005), New Perspectives on an Ol' Variable: (t,d) in British English. In: *Language Variation and Change* 17, 281–302.
- Travis, C. E. (2007), Genre Effects on Subject Expression in Spanish: Priming in Narrative and Conversation. In: *Language Variation and Change* 19(2), 1–35.
- Vangsnes, Ø. A. (2005), Microparameters for Norwegian Wh-grammars. In: *Linguistic Variation Yearbook* 5(1), 187–226.
- Vasko, A-L. (2005) UP CAMBRIDGE. *Prepositional Locative Expressions in Dialect Speech: A Corpus-based Study of the Cambridgeshire Dialect*. Helsinki: Société Néophilologique.
- Wagner, S. (2004), 'Gendered' Pronouns in English Dialects – a Typological Perspective. In: Kortmann 2004, 479–496.
- Westergaard, M. R. (2003), Word Order in Wh-questions in a North Norwegian Dialect: Some Evidence from an Acquisition Study. In: *Nordic Journal of Linguistics* 26, 81–109.
- Wright, J. (1898–1905), *The English Dialect Dictionary*. Oxford: Frowde.

*Lieselotte Anderwald, Kiel (Germany)
and Benedikt Szemrecsanyi, Freiburg (Germany)*

54. Contrastive corpus studies

1. Introduction
2. The revival of contrastive linguistics through corpus linguistics
3. Contrastive corpus analyses: A cross-section
4. The added value of multilingual corpora for contrastive studies
5. Links to other fields
6. Conclusion
7. Literature

1. Introduction

1.1. Survey

After a brief introduction and clarification of basic terms (in section 1), this article presents recent developments in modern corpus-based contrastive analysis and looks at the tools and corpora (easily) available for contrastive analysis (in section 2). It offers a

- Tagliamonte, S./Lawrence, H. (2000), 'I used to Dance, but I don't Dance Now': The Habitual Past in English. In: *Journal of English Linguistics* 28, 324–353.
- Tagliamonte, S./Smith, J. (2000), Old was, New Ecology: Viewing English through the Sociolinguistic Filter. In: Poplack, S. (ed.), *The English History of African American English*. Malden, MA: Blackwell, 141–171.
- Tagliamonte, S./Smith, J. (2002), 'Either it isn't or it's not': NEG/AUX Contraction in British Dialects. In: *English World-Wide* 23(2), 251–281.
- Tagliamonte, S./Smith, J. (2005), 'No Momentary Fancy!' The Zero 'Complementizer' in English Dialects. In: *English Language and Linguistics* 9(2), 289–309.
- Tagliamonte, S./Smith, J./Lawrence, H. (2005), 'No Taming the Vernacular! Insights from the Relatives in Northern Britain. In: *Language Variation and Change* 17, 75–112.
- Tagliamonte, S./Temple, R. (2005), New Perspectives on an Ol' Variable: (t,d) in British English. In: *Language Variation and Change* 17, 281–302.
- Travis, C. E. (2007), Genre Effects on Subject Expression in Spanish: Priming in Narrative and Conversation. In: *Language Variation and Change* 19(2), 1–35.
- Vangsnes, Ø. A. (2005), Microparameters for Norwegian Wh-grammars. In: *Linguistic Variation Yearbook* 5(1), 187–226.
- Vasko, A-L. (2005) UP CAMBRIDGE. *Prepositional Locative Expressions in Dialect Speech: A Corpus-based Study of the Cambridgeshire Dialect*. Helsinki: Société Néophilologique.
- Wagner, S. (2004), 'Gendered' Pronouns in English Dialects – a Typological Perspective. In: Kortmann 2004, 479–496.
- Westergaard, M. R. (2003), Word Order in Wh-questions in a North Norwegian Dialect: Some Evidence from an Acquisition Study. In: *Nordic Journal of Linguistics* 26, 81–109.
- Wright, J. (1898–1905), *The English Dialect Dictionary*. Oxford: Frowde.

*Lieselotte Anderwald, Kiel (Germany)
and Benedikt Szemrecsanyi, Freiburg (Germany)*

54. Contrastive corpus studies

1. Introduction
2. The revival of contrastive linguistics through corpus linguistics
3. Contrastive corpus analyses: A cross-section
4. The added value of multilingual corpora for contrastive studies
5. Links to other fields
6. Conclusion
7. Literature

1. Introduction

1.1. Survey

After a brief introduction and clarification of basic terms (in section 1), this article presents recent developments in modern corpus-based contrastive analysis and looks at the tools and corpora (easily) available for contrastive analysis (in section 2). It offers a

cross-section of corpus-based contrastive analyses to illustrate how useful a corpus approach can be in different linguistic sub-disciplines; it gives a few simple examples from German and English and emphasizes the student (researcher) perspective (in section 3). It summarizes the added value of multilingual corpora using German, English, French and Spanish EU texts as an illustration (in section 4). Finally, it puts contrastive corpus studies into the wider perspective of other fields (section 5).

1.2. Terminology

The term “contrastive linguistics” in this article refers to the comparison of different languages, although in an extended sense it could also be applied to (dialectal, social, historical, stylistic, etc.) varieties within a language. This broader view is taken in the comparison of natural language corpora and learner corpora (see article 15), of historical corpora (article 14) or multi-dimensional style analyses (article 38). Many corpus studies are inherently contrastive because they investigate text-type-specific or stylistic variation within the same language or different varieties of a language. Contrastive interlanguage studies (see Gilquin 2001) combine contrastive studies of mother-tongue varieties and the learner varieties of speakers with different mother-tongues (cf. also Granger 1996).

Contrastive corpus research in the narrow sense can be pursued using different types of bilingual or multilingual corpora:

- a translation corpus, where source texts from one language and target texts from a second language are aligned (see article 16) to form source language – target language pairs, juxtaposing a natural and a translated version of the same content and text or segment; or
- a comparable corpus, where natural texts from different languages with a similar topic, text-type and readership/audience are juxtaposed.

Unfortunately, the terminology used in contrastive corpus linguistics may be confusing: translation corpora are often referred to as parallel or comparable corpora (e.g. Baker 1993, 1996). Sometimes parallel corpora are considered as comprising two types, translation corpora and comparable corpora (Salkie 2003). Occasionally, the term translation corpus is applied to a text collection of translated texts without the originals, as in Baker’s Translational English Corpus and others (Olohan 2004, 59f.), but this is more relevant for translation studies than contrastive studies (cf. Baker 1995).

Moreover, sometimes all these corpora are called multilingual, although almost all of them are only bilingual. One truly multilingual corpus is the EU corpus (which contains legal and administrative texts from the European Union in many/all official languages of the Union, see also article 16), which is of course limited to European languages – and EU texts can be criticized for being unnatural and thus unsuitable for contrastive studies, as they have often been assembled by multinational author groups in French or English (so that the concept of an “original language” becomes blurred). Such politically motivated multilingual corpora, which often take the Canadian Hansards (in English and French) as their model, cover only relatively few text-types. Their obvious advantage is that they are easily available in electronic form, without copyright issues and usually translated by professionals.

The benefits and problems of translation-based data in contrastive linguistics have been widely discussed (e. g. Teubert 1996 vs. Mauranen 1997). The advantage of translation corpora is that the meaning in both texts is (almost) identical and that structural differences between source and target text should be due to contrasting language properties, conceptualisations and lexicalisations or grammaticalisations. Thus translation corpora seem the ideal basis for contrastive language studies. However, the target text may not reflect natural language since the translation process also plays a role and the translator may be influenced involuntarily by the source structures or change them using specific translation techniques (the specific translation variety is sometimes called *translationese*, cf. Baroni/Bernardini 2006). This is avoided when a comparable corpus of originals, natural texts in two or more languages with the same function, is analysed. Still, the issue arises whether the meaning of a construction in the two texts is really comparable in content and style. Thus a double strategy has developed over the last 15 years, comparing source-texts and target-texts in the same language, e. g. English source texts and German target texts as well as German source texts and English target texts (Johansson/Hofland 1994; Schmied 1994).

2. The revival of contrastive linguistics through corpus linguistics

2.1. Historical background

Comparing languages has a long tradition and although it played a major role in the development of (historical) philology in the 19th century, it was largely neglected during the first half of the 20th century. Contrastive linguistics had a short boom during the 1960s and 1970s, when it appeared attractive for applied purposes, especially for translation and language teaching. But the problem of finding a “*tertium comparationis*” for translations and the lack of predictive power (of learners’ errors) led to widespread disillusionment. Since the mid-1990s however, corpus-based contrastive work has shed new light on old contrastive linguistic issues and generated new questions in structural analysis as well as typological and stylistic perspectives.

Interestingly, the unification of corpus methodology and contrastive linguistics has not led to a new sub-discipline of contrastive corpus linguistics (as proposed by Aijmer/Altenberg 1996, 12), but rather to a new wave of corpus-based contrastive studies. Contrastive corpus linguistics is not a major keyword in introductions to corpus linguistics, but for contrastive specialists, corpus linguistics is a methodology that has penetrated most areas of (empirical) contrastive linguistics.

2.2. The corpus-based approach to contrastive linguistics

The new approach to contrastive linguistics was advocated by Sajavaara (1996), who proclaimed the failure of the earlier, structuralist contrastive linguistics and instead called for a socio-psycholinguistic basis, i. e. a cognitive, process-oriented approach that does not ignore the settings (*ibid.*, 21).

This cognitive orientation focuses on constructions where translators deviate from the structural equivalence (if they have a choice), since this reveals their awareness of optimising options (like explicitness, cf. Schmied/Schäffler 1996, 1997) or their awareness of target language preferences (cf. Schmied 1994, 1998).

Process-orientation, however, is not the only advantage of the corpus-linguistic approach to contrastive linguistics, it also adds a reliable quantitative dimension to contrastive linguistics and the related new approaches to typology including universals and grammaticalisation. With translation corpora, specialists not only have access to a vast array of examples (and do not have to invent or collect them impressionistically any more), but they can also describe gradable phenomena or language-specific preferences. This is particularly useful for closely related languages like English and German (and other members of the same language family or language type), whose structural inventory is almost identical but whose optimal choices may differ typically (e.g. German uses more modal adverbs at the expense of modal auxiliaries, compared to English, as exemplified below). Corpus-linguistics has freed contrastive linguistics from the need of native-speaker competence and the difficulty of direct equivalence since (ideally) the sheer number of occurrences makes individual “deviations” irrelevant. Now patterns or “rules” can be seen in high relative frequencies of structures (lexemes, collocations or colligations); exceptions are cases of low relative frequency, which may indicate rare or unnatural constructions. However, prototypical and non-prototypical, equivalent and non-equivalent constructions deserve linguistic analysis.

Over the last few years, corpus-based contrastive studies have been developed in various research centres: for English with Norwegian at Oslo (e.g. Johansson/Hofland 1994; Hasselgård/Oksefjell 1999; Johansson 2003, 2004), with Swedish at Lund (e.g. Johansson 1996; Altenberg 1999; Altenberg/Aijmer 2000), with German at Chemnitz, Dublin and Saarbrücken (e.g. Schmied 2004; Kenny 1999; Steiner 2005), with Portuguese at Oslo/Porto/Lisbon (e.g. Maia 2000; Frankenberg-Garcia 2004, 2005), with Spanish at León (e.g. Rabadán 2006), and with Italian at Bologna (e.g. Zanettin 2002; Bernardini 2000), for instance. Numerous relevant international conferences and their proceedings (e.g. Rábade/Doval Suárez 2002; Johansson/Oksefjell 1998; Borin 2002; Véronis 2000) bear witness to the rapid expansion of translation corpora. The journal *Languages in Contrast* is largely corpus-based and documents the progress made in the field.

2.3. Contrastive corpus processing

A major problem of all translation corpora is their alignment and processing (article 32), since for a reliable juxtaposition of structures “identical” text parts must be compared and calculated. For some smaller projects simple tools have been used that were developed for general use (e.g. ParaConc described in Barlow 1999). For most larger projects, specific tools have been developed (e.g. Hofland 1996; Ebeling 1998 for the Oslo multilingual project described below). Whereas simple tools only provide keyword-in-context (KWIC)-concordances and basic frequency statistics, more elaborate query systems offer combined search options with Boolean operators and restriction to subcorpora as well as diagrams for the visualisation of results. For the testing of hypotheses

concerning prototypical and non-prototypical equivalents the option (BUT) NOT is important ("List all cases where *help* is not translated as a form of *helfen* in German and *aider* in French"). This allows analysts to exclude some (frequent) constructions and to focus on the others (like *beitragen* "contribute" or *contribuer*).

These examples show that contrastive corpus studies do not require elaborate software, and so, that students and translators or other language service providers can make use of simple contrastive queries and their results. Even simple KWIC results from untagged corpora can be an interesting basis for language choices. Of course, lemmatizers and POS tagging programs are easily available nowadays, yet many users do not want to work on the data-bases (and many researchers cannot afford to do the intensive proof-reading necessary), but only receive quick results using a simple query interface and existing data-bases. For genre-specific structures, even parts of the existing standard corpora may be enough; in many cases specialised comparable mini-corpora are put together ad-hoc, analysed and discarded again when they have fulfilled their purpose.

2.4. Contrastive corpora as databases

The development of large multilingual corpora began in the early 1990s. The European Union, for example, funded two research projects, C.R.A.T.E.R. (McEnery/Oakes/Garside 1994) and Multext (Ide/Véronis 1994), both based on existing (highly specialised) parallel texts (from the International Telecommunications Union and the EU).

Since the big national reference corpora for European languages, which are usually modelled on the British National Corpus (BNC), are not really compatible (cf. Tamburini 2002 and article 20) despite the EAGLES guidelines developed over 10 years ago (the URL is given in the references), the following exemplary analyses have used smaller stratified translation corpora put together specifically for their purposes. Even then, few translation corpora are distributed freely nowadays, or can be consulted via the WWW. This is mainly due to copyright restrictions imposed by the publishers on literary texts, which corpus compilers would like to include, because traditional linguistic analysis is based on them. The English-German translation corpus and the multilingual EU corpus used below do not contain literary language but they are freely available on the WWW; other corpora are only accessible via intranet or completely off-line (cf. the URLs in section 7.2.).

Most multilingual corpora are based on the principles developed by Johansson/Hofland (1994) for the first Oslo Corpus: a stratified collection of contemporary (usually post-1990) originals of literary and non-literary texts and their translations, the latter often including political or EU texts. This corpus design has obvious advantages: natural and translated texts can be compared (if enough originals can be found in all categories, which is not always easy for languages with fewer speakers) and the translation quality is usually uncontroversial (except if, for instance, tourist brochures or users' manuals are included). But this design also has various disadvantages: the literary texts often cannot be used in searches over the WWW, the EU texts are not considered natural enough and translation options are not displayed, e.g. when several contrastive equivalents could be chosen by other translators. The only multilingual corpus that consists of multiple translations made explicitly is the Oslo Multilingual Corpus Project (Johansson

2002, 2004, 2007). Unfortunately, even this procedure of corpus-linguistic elicitation does not necessarily produce fully comparable results because translation (as a process in the translator's head) is of course a "black box" and full functional equivalence is impossible to ascertain. Generally, systematic annotation (semantic, pragmatic, POS tagging, etc.) of multilingual corpora has only just begun, and the issue of compatibility arises again on these other levels (see also article 55).

3. Contrastive corpus analyses: A cross-section

The following qualitative and quantitative analyses give examples from lexicology, grammar and discourse to demonstrate how contrasts between languages in complex phenomena like causativity and modality can be detected. This section places special emphasis on the student perspective and the difficulties of identifying comparative units, finding patterns and developing (restricted) queries (cf. Willemse et al. 2003). Finding patterns can be supported by corpus-linguistic statistics, but "seeing patterns" is a great accomplishment for students, not only because it helps them to distil their own "grammar" (unconsciously and consciously) when they are exposed to language usage, but also because it helps them to develop their skills in problem analysis and solution.

3.1. Contrastive lexicology: Demonstrating gradience and partial overlaps of close cognates

Most contrastive studies have focused on the lexicon (e.g. Altenberg/Granger 2002), since looking for key words in context can be done relatively easily. Frequency lists and collocational patterns, colligations or collexemes (article 43), and even patterns of lexical priming (Hoey 2005) can be obtained with relatively simple tools.

For related languages, contrastive analyses can help detect partial overlaps between close cognates like English *with* and German *mit* (e.g. Schmied 2002), to recognise lexeme splits or to distinguish between (obligatory) paradigmatic choices in a semantic field of one language (like the Spanish *ser/estar* for English *be*). The semantic structure of *think* comes out clearly when considering the equivalents in other Germanic languages, for instance. *Think* corresponds to three different verbs in Swedish, German and Dutch: in Swedish a distinction is made between *tänka* ('cognition'), *tycka* ('subjective evaluation') and *tro* ('belief') (Aijmer 1998); thus three main lexical entries appear to cover the space of one English lexeme – and a translation corpus is a better resource to make language users and learners aware of this than a bilingual dictionary because it shows gradience, overlaps and contexts more clearly. The following three examples from the same academic text in the bilingual English–German Translation Corpus available at <http://ell.phil.tu-chemnitz.de/search> illustrate the three meanings and three different translations in German (after the quotations the source is always indicated as the text-type ('ac' for academic), author and page): although *reden* "speak" in S1 is rather unusual for *conceive*, *nachdenken* for *think* in S2 is prototypical for *ponder*, and *annehmen* "assume" in S3 is possibly a clarification of *consider*.

(S1) It is actually a little simple minded to **think** of ‘the’ diameter of a supernova.

Es ist in der Tat ein wenig einfältig, von “dem” Supernovadurchmesser zu **reden**.

(ac/murdin: 461)

(S2) This is not such a very strange idea when you **think** about it.

Wenn man darüber **nachdenkt**, ist dies kein sehr ungewöhnlicher Gedanke. (ac/murdin: 737)

(S3) Most astronomers regard the claim with scepticism, and **think** that nothing really significant was seen.

Doch die meisten Astronomen betrachten diese Aussage mit Skepsis und **nehmen an**, daß nichts wirklich Signifikantes beobachtet worden ist. (ac/murdin: 825)

Different translation equivalents can also be used for disambiguating meanings in order to establish standard collocations. For instance, the German *Haushalt* is mainly a political term in the Chemnitz English–German Translation Corpus: generally, *Haushalt* has two equivalents in English, *household* and *budget(ary)*, the latter being restricted to political texts (Table 54.1). Thus text-type helps to disambiguate the two meanings, which can also be rendered as *home* and *fiscal* in English, respectively.

The example also demonstrates how important the stratification of corpora is to avoid generalising genre-specific usages.

Tab. 54.1: Occurrences of *Haushalt* in German and its two main equivalents in English (in the Chemnitz English–German Translation Corpus)

Text Type	German <i>Haushalt</i>	English <i>household</i>	English <i>budget</i>
EU Documents	78	8	50
Academic Texts	4	4	—
Public Speeches/Art.	12	1	9
Tourist Brochures	3	2	—
Total	97	15	59

3.2. Contrastive grammar: Analysing language-specific phenomena like stages in grammaticalisation

Because morphological affixes and function words have higher frequencies, contrastive grammar often has more convincing results than contrastive lexicology, but it has to rely more on wildcards (if the corpus is not lemmatised), especially for languages whose inflectional morphology is not as restricted as in English (like the German *helfe/hilfst/hilft/helfen/helft* for the simple present of *helfen* ‘to help’).

Although English and German share many structural features due to their common origin and common West Germanic areal type, the auxiliary – catenative – full verb cline has expanded considerably in English over the last few hundred years and continues to do so, e. g. in the case of *help*. *Help* is a catenative, i. e. a verb that takes non-tensed (infinitive, participle) clauses as complements (cf. Huddleston/Pullum 2002, especially 1194–1245), but there is a trend towards using *help* without infinitive *to* like an auxiliary. This is a major grammatical change, since catenatives are considered main verbs and the

following full verb with or without infinitive marker is considered a complement. This makes dropping the infinitive particle/conjunct *to* after *help* an interesting case of grammaticalisation. In contrast to *dare*, which appears to move away from auxiliary status, *help* seems to move towards it (cf. Mair/Leech 2006). As in other cases of grammaticalisation, this syntactic change is accompanied by a semantic change away from the traditional meaning ‘support’ towards the new meaning ‘contribute’, which is partly rendered as *helfen* in German (S4) and partly not (S5). Sometimes the translation into German makes the auxiliary status of *help* explicit by the use of a corresponding auxiliary *können* (S6).

- (S4) There was another important factor that **helps** explain the sharp rally in interest rate futures in October 1987.
 Dies stellte einen weiteren wesentlichen Faktor dar, der die heftige Rally bei Zinsfutures im Oktober 1987 erklären **hilft**. (ac/Murphy: 652)
- (S5) This **helps** to account for the remarkable degrees to which mammals diversified and the speed at which they did so.
 Von daher **läßt** sich teilweise erklären, in welch erstaunlichem Maße und mit welcher Geschwindigkeit die Säugetiere sich diversifizierten. (ac/Crosby 21).
- (S6) Imagining the 155-meter-high Washington Monument completely hidden inside the new lava dome **helps** us comprehend its immense size.
 Wenn man sich das 155 Meter hohe Washington Monument vorstellt, das im Innern des neuen Lavadoms völlig versteckt wäre, **kann** man seine ungeheure Größe begreifen.
 (ac/Decker/Decker: 836)

A comparative corpus-linguistic analysis can contribute to this discussion on grammaticalisation in many ways: a bilingual search for the forms of English *help* (*help/helps/helped*) in the translation corpus shows (in Table 54.2) that they occur very often when they are NOT rendered by any of the German inflected forms of *helfen* (*hilft/hilfst/helfelhelfen/helft/half/halfst/halfen/halft/geholfen*); English *help* corresponds to German *helfen* in little more than a fifth of all cases.

Tab. 54.2: The English forms of *help* translated as *helfen* in the Chemnitz English–German Translation Corpus

Text Type	English <i>help</i>	German <i>helfen</i>
EU Documents	77	8
Academic Texts	62	11
Public Speeches/Art.	81	24
Tourist Brochures	31	10
Total	251	56

The text-type comparison reveals that the equivalence occurs least often in EU documents: in 8 out of 77 cases altogether, i. e. in 85 percent the stereotypical equivalent is not used. The unexpectedly low figure for the correspondence of related terms illustrates the wide variation that occurs in freely translated text passages.

Of course, translators legitimately try to vary their language and may choose other constructions with related adjectives (like *hilfreich* ‘helpful’ in S7) and nouns (*Hilfe* ‘help’ in S8), but the cases where a different, morphologically unrelated German word is chosen like *beitragen* ‘contribute’ (in S9, or *teilweise* ‘partly’ in S5 above) correlate with the new syntactic pattern mentioned above.

- (S7) This should provide a useful framework and, at the same time, **help** point out the direction we'll be going.
Dies soll ebenso einen sinnvollen Rahmen abstecken wie als **hilfreiche** Orientierung dienen, welche Wegrichtung wir einschlagen werden. (John Murphy 1991/92)
- (S8) Perhaps the best way of interpreting Scotland's wild places is with the **help** of the Countryside Ranger Service.
Die beste Art, Schottlands unberührte Natur zu erkunden, ist mit **Hilfe** des Countryside Ranger Service. (Scottish Tourist Board, 1993)
- (S9) Such economic growth can **help** cut off the oxygen of terrorism.
Ein derartiges Wirtschaftswachstum kann dazu **beitragen**, dem Terrorismus den Boden zu entziehen. (Speech John Major 1994)

The reverse analysis shows that German *beitragen* is not (yet) rendered as *help* in many cases; a few cases are restricted to EU documents and political speeches/articles, but overall, the prototypical *contribute* is much more common. This example illustrates that corpus searches can show to what extent syntax and semantics overlap – although the infinitive without *to* can only be searched for easily in a POS tagged (and carefully post-edited!) corpus.

There are however many cases where semantic elements are combined in unconventional ‘translations’; in S10 *helped to assure that ... would not* is rendered as the antonym *verhinderten* (i. e. ‘prevented’) and the German translation adds a much more explicit phrase (“and thus to compensate for the wood cutting of thousands of trees”) to the English original. This is not a language-specific, but a translation-specific structure.

- (S10) They **helped** to assure that seedlings would not grow into trees to replace the thousands cut down in answer to European needs in the islands and elsewhere.
Sie **verhinderten**, daß nachsprießende Sämlinge sich zu Bäumen auswachsen und damit den Holzeinschlag ausgleichen konnten, dem Tausende von Bäumen zum Opfer gefallen waren, um den Holzbedarf der Europäer auf den Inseln und anderswo zu befriedigen. (ac/Crosby: 947)

Other corpus-linguistic studies have exemplified similar surprising cases, where seemingly equivalent constructions turn out to be incongruent. Viberg (1996), for instance, pointed out the low intertranslatability for English *go* and Swedish *gå*: they overlap only for the most basic meaning ‘motion of a human being’, not in the many ‘metaphorical’ meanings.

For many syntactic comparisons, a tagged corpus is necessary, especially when it comes to determining whether German, for instance, is really “giving us a ‘tighter fit’ between surface form and semantic representation” (Hawkins 1986, 122). The heated, more theoretical (generative) than empirical debate on this issue should be complemented by quantitative analyses of raising constructions, WH-extraction, pied piping and NP deletions in a stratified English–German translation corpus. However this goes well beyond the simple inductive ‘discovery’ perspective adopted in this section.

3.3. Contrastive discourse and style analysis: Presenting and structuring information

Research questions studied in monolingual corpora can become even more focussed through contrastive analysis of multilingual corpora. The presentation of information in discourse often displays subtle preferences even in closely related languages like Germanic, Romance or Slavonic language families (Hasselgård et al. 2002). Coherence is a feature of texts in all languages, but to what extent formal devices of cohesion are used to signal a universal phenomenon deserves close comparative study.

The first example is clause connectors, a central feature of texts that helps in information processing (but does not have to be stated explicitly, i. e. can have no cohesion feature on the surface). Although they serve one large (con)textual purpose, they range from adverbs to conjunctions and relative markers in terms of word class. Since they can be extracted through relatively simple queries, they have enjoyed great research popularity for a long time. Thus it can be demonstrated that German *also* is not translated by English *also*, but by *thus* or *so*, which, even taken together, only account for a small part of German *so* (see Table 54.3).

Tab. 54.3: The English forms of *so* translated as German *so* in the Chemnitz English–German Translation Corpus.

Text Type	English	German
EU Documents	135	36
Academic Texts	767	336
Public Speeches/Art.	184	49
Tourist Brochures	68	29
Total	1154	450

Students who may be misled by such partial ‘false friends’ are easily convinced by the weight of the corpus evidence. These simple examples also show that the traditional word class adverb has to be analysed in functional terms, since German *also* and *so* function more often as causal conjuncts than their English equivalents, which even includes zero. The translations are rarely as prototypical in English, German, French and Spanish as in S11:

- (S11) **English:** I regret this, but the vote has already been taken and the decision is made, **so** let us leave the matter there.
German: Ich bedauere das, aber die Abstimmung ist durchgeführt worden, die Entscheidung ist gefallen, **also** lassen wir die Dinge.
French: Je le déplore mais le vote a été effectué, la décision est tombée et nous devrions **donc** en rester là.
Spanish: Lo lamento, pero la votación se ha realizado, se ha adoptado la decisión y, por consiguiente, dejemos **así** las cosas. (ep-00-01-17.txt: 100)

The second example is again surface-related and a typical feature of many Indo-European languages, especially when word order cannot be used to focus on particular parts

of the information provided by the writer, i.e. cleft constructions for emphasis and focussing (e.g. *it*-clefts and WH-clefts = pseudoclefts). Despite similar structural options, the amount, functions and type of cleft constructions have been studied intensively in English, German and Norwegian (Johansson 2001; Ahlemeyer/Kohlhof 1999). The latter study found that only one third of the English clefts were translated by their direct German equivalent and that the structural differences between the two languages could not account for that, concluding:

“Clefts sentences are marked syntactic structures in English to which a range of specific discourse functions have been attributed in the linguistic literature. It is these discourse functions that must be captured in the German translation as well” (ibid., 12).

This leads to more comprehensive contrastive style analyses, as in the construction in S12, which illustrates the information packaging in a focussing construction (*it is ... who*) – and a stereotypical case of nominalisation, which German is infamous for.

- (S12) Of course, in the heat of an insurrection against an oppressive system **it is those who** are most outspoken and most courageous – the Rosa Luxemburgs and the Karl Liebknechts – who call hundreds of thousands on to the streets.

In der Hitze eines Aufstands gegen ein unterdrückerisches System **sind es** natürlich **die** Freimütigsten und Mutigsten – die Rosa Luxemburgs und Karl Liebknechts –, die Hunderttausende auf die Straßen rufen.
(ac/harm: 777)

Our third example of information presentation is less surface-related and yet it shows that stylistic choices depend largely on the impression the writer intends to make on the reader in a specific context. For political texts, agency and thus responsibility are important aspects. A corpus-linguistic analysis of British and Spanish newspapers for instance (Marín-Arrese et al. 2002) shows similar de-personalisation strategies depending on the ideological orientation of journalists. In principle, the conservative papers (*ABC, La Vanguardia, The Times*) had more often an omission of agent than the liberal ones (*El País, The Guardian*). Both languages used nominalisations, but Spanish had more passive *se* and English more periphrastic passive. S13 is an example from the EU translation corpus, where French also uses the passive construction, but German does not.

- (S13) **English:** Mrs Lynne, you are quite right and I shall check whether this has actually not **been** done.

German: Frau Lynne, Sie haben völlig recht, und ich werde prüfen, ob all dies wirklich so ist.

French: Madame Lynne, vous avez parfaitement raison et je vais vérifier si tout cela n’ a effectivement pas **été** fait.

Spanish: Señora Lynne, tiene toda la razón del mundo, y verificaré si estas cosas no **se** han hecho de verdad.
(p-00-01-17: 33)

Lastly, a really multilingual approach has been followed by Fabricius-Hansen (1998). She substantiated her claims on information density by comparing translations from German into both Norwegian and English and identified strategies of information splitting in particular in Norwegian and English, like backward information splitting when presuppositions are extracted (ibid., 231).

The discourse examples can also demonstrate the limits of corpus-based contrastive analyses, because any fine-grained functional analysis may include specific interpretations and inferences that are difficult to substantiate by using a surface-based research methodology.

4. The added value of multilingual corpora for contrastive studies

4.1. Multilingual corpus comparisons for the fine-tuning of similarities and contrasts

This section demonstrates the added value of multilingual translation corpora compared to monolingual and even bilingual translation corpora, since the value of multilingual corpora has sometimes been called into question (e.g. Santos 2003).

Apart from the practical applications of translation corpora (for language teaching and learning, lexicography and translation), translation corpora can display differences between two or more languages. This increases our knowledge of typological and language-specific features as well as of cultural differences and language universals. The contrastive perspective however also gives us deeper insights into individual languages, which allows the fine-tuning of qualitative and quantitative similarities and differences and which are difficult to obtain through monolingual introspection. Johansson (2002, 47) states: "if we want to gain insight into language and translation generally, and at the same time highlight the characteristics of each language, it is desirable to extend the comparison beyond language pairs". In typologically similar languages like English and German on the one hand, and French and Spanish on the other, contrasts are often not clear-cut but gradient and a matter of choice between options. Thus German learners of English know that modal auxiliaries in English (e.g. *may*) are often expressed as modal adverbs in German (e.g. *vielleicht*); a multilingual search may make them aware of further alternatives if they look at the French equivalent text, where a subjunctive signals hypotheticality (in S14; examples in this section are taken from the multilingual English-German-French-Spanish EU Parliament corpus available at <http://ell.phil.tu-chemnitz.de/multiSearch>).

(S14) **English:** Mr President, ladies and gentlemen, Mrs Gradin, before coming to the individual requests I would like to situate this report within a wider framework, even if this **may** seem presumptuous to the eyes and in the ears of some Members.

German: Herr Präsident, liebe Kolleginnen und Kollegen, liebe Frau Kommissarin Gradin! Bevor ich zu unseren einzelnen Forderungen komme, möchte ich diesen Bericht in einen größeren Zusammenhang stellen, auch wenn dies in den Augen und Ohren der einen oder anderen **vielleicht** anmaßend erscheint.

French: Monsieur le Président, chers collègues, Mme le commissaire, avant que je n'en vienne à nos exigences, j'aimerais placer ce rapport dans un contexte plus large même si cela **paraît** prétentieux aux yeux de l'un ou l'autre. (ep-96-09-19: 19)

Apart from the modality aspect, students can discuss contrasting conventions of address in this example. Simard (2000, 50) is right when he discusses multilingual text alignment with the following justification: "while trilingual or multilingual text alignments may not

be interesting in themselves, any additional version of a translated text should be viewed as additional information that can and should be used to produce better bilingual alignments, and therefore a better knowledge of bilingual translational equivalences”.

However, as stated above, the equivalence problem is more relevant for the translator than for the contrastive linguist who can see ‘critical cases’ as an inductive starting point for further considerations – and other research methodologies. More possible equivalence problems should be seen as a source of inspiration.

4.2. Multilingual corpus searches as discovery procedures

Yet, multilingual corpora (in the narrow sense as translation corpora with more than two languages) have even more advantages. They juxtapose a multitude of structures with similarities and differences, combining concrete examples with comparative figures. Subjective linguistic intuition (*Spachgefühl*) would hardly be able to account for all that. This is a particular advantage when non-native languages are contrasted. Thus despite the query problems caused by multifunctional auxiliaries (e. g. in English *have* may signal “time/temporality” or “causativity”, both attested in S15) and inflectional morphology (e. g. *lassen/läßt*, etc. in German or *faire/faisait*, etc. in French), a combination of open queries with English: *ha** + German: *la** + French *faï** provides some interesting prototypical results:

- (S15) **English:** I **have** also noted that my dear friend, Mr Dell'Alba, **has had** his grumbles **printed** as an appendix to my opinion.

German: Ich habe auch zur Kenntnis genommen, daß mein spezieller Freund aus dem Parlament, Herr Dell' Alba, seine Kritik als Anlage zu meiner Stellungnahme **hat drucken lassen**.

French: J'ai également noté que mon cher ami député M. Dell'Alba **avait fait imprimer** ses rouspétances en annexe à mes remarques. (ep-00-05-03: 2651)

However, this query also extracts unwanted cases (as in S16), which correspond to the forms *havel/haben/faire* but not the intended meaning ‘causative’. At least, it makes students aware of the time/temporal function of *had* in English and the idiomatic usages of *faire* in French in the example below. However, the ‘problem’ could have been avoided by adding a Spanish verb like *hacer* and *causar* to the search option above; then this case would not have been included in the search results.

- (S16) **English:** Just eight days earlier, the European Parliament **had** adopted a resolution in the topical and urgent debate which runs to about three pages in length.

German: Das Europäische Parlament hatte genau acht Tage vorher eine Entschließung in der Dringlichkeitsdebatte verabschiedet, die ungefähr drei Seiten **lang** ist.

French: Huit jours auparavant, le Parlement européen avait adopté une résolution au cours du débat d'actualité, une résolution qui doit **faire** environ trois pages.

Spanish: Exactamente ocho días antes, el Parlamento Europeo aprobó una resolución en el debate de urgencia que abarca aproximadamente tres páginas. (ep-97-07-14: 39)

In contrast to the bilingual comparison (in section 3.2. above), a multilingual comparison of *have* grammaticalisation confirms more clearly the expanded function of English *help*.

In many sample sentences (S17) and in the small statistical comparison (in Table 54.4), the correspondence of *help* to *contribute/beitragen/contribuer* becomes clear.

- (S17) **English:** We also hope that the communication will **help to** build a better common understanding of how to manage risks and to dispel fears that the precautionary principle might be used in an arbitrary way or as a disguised form of trade protectionism.

German: Wir hoffen außerdem, daß die Mitteilung dazu **beitragen** wird, die Möglichkeiten des Risikomanagements besser zu verstehen und Befürchtungen auszuräumen, das Vorsorgeprinzip könnte willkürlich oder als versteckte Form des Handelsprotektionismus eingesetzt werden.

French: Nous espérons aussi que cette communication **contribuera** à instaurer une meilleure compréhension commune de la manière de gérer les risques et à dissiper les craintes selon lesquelles le principe de précaution pourrait être utilisé de façon arbitraire ou pourrait constituer une forme déguisée de protectionnisme commercial.

Spanish: Esperamos también que la comunicación **favorezca** un mejor entendimiento común de cómo deben gestionarse los elementos de riesgo y disipe los temores de que el principio de precaución pueda ser utilizado de manera arbitraria o como una forma encubierta de protecciónismo comercial. (ep-00-02-02.txt: 267)

Tab. 54.4: Equivalents of English *help* in German, French and Spanish in the EU Translation Corpus

<i>help to</i>	30	<i>translated</i>	as
<i>beitragen zu</i>	23	<i>helfen</i>	4
<i>contribuer à</i>	26	<i>aider</i>	4
<i>contribuir a</i>	17	<i>ayudar</i>	9

Finally, even gaps or fillers in a structure come to light in multilingual juxtapositions. S18 shows the verbal structures of *I am sure* in English, French and Spanish in contrast to the adverbial in German, but it also brings out the discourse function of the French pronoun *en* (in contrast to its partitive functions), since it has no equivalent in the other three languages.

- (S18) **English:** But he will appreciate – with his customary generosity **I am sure** – that for us to be able to listen to the response to what we are proposing, it is first necessary to transmit what we are proposing.

German: Jedoch wird er **sicherlich** mit gewohnter Großzügigkeit anerkennen, daß die Voraussetzung dafür, daß wir überhaupt eine Reaktion auf unsere Vorschläge entgegennehmen können, wir unsere Vorschläge erst einmal übermitteln müssen.

French: Mais il admottra avec sa générosité habituelle, **j'en suis sûr** que pour que nous soyons à même d'écouter vos réactions à nos propositions, il faut tout d'abord que nous vous les transmettions.

Spanish: Pero **seguro** que sabrá apreciar – con su acostumbrada generosidad – que para que nosotros podamos escuchar la respuesta a lo que estamos proponiendo es necesario que antes transmitamos lo que estamos proponiendo. (ep-00-01-18.txt: 1436)

A similar contrast between the “Romance” languages French, Spanish and English (?) and the Germanic German occurs when clause connections through gerunds are compared: no problem for the former, but a rather complex construction in the latter (S19). A comprehensive analysis, however, requires a tagged corpus.

- (S19) **English:** Mr President, I will try to be brief **in answering** the Member's question. ...
German: Herr Präsident, ich werde mich um eine kurze Antwort bemühen **und** möchte der Frau Abgeordneten **sagen**, dass ...
French: Je vais m'efforcer, Monsieur le Président, d'apporter une réponse brève **en disant** à Mme la député que ...
Spanish: Voy a esforzarme, señor Presidente, por dar una respuesta breve **diciendo** a su Señoría que ...
(ep-00-02-15.txt: 1129)

These examples, of course, provide only illustrative evidence of the opportunities that multilingual corpora often provide: more striking examples, clearer patterns and thus better understanding of parallel and non-parallel structures in contrastive linguistics.

5. Links to other fields

The field of contrastive corpus studies sketched out so far intertwines with a number of other corpus-oriented areas of research and beyond.

Translation or parallel corpora (article 16) are the basis for contrastive analyses as well as human and machine translation (article 55 and article 56, respectively). Whereas humans study translation through parallel corpora, translation programmes compare language structures. Incongruencies between translation corpora may be attributed to three reasons (which cannot always be kept apart discretely):

- differences persist because the translator produced a different text consciously or unconsciously (translationese due to universal features of the translation process, which may lead to more explicit or more or less comprehensible texts, e. g. Olohan/Baker 2000);
- differences persist because structures are transferred from the source language to the target language (translationese due to interferences, which may lead to unnatural structures or preferences in the target language); and
- differences persist because the target language prefers other structures naturally.

The first reason is particularly relevant for translation studies, the second for language learning and teaching, and the third for contrastive linguists.

Contrastive corpus-linguists need the input of corpus compilers and tool developers, just like corpus analysts supply an input for language service providers in the widest sense, from translators to teachers. This is because the results of corpus-linguistic studies cannot only be used in language descriptions, but also in applied linguistic areas like contrastive grammar (cf. Schmied 1999), lexicography (article 8) and language (and translation) teaching and learning (article 7 and Barlow 2000). Nowadays, multilingual lexicology (Vintar/Hansen 2005) and lexicography (Teubert 2001, 2002) seem to be particularly underdeveloped. Future bilingual dictionaries (cf. Schmied/Fink 2000 for examples of *with = mit*), for instance, should include:

- quantitative and style-specific equivalents (and differences), providing for bilingual dictionaries what Biber et al.'s *Longman Grammar of Spoken and Written English* (1999) does for grammar;

- a contrastive annotation system in each subentry that marks clear semantic and syntactic equivalents briefly and unambiguously and concentrates on the semi- or non-equivalents;
- authentic corpus examples of prototypical and non-prototypical equivalents; and
- a CD-ROM with a corpus and a query system to provide more examples of prototypical and non-prototypical equivalents and simple comparative statistics.

So far, this exploitation of contrastive studies for applied purposes has hardly begun.

6. Conclusion

Multilingual corpora, and particularly translation corpora, have managed to re-invigorate contrastive linguistics in all linguistic sub-disciplines. They can also be used to find out more about language phenomena than would not be possible with monolingual and even bilingual corpora alone. Often interesting features of a particular language, can be noticed when they are seen in contrast. What individual researchers may not see through introspection they may also overlook by analysing only monolingual or even bilingual corpora (especially if the two languages are structurally similar).

At the moment, this is only a promising beginning and more corpora need to be compiled, more (syntactic, semantic and pragmatic) tags need to be inserted into corpus texts and many more detailed analyses need to be carried out before we can come to a more comprehensive understanding of the qualitative and quantitative contrasts between languages.

7. Literature

7.1. References

- Ahlemeyer, B./Kohlhof, I. (1999), Bridging the Cleft: An Analysis of the Translation of English *it-clefts* into German. In: *Languages in Contrast* 2(1), 1–25.
- Aijmer, K. (1998), Epistemic Predicates in Contrast. In: Johansson/Oksefjell 1998, 277–295.
- Aijmer, K. (1999), Epistemic Possibility in an English-Swedish Contrastive Perspective. In: Hasselgård/Oksefjell, 1999, 301–323.
- Aijmer, K./Altenberg, B. (eds.) (2004), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, Göteborg 22–26 May 2002. Amsterdam: Rodopi.
- Aijmer, K./Altenberg, B./Johansson, M. (eds.) (1996), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.
- Altenberg, B. (1999), Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences. In: Hasselgård/Oksefjell, 1999, 249–268.
- Altenberg, B./Aijmer, K. (2000), The English-Swedish Parallel Corpus: A Resource for Contrastive Research and Translation Studies. In: Mair, C./Hundt, M. (eds.), *Corpus Linguistics and Linguistic Theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20)*, Freiburg im Breisgau 1999. Amsterdam/Philadelphia: Rodopi, 15–33.

- Altenberg, B./Granger, S. (eds.) (2002), *Lexis in Contrast: Corpus-based Approaches*. Amsterdam: John Benjamins.
- Baker, M. (1993), Corpus Linguistics and Translation Studies: Implications and Applications. In: Baker, M./Francis, G./Tognini Bonelli, E. (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233–250.
- Baker, M. (1995), Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. In: *Target* 7(2), 223–245.
- Baker, M. (1996), Corpus-based Translation Studies. The Challenges that Lie Ahead. In: Somers, H. (ed.), *Terminology, LSP and Translation*. Amsterdam: John Benjamins, 175–186.
- Barlow, M. (1999), MonoConc 1.5 and ParaConc. In: *International Journal of Corpus Linguistics* (4)1, 319–327.
- Barlow, M. (2000), Parallel Texts in Language Teaching. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 106–115.
- Baroni, M./Bernardini, S. (2006), A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. In: *Literary and Linguistic Computing* 21, 259–274.
- Bernardini, S. (2000), *Competence Capacity Corpora. A Study in Corpus-aided Language Learning*. Via Marsala: Cooperativa Libraria Universitaria Editrice Bologna.
- Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow, UK: Pearson Education.
- Borin, L. (ed.) (2002), *Parallel Corpora, Parallel Worlds*. (Language and Computers 43.) Amsterdam: Rodopi.
- Ebeling, J. (1998), The Translation Corpus Explorer: A Browser for Parallel Texts. In: Johansson/Oksefjell 1998, 101–112.
- Fabricius-Hansen, C. (1998), Informational Density and Translation, with Special Reference to German – Norwegian – English. In: Johansson/Oksefjell 1998, 197–234.
- Frankenberg-Garcia, A. (2004), Lost in Parallel Concordances. In: Aston, G./Bernardini, S./Stewart, D. (eds.), *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins, 213–229.
- Frankenberg-Garcia, A. (2005), Pedagogical Uses of Monolingual and Parallel Concordances. In: *English Language Teaching Journal* 59(3), 189–198.
- Gilquin, G. (2001), The Integrated Contrastive Model. Spicing up your Data. In: *Languages in Contrast* 3, 95–123.
- Granger, S. (1996), From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora. In: Aijmer/Altenberg/Johansson 1996, 37–51.
- Hansen-Schirra, S. (2003), Linguistic Enrichment and Exploitation of the Translational English Corpus. In: *Proceedings of Corpus Linguistics 2003*. Lancaster, UK, 288–297.
- Hasselgård, H. (1998), Thematic Structure in Translation between English and Norwegian. In: Johansson, S./Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*. Amsterdam: Rodopi, 145–168.
- Hasselgård, H./Oksefjell, S. (eds.) (1999), *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi.
- Hasselgård, H./Johansson, S./Behrens, B./Fabricius-Hansen, C. (eds.) (2002), *Information Structure in a Cross-linguistic Perspective*. Amsterdam: Rodopi.
- Hawkins, J. (1986), *A Comparative Typology of English and German: Unifying the Contrasts*. London/Sydney: Croom Helm.
- Hoey, M. (2005), *Lexical Priming. A New Theory of Words and Language*. London: Routledge.
- Hofland, K. (1996), A Program for Aligning English and Norwegian Sentences. In: Hockey, S./Ide, N./Perissinotto, G. (eds.), *Creating and Using English Language Corpora*. Amsterdam: Rodopi, 25–37.
- Huddleston, R./Pullum, G. K. (2002), *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

- Ide, N./Véronis, J. (1994), MULTEXT: Multilingual Text Tools and Corpora. In: *Proceedings of COLING 1994*. Kyoto, Japan, 588–592.
- Iglesias-Rábade, L./Suárez, S. M. D. (eds.) (2002), *Studies in Contrastive Linguistics. Proceedings of the Second International Contrastive Linguistics Conference (ICLC 2)*. Universidade de Santiago de Compostela: USC Publishing Services.
- Johansson, M. (1996), Contrastive Data as a Resource in the Study of English Clefts. In: Aijmer/Altenberg/Johansson 1996, 127–150.
- Johansson, S. (2001), The German and Norwegian Correspondences to the English Construction Type *that's what*. In: *Linguistics* 39(3), 583–605.
- Johansson, S. (2002), Towards a Multilingual Corpus for Contrastive Analysis and Translation Studies. In: Borin 2002, 47–59.
- Johansson, S. (2003), Reflections on Corpora and their Uses in Cross-linguistic Research. In: Zanettin, F./Bernardini, S./Stewart, D. (eds.), *Corpora in Translator Education*. Manchester: St Jerome Publishing, 133–144.
- Johansson, S. (2004), Viewing Languages through Multilingual Corpora, with Special Reference to the Generic Person in English, German, and Norwegian. In: *Languages in Contrast* 4(2), 261–280.
- Johansson, S. (2007), *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.
- Johansson, S./Ebeling, J./Hofland, K. (1996), Coding and Aligning the English-Norwegian Parallel Corpus. In: Aijmer/Altenberg/Johansson 1996, 87–112.
- Johansson, S./Hofland, K. (1994), Towards an English-Norwegian Parallel Corpus. In: Fries, U./Tottie, G./Schneider, P. (eds.), *Creating and Using Language Corpora*. Amsterdam/Atlanta: Rodopi, 25–37.
- Johansson, S./Hofland, K. (2000), The English–Norwegian Parallel Corpus: Current Work and New Directions. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 134–147.
- Johansson, S./Oksefjell, S. (eds.) (1998), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.
- Kenny, D. (1999), The German-English Parallel Corpus of Literary Texts (GEPCOLT): A Resource for Translation Scholars. In: *Teanga* 18, 25–42.
- Maia, B. (2000), Making Corpora – a Learning Process. In: Bernardini, S./Zanettin, F. (eds.), *I corpora nella didattica della traduzione*. Bologna: CLUEB, 47–46.
- Maia, B./Haller, J./Ulrych, M. (eds.) (2002), *Training the Language Services Provider for the New Millennium*. Porto: Universidade do Porto.
- Mair, C./Leech, G. (2006), Current Changes in English Syntax. In: Aarts, B./McMahon, A. (eds.), *The Handbook of English Linguistics*. Oxford: Blackwell, 318–342.
- Marín-Arrese, J./Martínez-Caro, E./Neff, J./Pérez de Ayala, S./Blanco, M. L./Molina, C. (2002), Mystification of Agency and Primary Responsibility in Newspaper Discourse in English and Spanish: A Comparable Corpus Study. In: Rábade/Suárez 2002, 599–609.
- Mauranen, A. (1997), Hedging in Language Revisers' Hands. In: Markkanen, R./Schröder, H. (eds.), *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Berlin: Walter de Gruyter, 115–133.
- McEnery, A./Oakes, M. P./Garside, R. G. (1994), The Use of Approximate String Matching Techniques in the Alignment of Sentences in Parallel Corpora. In: Vella, A. (ed.), *The Proceedings of MT – 10 Years on*. Cranfield, UK, 55–67.
- Neumann, S./Hansen-Schirra, S. (eds.) (2003), *Proceedings of the Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives (Lancaster, 27 March 2003)*. Lancaster: UCREL.
- Olohan, M./Baker, M. (2000), Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation? In: *Across Languages and Cultures* 1, 141–172.
- Olohan, M. (2004), *Introducing Corpora in Translation Studies*. London: Routledge.

- Rabadán, R. (2006), Translating the ‘Predictive’ and ‘Hypothetical’ Meanings English-Spanish. In: *Meta* 52(3), 484–502.
- Rábade, L. I./Doval Suárez, S. M. (eds.) *Studies in Contrastive Linguistics*. Santiago de Compostela: Universidad di Santiago de Compostela.
- Ramón García, N. (2003), *Estudio contrastivo inglés-español de la caracterización de sustantivos*. León: Universidad de León.
- Sajavaara, K. (1996), New Challenges for Contrastive Linguistics. In: Aijmer/Altenberg/Johansson 1996, 17–36.
- Salkie, R. (2003), Using Parallel Corpora in Translation. In: The Guide to Good Practice for Learning and Teaching in Languages, Linguistics and Area Studies. LTSN Subject Centre for Languages, Linguistics and Area Studies. Available at: <http://www.lang.ltsn.ac.uk/resources/good-practice.aspx?resourceid=1444>.
- Santos, D. (1997), The Importance of Vagueness in Translation: Examples from English to Portuguese. In: *Romansk Forum* 5, 43–69.
- Santos, D. (2003), Against Multilinguality. In: Neumann/Hansen-Schirra 2003, 7–16.
- Schmied, J. (1994), Translation and Cognitive Structures. In: *Hermes, Journal of Linguistics* 13, 169–181.
- Schmied, J. (1998), To Choose or Not to Choose the Prototypical Equivalent. In: Schulze, R. (ed.), *Making Meaningful Choices in English. On Dimensions, Perspectives, Methodology, and Evidence*. Tübingen: Gunter Narr, 207–222.
- Schmied, J. (1999), Applying Contrastive Corpora in Modern Contrastive Grammars: The Chemnitz Internet Grammar of English. In: Hasselgård/Oksfjell 1999, 21–30.
- Schmied, J. (2002), Prototypes, Transfer and Idiomaticity: An Empirical Analysis of Local Prepositions in English and German. In: Rábade/Doval Suárez 2002, 947–959.
- Schmied, J. (2004), Translation Corpora in Contrastive Research, Translation and Language Teaching. In: *TradTerm* 10. São Paulo: Humanitas FFLCH/USP, 83–115.
- Schmied, J./Fink, B. (2000), Corpus-based Contrastive Lexicology: The Case of English *with* and its German Translation Equivalents. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 157–176.
- Schmied, J./Schäffler, H. (1996), Approaching Translationese through Parallel and Translation Corpora. In: Percy, C./Lancashire, I./Meyer, C. (eds.), *Synchronic Corpus Linguistics*. Amsterdam: Rodopi, 41–56.
- Schmied, J./Schäffler, H. (1997), Explicitness as a Universal Feature of Translation. In: Ljung, M. (ed.), *New Ways in Corpus Linguistics*. Amsterdam: Rodopi, 21–34.
- Simard, M. (2000), Multilingual Text Alignment: Aligning Three or More Versions of a Text. In: Véronis, J. (ed.), *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers, 49–67.
- Steiner, E. (2005), *Explication, its Lexicogrammatical Realization, and its Determining (Independent) Variables – towards an Empirical and Corpus-based Methodology*. SPRIKreports 36. Available at: http://www.hf.uio.no/forskningsprosjekter/sprik/docs/pdf/Report_36_ESteiner.pdf.
- Tamburini F. (2002), A Dynamic Model for Reference Corpora Structure Definition. In: *Proc. Third International Conference on Language Resources and Evaluation – LREC2002*. Las Palmas, Canary Islands, Spain, 1847–1850.
- Teich, E. (2003), *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Walter de Gruyter.
- Teich, E./Hansen, S./Fankhauser, P. (2001), Representing and Querying Multi-layer Annotated Corpora. In: *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia, PA, 228–237.
- Teubert, W. (1996), Comparable or Parallel Corpora? In: *International Journal of Lexicography* 9(3), 38–64.
- Teubert, W. (2001), Corpus Linguistics and Lexicography. In: *International Journal of Corpus Linguistics* 6, 125–153.
- Teubert, W. (2002), The Role of Parallel Corpora in Translation and Multilingual Lexicography. In: Altenberg/Granger 2002, 189–214.

- Véronis, J. (ed.) (2000), *Parallel Text Processing. Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.
- Viberg, Å. (1996), Cross-linguistic Lexicology. The Case of English *go* and Swedish *gå*. In: Aijmer/Altenberg/Johansson 1996, 153–182.
- Vinay, J.-P./Darbelnet, J. (1995), *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam: John Benjamins.
- Vintar, S./Hansen, S. (2005), Cognates – Free Rides, False Friends or Stylistic Devices: A Corpus-based Comparative Study. In: Barnbrook, G./Danielsson, P./Mahlberg, M. (eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: Birmingham University Press.
- Willems, D./Defrancq, B./Colleman, T./Noël, D. (2003), *Contrastive Analysis in Language – Identifying Linguistic Units of Comparison*. London: Palgrave Macmillan.
- Zanettin, F. (2002), DIY Corpora: The WWW and the Translator. In: Maia/Haller/Ulrych 2002, 239–248.

7.2. WWW resources

URLs accessed on May 9, 2007.

The EAGLES corpus guidelines: <http://www.ilc.cnr.it/EAGLES/browse.html>

The Oslo ‘Languages in Contact’ page: <http://www.hf.uio.no/forskningsprosjekter/sprik/english/index.html>

The Chemnitz English/German translation corpus: <http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/project.htm>

The Chemnitz multilingual search tools: <http://ell.phil.tu-chemnitz.de/>

The English-Norwegian parallel corpus: <http://www.hf.uio.no/iba/prosjekt/>

The Linguateca distributes language resource center for Portuguese: <http://www.linguateca.pt>

The English-Swedish parallel corpus: <http://www.englund.lu.se/content/view/66/127/>

TransSearch: <http://www.tsrali.com/?UTLanguage=en>

Josef Schmied, Chemnitz (Germany)

55. Corpora in human translation

1. Introduction
2. Corpus design and availability of corpora
3. Corpus processing
4. Applications in translation and translatology
5. Links to other fields
6. Literature

1. Introduction

Human translation is one of the fields in which the use of corpora has had a growing impact in the last couple of decades (for the use of corpora in machine translation see article 56). This is reflected in a wealth of publications, as well as thematically dedicated

- Véronis, J. (ed.) (2000), *Parallel Text Processing. Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.
- Viberg, Å. (1996), Cross-linguistic Lexicology. The Case of English *go* and Swedish *gå*. In: Aijmer/Altenberg/Johansson 1996, 153–182.
- Vinay, J.-P./Darbelnet, J. (1995), *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam: John Benjamins.
- Vintar, S./Hansen, S. (2005), Cognates – Free Rides, False Friends or Stylistic Devices: A Corpus-based Comparative Study. In: Barnbrook, G./Danielsson, P./Mahlberg, M. (eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: Birmingham University Press.
- Willems, D./Defrancq, B./Colleman, T./Noël, D. (2003), *Contrastive Analysis in Language – Identifying Linguistic Units of Comparison*. London: Palgrave Macmillan.
- Zanettin, F. (2002), DIY Corpora: The WWW and the Translator. In: Maia/Haller/Ulrych 2002, 239–248.

7.2. WWW resources

URLs accessed on May 9, 2007.

The EAGLES corpus guidelines: <http://www.ilc.cnr.it/EAGLES/browse.html>

The Oslo ‘Languages in Contact’ page: <http://www.hf.uio.no/forskningsprosjekter/sprik/english/index.html>

The Chemnitz English/German translation corpus: <http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/project.htm>

The Chemnitz multilingual search tools: <http://ell.phil.tu-chemnitz.de/>

The English-Norwegian parallel corpus: <http://www.hf.uio.no/iba/prosjekt/>

The Linguateca distributes language resource center for Portuguese: <http://www.linguateca.pt>

The English-Swedish parallel corpus: <http://www.englund.lu.se/content/view/66/127/>

TransSearch: <http://www.tsrali.com/?UTLanguage=en>

Josef Schmied, Chemnitz (Germany)

55. Corpora in human translation

1. Introduction
2. Corpus design and availability of corpora
3. Corpus processing
4. Applications in translation and translatology
5. Links to other fields
6. Literature

1. Introduction

Human translation is one of the fields in which the use of corpora has had a growing impact in the last couple of decades (for the use of corpora in machine translation see article 56). This is reflected in a wealth of publications, as well as thematically dedicated

workshops such as the Conference on Corpus Use and Learning to Translate (Bertinoro 2000), Research Models in Translation Studies (Manchester 2000), Symposium on Corpora and Translation Studies (Surrey 2002), and Corpus-based Translation Studies: Research and Applications (Pretoria 2003). Corpora have also had a strong presence at the major conferences in both translation studies (e.g. the Conference of the European Society of Translation Studies (EST)) and corpus linguistics (e.g. the Corpus Linguistics conference, Meeting of the International Computer Archive of Modern/Medieval English (ICAME), and the Conference on Language Resources and Evaluation (LREC)). Corpus-based translation studies is thus beginning to form a major paradigm both within the discipline of translation studies and across its boundaries.

As for many other fields, so also for translation studies, the primary value of employing corpora is the opportunity to investigate large amounts of data and conduct empirical research on translations. With the move from observation of text samples to the investigation of larger sets of texts, new methodological and technical challenges emerge for the discipline.

On the technical plane, translation studies benefits from existing corpus-linguistic techniques, such as keyword-in-context (KWIC) concordances, automatic frequency counts of words and the like (cf. articles 33 and 36). While the use of such tools has become common practice in corpus-based translation studies, more sophisticated corpus techniques, notably tools for corpus annotation, corpus maintenance and corpus query as they have been developed for monolingual corpora, have only recently started to be employed (cf. Neumann/Hansen-Schirra 2003 and section 3).

The principal challenges corpus-based translation studies faces, however, are methodological ones. The primary issue is to decide which of its research questions lend themselves to corpus-linguistic, empirical procedures and which do not. Translation studies is a heterogeneous discipline, which imports methods from various fields and adapts them to its own purposes. A theory of translation may thus focus on linguistic issues and hence push towards contrastive linguistics (e.g. Snell-Hornby 1988) or psycholinguistics (e.g. Lörscher 1991); or it may focus on particular text types, such as literary translation, and thus push towards comparative stylistics (Vinay/Darbelnet 1995) or hermeneutics (Toury 1995); or it may focus on rather abstract issues, such as 'equivalence' (Nida 1964), and thus push more towards language philosophy. Whatever the focus, corpus-based work potentially offers the opportunity to make analysis procedures transparent and analysis results replicable. Here, the stronger the interest in linguistic micro-analysis as a way of answering the research questions posed, the more corpus-based procedures of analysis lend themselves to application in translation research. Thus, old questions can be addressed in novel ways and new questions can be asked. One of the old questions concerns the units of translation. Text-translation alignment, the technical basis of parallel concordance programs (e.g. Barlow 1999), suggests that this issue is resolved, when in fact it is still on the research agenda, both in computational linguistics (cf. article 56) and in linguistics proper (see e.g. Fabricius-Hansen 1998, 1999) as well as from the point of view of translational equivalence. Another issue is the detection of universal laws of translation (e.g. Toury's laws of interference and growing standardization; Toury 1995). Such generalizing observations have been recast in corpus-based translation studies as 'universals of translation' (cf. Mauranen/Kujamäki 2004), so that we finally have the opportunity to test their validity. In this area of corpus-based translation studies, monolingually comparable corpora are widely employed (cf. section 2).

But there are also some new questions which can only be asked now that the technology for analysing large sets of translation data is available. These include language change through translation (e.g. Baumgarten/House/Probst 2004; cf. also article 52), translation as functional or registerial variation (Neumann 2003, Teich 2003), or issues of cognitive processing (Hansen 2003; see also section 4).

Corpus-based work can thus shed light on traditional research topics and inspire new aspects to be investigated, pushing the field further both methodologically and in terms of theory. The remainder of this article is organised as follows. Section 2 presents the principal corpus designs researchers use in corpus-based translation studies. Section 3 describes the kinds of corpus processing techniques typically employed as well as some additional techniques that are potentially useful. Section 4 presents a selection of applications of corpus-based work on translation. Finally, section 5 concludes with some remarks on the relation of corpus-based translation studies to other fields.

2. Corpus design and availability of corpora

Corpus-based research on translations requires particular types of corpus design. The present section describes the corpus designs most commonly employed and comments on the availability of suitable corpus resources.

2.1. Types of corpus design

Two types of corpus design are most commonly used in corpus-based translation studies: the *parallel corpus* and the monolingually *comparable corpus* (cf. articles 16, 54, 32).

Parallel corpora consist of source language (SL) texts and translations of those texts into a target language (TL) (see Figure 55.1). They are commonly employed in bilingual lexicography (cf. article 8) and more recently also in machine translation (cf. article 56). In translation research, parallel corpora are used to provide information on language-pair specific translation behaviour, to observe equivalence relations between lexical items or grammatical structures in the source and target languages/texts, or, in a learners' corpus (cf. section 4), to investigate translation problems and translation mistakes.

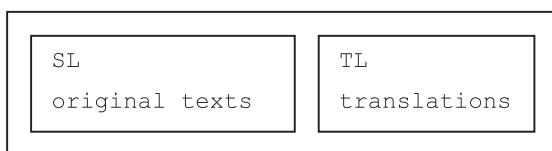


Fig. 55.1: Parallel corpus

Recent parallel corpus initiatives are moving towards more than one language pair, e.g. the Oslo Multilingual Corpus (OMC) of the SPRIK (Språk i kontrast) project (Johansson 2002). Also, there are a few treebanking efforts, i.e., annotation of parallel corpora in terms of syntactic structure (e.g. the Sofie Treebank (Samuelsson/Volk 2005) or the CroCo Corpus (Hansen-Schirra/Neumann/Vela 2006)). Parallel treebanks are used for

research on translations, for translator training as well as for machine translation and multilingual grammar induction.

Monolingually comparable corpora (short: comparable corpora) are a more recent idea (Baker 1995). They are collections of translations (from one or more source languages) into one target language and original texts in the target language (cf. Figure 55.2). Comparable corpora “should cover a similar domain, variety of language and time span, and be of comparable length” (Baker 1995, 23).

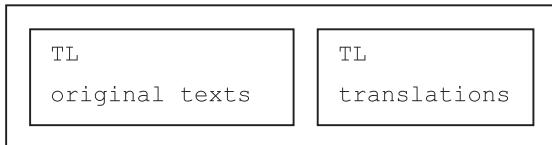


Fig. 55.2: Comparable corpus

Comparable corpora have “the potential to reveal most about features specific to translated text, i. e., those features that occur exclusively, or with unusually low or high frequency, in translated text as opposed to other types of text production, and that cannot be traced back to the influence of any one particular source text or language” (Kenny 1997).

Among the features researchers have posited comparing translations and texts originally produced in the target language are explicitation, simplification, normalisation/conservatism (Baker 1995, 1996; Laviosa-Braithwaite 1996; Kenny 1998; Mauranen 1997; Teich 2003; Hansen 2003). The main application of comparable corpora is the investigation of the specific (and possibly universal) properties of translations. The fact that translations exhibit linguistic properties that distinguish them from texts that are *not* translations is sometimes also referred to as *translationese* (see e. g. Baroni/Bernardini 2006). Comparable corpora can also be useful in translation training and translation practice. For details on applications of this type of corpus, see section 4.

The most recent type of corpus design is a combination of parallel and comparable corpora. Such a combination will automatically contain a third subcorpus: a *multilingually comparable corpus* (cf. Figure 55.3).

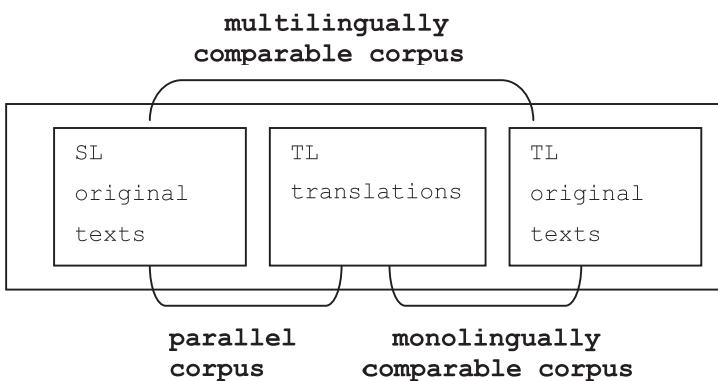


Fig. 55.3: Combined parallel-comparable corpus

This combination of corpora can be used for cross-linguistic comparison of original texts, cross-linguistic comparison of original and translated texts, cross-linguistic comparison of translated texts and monolingual comparison of original and translated texts. For example, an analysis of a multilingually comparable corpus allows establishing a *tertium comparationis* for an analysis of parallel texts (cf. Teich 2003). Reference to multilingually comparable texts can also be useful in translation teaching and practice (cf. section 4).

Methodologically, the primary issue concerning comparable corpora, both monolingual and multilingual, is to define the notion of *comparability*. The favoured solution is to use the concept of register, i. e., functional variation or variation according to situational context. TEC (Translational English Corpus; see section 2.2.), for example, has followed the design of the British National Corpus (BNC) in terms of registerial distinctions. Other projects have used Halliday's (Halliday/Matthiessen 2004) or Biber's (Biber 1995) register features as a basis for defining comparability (Teich 2003, Hansen 2003; cf. also article 38).

2.2. Availability of corpus resources

Of the three types of corpora described in the previous section, parallel corpora are the most wide-spread. Among the best known parallel corpus resources are the French-English Canadian Hansard, parts of which are available in aligned form (Germann 2001), the English-Norwegian Parallel Corpus (ENPC; Johansson/Ebeling/Hofland 1996), or the English-Swedish Parallel Corpus (ESPC; Aijmer/Altenberg/Johansson 1996).

Monolingually comparable corpora are less frequent than parallel corpora. Among the best known corpora of this kind are the Corpus of Translated Finnish (CTF; Mauranen/Kujamäki 2004) and the Translational English Corpus (TEC; Baker 1995) with (part of) the BNC providing the comparable texts.

While there is quite an amount of electronic text material available on the web that potentially qualifies for translation research (cf. Web-as-corpus; see article 18), the availability of corpora is actually limited.

Apart from copyright restrictions, other factors impeding the ready use of electronically available resources are coverage of languages and coverage of registers. In terms of language coverage, there is a clear focus on English. Also, parallel corpora more often than not consist of texts from only one language pair, with the exception of materials from the European Union (e. g. the Europarl Corpus (Koehn 2005); the MLCC – Multilingual Corpora for Co-operation, distributed via ELRA; or TRACTOR, distributed via TELRI). Multilingually comparable material is generally available, too, but for a combined parallel-monolingually comparable corpus, comparability again becomes an issue. For instance, we may be able to obtain a multilingually comparable corpus of newspaper texts (such as the corpus distributed by the European Corpus Initiative (ECI)), but the problem then is to find an equivalent parallel corpus.

In terms of register coverage, most available material is written-to-be-read. Also, the register distinctions drawn by some corpora may not be fine-grained enough (e. g. TEC, following the register grid of the BNC, only distinguishes between fiction, biography and news), e. g. if one is interested in translations in a particular field of discourse. This causes corpora to be ill-suited for many research questions in translation studies which

require well-balanced and registerially sufficiently differentiated data to arrive at empirically valid corpus analysis results. Furthermore, as to spoken data, be it written-to-be-spoken or spontaneously spoken, corpora of (simultaneously or consecutively) interpreted text are, at this stage, seriously underrepresented. In fact, empirical research on interpretation is generally still rare (e.g. Lambert/Moser-Mercer 1994; Shlesinger 1998; Meyer 1998; Pöchhacker 2000; Meyer et al. 2003).

With these factors limiting the ready intake of available electronic material, researchers in translation studies in fact often have to compile corpora that suit their particular needs from scratch. This then requires dealing with all the issues involved in such an endeavour, from sampling and obtaining data, through possibly digitising text and taking care of copyright issues, to corpus processing in more than one language.

3. Corpus processing

In translation studies, most corpus-based research is carried out on the basis of raw text. Only the encoding of meta-information is a usual practice (cf. section 3.3.). Recently, however, the use of annotated corpora is becoming more common because this is the only way of empirically investigating grammatical and semantic, as well as discourse or register features (cf. Hansen/Teich 2001, Hansen-Schirra 2003).

Investigating a corpus across languages requires several processing steps. Possible methods and procedures are described in the remainder of this section.

3.1. Alignment

For the analysis of a parallel corpus (see also article 16), the units of translation (i.e. source language text units and their translational equivalents) need to be aligned. There are various alignment programs freely available (e.g. Hofland 1996); additionally, aligners are often incorporated in translation memories (databases with source language segments and their aligned target language equivalents; e.g. the Translator's Workbench by Trados, Heyn 1996). The most commonly implemented technique is sentence-by-sentence alignment. The aligned sentences are stored either in one file, where the source language sentence and its translation are represented in a tab separated vector format, or in two separate files, linking the sentences through identical line numbers. Files of aligned texts can be exported to translation workbenches (i.e. translation memories including terminology programs) and to standard applications, such as MS Excel or MS Access.

Other alignment techniques apply to paragraphs (cf. Mihailov 2001) or words (e.g. the word aligner GIZA++, Och/Ney 2003), the latter being a prerequisite for statistical machine translation. The TreeAligner (Volk et al. 2006) can be used for parallel treebanking, i.e. to align bilingual sentence pairs already annotated in terms of syntactic structure.

3.2. Linguistic annotation

For the analysis of grammatical and semantic features, as well as discourse features, texts have to be linguistically annotated with the relevant information first, before any analysis can be carried out. Depending on how abstract the linguistic features to be

analysed are, linguistic corpus annotation can be done automatically, semi-automatically or manually.

There are a number of automatic annotation techniques. The most reliable ones are part-of-speech taggers (cf. articles 23 and 24) and morphological analysers (article 25), following either a rule-based or a statistical approach. For the former, the rules have to be modified according to the language which is to be analysed; and for the latter a language-specific training is required on the basis of which probabilities concerning the occurrence of different tags can be calculated.

PoS tagging is rather common in corpus-based translation studies, but other types of annotation, such as syntactic or semantic tagging, while potentially useful, are less widely practised. This has a number of reasons. First, for the types of research questions most commonly posed in corpus-based translation studies, it is often sufficient to use PoS-tagged text or even raw text (cf. section 4). Second, syntactic (articles 13 and 28) and even more so semantic annotation (cf. e. g. article 26) always involves manual effort: fully automatic procedures are not available or not sufficiently reliable, so that annotations have to be either checked by humans or carried out manually. Third, another impeding factor in corpus-based analysis of translation is multilinguality. Many of the available tools are not entirely language-independent. When applied to a new language, they have to be adapted to that language. For instance, in PoS tagging, tag sets which are both suitable for the new language and comparable to the tag set of the first language have to be designed and the tagger needs to be trained. More information on the preprocessing of multilingual corpora can be found in article 32.

3.3. Representation and mark-up

In terms of corpus encoding, corpora in translation studies commonly adhere to standards such as the one developed by the Text Encoding Initiative (TEI), employing SGML or XML for representation (cf. article 22). This guarantees the exchangeability and searchability of the corpus and its mark-up. The text body is annotated for headings, sentences, paragraphs etc. Each text is encoded in terms of a header that provides meta-information on title, author, publication, register information, etc. In case of a translation corpus, it is important to encode information on the translator, the translation process, etc. Figure 55.4 displays a header as it is used in the Translational English Corpus (see section 2.2.).

The TEC header includes information on the title of the book, the translator, the translation output, the translation process, the author of the original and the source language text. This information is important to filter the corpus and divide it into subcorpora and to enable corpus queries according to register or independent variables, such as publication date, nationality of the author or sex of the translator. Thus, translations of female full-time, in-house translators who translate into their mother tongue can, for instance, be compared to translations of male part-time, free-lance translators who translate into a foreign language.

In parallel translation corpora, the alignment of source and target language segments is also represented in the encoding. Figure 55.5 illustrates the corpus encoding of the ELAN Slovene-English Corpus (cf. Erjavec 1999).

```

<Header>
  <title>
    <filename></filename>
    <subcorpus></subcorpus>
    <collection></collection>
    <editor></editor>
  </title>
  <translator>
    <name></name>
    <gender></gender>
    <sexualOrientation></sexualOrientation>
    <Nationality></Nationality>
    <employment></employment>
    <status></status>
  </translator>
  <translation>
    <mode></mode>
    <extent></extent>
    <publisher></publisher>
    <pubPlace></pubPlace>
    <date></date>
    <copyright></copyright>
    <sponsor></sponsor>
    <reviews></reviews>
    <comments></comments>
  </translation>
  <translationProcess>
    <direction></direction>
    <mode></mode>
    <type></type>
  </translationProcess>
  <author>
    <name></name>
    <gender></gender>
    <sexualOrientation></sexualOrientation>
    <Nationality></Nationality>
  </author>
  <sourceText>
    <language></language>
    <mode></mode>
    <status></status>
    <publisher></publisher>
    <pubPlace></pubPlace>
    <date></date>
    <comments></comments>
  </sourceText>
</Header>
```

Fig. 55.4: TEC header

The corpus is divided into translation units (tu), each including the original and the translated segment. Most of the alignment is sentence-based, i.e. each translation unit covers one sentence and its translation. In addition, words (w) and punctuation (c) elements are distinguished.

```
<tu lang="en-sl" id="ligs.301">
<seg lang="en"> <w>Many</w> <w>text</w> ... </seg>
<seg lang="sl"> <w>Za</w> <w>Linux</w> <w>je</w> ... </seg>
</tu>
...
<tu lang="en-sl" id="gnpo.301">
<seg lang="en"> <w>Usage</w> <c>:</c> <w>%s</w> ... </seg>
<seg lang="sl"> <w>Uporaba</w> <c>:</c> <w>%s</w> ... </seg>
</tu>
```

Fig. 55.5: Translation units in ELAN

Apart from meta-information and alignment, linguistic annotation can potentially also be encoded in a mark-up language, such as SGML or XML. Similar to other areas of corpus linguistics, this becomes especially advisable when there is more than one level of annotation (multi-level or multi-layer corpus; cf. Teich/Hansen/Fankhauser 2001). Within this context, the German-English/English-German CroCo Corpus uses XCES (the Corpus Encoding Standard for XML; cf. article 22) to represent multi-layer annotation and alignment (cf. Hansen-Schirra/Neumann/Vela 2006).

3.4. Querying

When working with translation corpora, raw text, part-of-speech tags as well as raw text in combination with part-of-speech tags can be queried for. As a result, the matching instances as well as their previously aligned equivalents (cf. section 3.1. above) are displayed in a KWIC output. Query tools which support working with multilingual texts are, for example, the parallel concordancer ParaConc (Barlow 1999), the Translation Corpus Explorer (Ebeling 1998) or the Corpus Query Processor (Christ 1994). As a result, both the target language matches and their aligned source language segments are displayed in the concordance.

4. Applications in translation and translatology

In translation studies, the role of corpora has traditionally been restricted to use in the computational branches of the discipline. In particular, corpora have been used in the fields of terminology and lexicography (see also article 8) as well as for the development of translation aids, e. g. translation memories or machine translation programs (see article 56). Only recently, large electronically available corpora have come to be used in translator education and training, translation practice and translation research.

4.1. Corpora in translator education and training

The basic idea of using corpora in translator education and training is that a parallel corpus consists of a more comprehensive and diverse variety of source language items and possible translation solutions than a dictionary could ever display (cf. Zanettin/

Bernardini/Stewart 2003). Thus, in translator training, parallel corpora are explored for terminology look-up (Pearson 2000, Danielsson/Ridings 2000, Maia 2003 as well as Bowker/Pearson 2002 in a Languages for Specific Purposes (LSP) context). On the other hand, they are also employed for teaching the usage of collocations (Teubert 2003 and Barlow 2000 using parallel corpora) as well as register- and typology-specific patterns of the target language (Pearson 2003 and Bowker 1999 using comparable corpora). Hansen/Teich (2002) show how an English-German translation reference corpus annotated with part-of-speech tags can be used to look up not only lexical items but also grammatical structures. Furthermore, a parallel corpus can show translation students and language learners how to deal with translation problems (see Pearson 2003 for English-German and Johansson/Hofland 2000 for English-Norwegian) and how to avoid typical mistakes (see, for instance, Vintar/Hansen 2005 analysing cognates in parallel texts).

One interesting approach to providing students with insights into possible translation strategies is to collect several translations of one and the same source language text (cf. Teubert 2001). A similar scenario is introduced by Johansson (2003), where translations into several languages are collected from one and the same source language text. Here, students are able to learn to which degree the linguistic structures of the source language text can be preserved in the target language or how they have to be transferred according to the norms of the target language.

Another common method of teaching and studying translation is the use of learner corpora (cf. article 15). Here, several translations of one and the same source language text produced by students are collected. A very easy way to collect such learner texts is to submit and store them electronically as proposed by Bowker/Bennison (2003) in their work on the Student Translation Archive. Also, possible translation errors or peculiarities can be tagged and explored in such a way that the students can learn from the translation behaviour of other learners and translators (cf. Malmkær 2003).

4.2. Corpora in translation practice

In translation practice, we have to distinguish between the translation of highly repetitive texts (such as manuals or instructions) and the translation of creative writing (e.g. fiction) or expository prose (e.g. (popular-) scientific texts). In the context of translating software manuals, for instance, it can be worthwhile to compile previously translated manuals, align them, and store them in a database, which can then be used for reference. Such translation memories are usually equipped with alignment and terminology management tools (e.g. Translator's Workbench by Trados, Déjà vu by Atril or Star Transit) and are thus useful both for terminology look-up and for the pre-translation of phrases and even of whole sentences.

But even for texts that are not highly repetitive, multilingual and especially parallel corpora can support the translation process. A parallel corpus can be employed as a multilingual lexical resource, being more comprehensive and diverse than dictionaries. A multilingually comparable corpus can be used for exploring register use as well as typological differences. This is extremely helpful for the translation of special purpose texts and the acquisition of highly specialised terminology since term banks and glossaries can be built up easily (Bowker/Pearson 2002). If a corpus is linguistically annotated,

it can also be used to help solve grammatical or semantic translation problems (Vela/Hansen-Schirra 2006). For the translation of literary texts, the investigation of a comparable corpus can reveal the personal style of an author or translator that may be incorporated in the translation (Baker 2000).

When speaking about corpora as reference resources for professional translators, another useful resource, which is freely available and easily accessible for many different languages, should be mentioned in this context: the World Wide Web (cf. Kilgarriff 2001 and article 18). Lexical look-up can, for example, be initiated via a search engine like Google. The matches are displayed in the form of links to web sites, which can be investigated. An easier way to linguistically explore the Web is the online tool WebCorp, which displays the matches in the form of a concordance, a KWIC output. Another useful application of the Web as a translation aid is to search for multilingual web sites, i.e. translated web sites, from which parallel texts can easily be downloaded, aligned and used as a parallel corpus or a translation memory.

4.3. Corpora in translation research

As mentioned at the beginning of this article, the importance of corpus linguistics has only recently been considered in the theoretical and descriptive branches of translation studies (cf. Olohan 2004). In related disciplines (e.g. contrastive linguistics), researchers even try to ban translations from corpora because translated text is regarded as inferior compared to originals and it is not considered worth investigating because it is generally constrained by the presence of a fully articulated text in another language. Sager (1984) suggests to examine translations as a special kind of text production and to look into their special characteristics. Here, the value of a translation is still regarded as being dependent on that of its original text. In contrast, Baker (1995) tries to exclude the influence of the source language on a translation in order to analyse characteristic patterns of translations independently of the source language. Within this context Baker (1996) formulates the following hypotheses on the universal features of translations: explicitation (translations are more explicit than originals), simplification (translations are easier to understand and more readable), normalisation (translations strongly adhere to the usage norms of the target language) and levelling out (translations are more alike than the individual texts in a corpus of originals). Other universals, like the ‘unique items’ hypothesis (cf. Tirkkonen-Condit 2004) or the SL interference observation (e.g. Teich 2003, Mauranen 2004 or Eskola 2004), have been added to the list of specific properties of translations.

These hypotheses are tested in several studies using for instance TEC and BNC as comparable corpora (cf. section 2.2.): Laviosa-Braithwaite (1996) tests simplification and levelling out by analysing average sentence length, lexical density and type-token ratio. Olohan/Baker (2000) find evidence of explicitation in TEC investigating the use of *that*-connectives in contrast to zero-connectives of the reporting verbs *say* and *tell*. Furthermore, Olohan tests the hypothesis of explicitation on the basis of the use of contractions (Olohan 2003) and optional syntactic elements in translations (Olohan 2004). Hansen (2003) finds evidence of normalisation in a tagged version of TEC and explains this result with the help of a psycholinguistic experiment.

Based on comparable corpora, translation universals, such as explicitation, simplification or normalisation, are also tested for languages other than English (e. g. Bernardini/Zanettin 2004 or Mauranen/ Kujamäki 2004).

Using a corpus of English source language texts, German translations, and comparable German originals, Hansen/Teich (1999) investigate the above mentioned translation features explicitation, simplification, normalisation and levelling out. The use of a combined parallel-comparable corpus allows them to identify the influence of the source language texts on the translations and on the target language. Teich (2003), for instance, finds a special kind of source language interference (the typical language use of the source language “shines through” in the German translations).

Parallel corpora are used for the investigation of information structure in English and German texts (e. g. Doherty 1999), thematic structure (Hasselgård 1998), information packaging (e. g. Steiner 2002, 2004 for English-German and Fabricius-Hansen 1999 for German-English and German-Norwegian), explicitation in English and Norwegian parallel texts (Johansson 1995) and the relation between explicitness/explicitation and cohesion in German-English and English-German translations (Hansen-Schirra/Neumann/Steiner 2006).

Finally, corpus-based methods are used in order to investigate translator's style (Baker 2000, Olohan 2004), creativity in translation (Kenny 1998) as well as intercultural issues (Mauranen 1997). Such analyses can serve as input for translators' education and in translation practice as well as provide a basis for translation criticism.

5. Links to other fields

The field of corpus-based translation studies has links with a number of other corpus-oriented areas of research.

In terms of research objectives and methodologies, there is a high degree of complementarity with modern contrastive linguistics (cf. article 54). This is documented by the work of a number of research groups that use contrastive analysis to inform research on translation and vice versa, e. g. the Oslo SPRIK group (e. g. Johansson 2002), the Saarbrücken corpus-based translation group (e. g. Hansen-Schirra/Neumann/Steiner 2006) and the Hamburg research center on multilinguality (*Sonderforschungsbereich 538 Mehrsprachigkeit*; e. g. House 2002).

In terms of techniques, there are links with the area of tool development for corpus linguistics: corpus-based translation studies makes use of standard corpus technology, such as KWIC concordances (article 33), PoS tagging and higher level annotation (articles 23 and 24), and text-translation alignment (article 56). It is also a potential user of more sophisticated techniques, such as linguistic database technology based on XML. The use of such techniques in corpus-based translation research is still rare, however; but the research issues of corpus-based translation studies can also act as a driving force for the development of new computational tools (e. g. alignment). Finally, in terms of corpus encoding, researchers in corpus-based translation studies recognize and adhere to mark-up standards, such as TEI and their various incarnations in SGML and XML, thus linking up with Humanities Computing (article 22).

In summary, the field of corpus-based translation studies is a rather young discipline, to which corpora have been introduced only about a couple of decades ago. An appro-

priate reflection of the field at its present stage is Olohan's recent monograph *Introducing Corpora in Translation Studies* (Olohan 2004). With the advent of corpora, translation studies encounters new opportunities and new challenges, methodological and technical as well as theoretical. From the perspective of methodology, the primary question is for which types of research questions corpora may be suitably employed; from the technical perspective, the primary issue is to decide on the computational tools to be used for analysis and, if necessary, to implement tools specifically suited to translation analysis (e.g. aligners, parallel concordancers, multilingual databases). From a theoretical perspective, the application of corpora in translation studies can be used to test existing theories and models of translation, to refine them, and to push them further by raising new issues, such as language contact and language change through translation.

6. Literature

- Aijmer, K./Altenberg, B./Johansson, M. (eds.) (1996), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.
- Baker, M. (1995), Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. In: *Target* 7(2), 223–245.
- Baker, M. (1996), Corpus-based Translation Studies: The Challenges that Lie Ahead. In: Somers, H. L. (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins, 175–186.
- Baker, M. (2000), Towards a Methodology for Investigating the Style of a Literary Translator. In: *Target* 12(2), 241–266.
- Barlow, M. (1999), MonoConc 1.5 and ParaConc. In: *International Journal of Corpus Linguistics* 4(1), 319–327.
- Barlow, M. (2000), Parallel Texts in Language Teaching. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 106–115.
- Baroni, M./Bernardini, S. (2006), A New Approach to the Study of Translationese: Machine-learning the Difference Between Original and Translated Text. In: *Literary and Linguistic Computing* 21(2), 259–274.
- Baumgarten, N./House, J./Probst, J. (2004), English as Lingua Franca in Covert Translation Processes: A Project Report. In: *The Translator* 10(1), 83–108.
- Bernardini, S./Zanettin, F. (2004), When is a Universal not a Universal? Some Limits of Current Corpus-based Methodologies for the Investigation of Translation Universals. In: Mauranen/Kujamäki 2004, 51–62.
- Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
- Bowker, L. (1999), Exploring the Potential of Corpora for Raising Language Awareness in Student Translators. In: *Language Awareness* 8, 160–173.
- Bowker, L./Bennison, P. (2003), Student Translation Archive: Design, Development and Application. In: Zanettin/Bernardini/Stewart 2003, 103–117.
- Bowker, L./Pearson, J. (2002), *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routledge.
- Christ, O. (1994), A Modular and Flexible Architecture for an Integrated Corpus Query System. In: *Proceedings of COMPLEX'94*. Budapest, Hungary, 23–32.
- Danielsson, P./Ridings, D. (2000), Corpus and Terminology: Software for the Translation Program at Göteborgs Universitet or Getting Students to Do the Work. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 65–72.

- Doherty, M. (1999), Clefts in Translations between English and German. In: *Target* 11(2), 289–315.
- Ebeling, J. (1998), The Translation Corpus Explorer: A Browser for Parallel Texts. In: Johansson, S./Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*. Amsterdam: Rodopi, 101–112.
- Erjavec, T. (1999), Making the ELAN Slovene/English Corpus. In: *Proceedings of the Workshop "Language Technologies – Multilingual Aspects"*. Ljubljana, Slovenia, 23–30.
- Eskola, S. (2004), Unusual Frequencies in Translated Language: A Corpus-based Study on a Literary Corpus of Translated and Non-translated Finnish. In: Mauranen/Kujamäki 2004, 83–99.
- Fabricius-Hansen, C. (1998), Information Density and Translation, with Special Reference to German – Norwegian – English. In: Johansson, S./Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi, 197–234.
- Fabricius-Hansen, C. (1999), Information Packaging and Translation: Aspects of Translational Sentence Splitting (German – English/Norwegian). In: Doherty, M. (ed.), *Sprachspezifische Aspekte der Informationsverteilung*. (Studia Grammatica 47.) Berlin: Akademie-Verlag, 175–214.
- Germann, U. (2001), *Aligned Hansards of the 36th Parliament of Canada*. Available at: <http://www.isi.edu/natural-language/download/hansard>.
- Halliday, M. A. K./Matthiessen, C. M. I. M. (2004), *An Introduction to Functional Grammar*. Oxford: Oxford University Press.
- Hansen, S. (2003), *The Nature of Translated Text – an Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. (Saarbrücken Dissertations in Computational Linguistics and Language Technology 13.) Saarbrücken: Universität des Saarlandes.
- Hansen, S./Teich, E. (1999), Kontrastive Analyse von Übersetzungskorpora: Ein funktionales Modell. In: Gippert, J. (ed.), *Sammelband der Jahrestagung der GLDV 99*. Prag: Enigma Corporation, 311–322.
- Hansen, S./Teich, E. (2001), Multi-layer Analysis of Translation Corpora: Methodological Issues and Practical Implications. In: *Proceedings of EUROLAN 2001 Workshop on Multi-layer Corpus-based Analysis*. Iasi, Romania, 44–55.
- Hansen, S./Teich, E. (2002), The Creation and Exploitation of a Translation Reference Corpus. In: *Proceedings of the First International Workshop on Language Resources for Translation Work and Research*. Las Palmas de Gran Canaria, Spain, 1–4.
- Hansen-Schirra, S. (2003), Linguistic Enrichment and Exploitation of the Translational English Corpus. In: *Proceedings of Corpus Linguistics 2003*. Lancaster, United Kingdom, 288–297.
- Hansen-Schirra, S./Neumann, S./Steiner, E. (2006), Cohesion and Explicitation in an English-German Translation Corpus. In: *Proceedings of SPRIK Conference 2006: Explicit and Implicit Information in Text – Information Structure across Languages*. Oslo, Norway, 8–11.
- Hansen-Schirra, S./Neumann S./Vela, M. (2006), Multi-dimensional Annotation and Alignment in an English–German Translation Corpus. In: *Proceedings of the workshop on Multi-dimensional Markup in Natural Language Processing (NLPXML-2006)*. Trento, Italy, 35–42.
- Hasselgård, H. (1998), Thematic Structure in Translation between English and Norwegian. In: Johansson, S./Oksefjell, S. (eds.), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*. Amsterdam: Rodopi, 145–167.
- Heyn, M. (1996), Integrating Machine Translation into Translation Memory Systems. In: *European Association for Machine Translation – Workshop Proceedings, ISSCO*. Geneva, Switzerland, 111–123.
- Hofland, K. (1996), A Program for Aligning English and Norwegian Sentences. In: Hockey, S./Ide, N./Perissinotto, G. (eds.), *Research in Humanities Computing*. Oxford: Oxford University Press, 165–178.
- House, J. (2002), Maintenance and Convergence in Covert Translation English–German. In: Hasselgård, H./Johansson, S./Behrens, B./Fabricius-Hansen, C. (eds.), *Information Structure in a Cross-linguistic Perspective*. Amsterdam: Rodopi, 199–213.

- Johansson, S. (1995), Mens sana in corpore sano: On the Role of Corpora in Linguistic Research. In: *The European English Messenger* 4(2), 19–25.
- Johansson, S. (2002), Towards a Multilingual Corpus for Contrastive Analysis and Translation Studies. In: Borin, E. (ed.), *Parallel Corpora, Parallel Worlds*. Amsterdam: Rodopi, 47–59.
- Johansson, S. (2003), Reflections on Corpora and their Uses in Cross-linguistic Research. In: Zanettin/Bernardini/Stewart 2003, 133–144.
- Johansson, S./Ebeling, J./Hofland, K. (1996), Coding and Aligning the English–Norwegian Parallel Corpus. In: Aijmer/Altenberg/Johansson 1996, 87–112.
- Johansson, S./Hofland, K. (2000), The English–Norwegian Parallel Corpus: Current Work and New Directions. In: Botley, S. P./McEnery, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 134–147.
- Johansson, S./Oksefjell, S. (eds.) (1998), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.
- Kenny, D. (1997), Creatures of Habit? What Collocations Can Tell us about Translation. Poster presented at ACH/ALLC 97, Ontario, Canada. Text available at: <http://www.cs.queensu.ca/achallc97/papers/a006.html>.
- Kenny, D. (1998), Creatures of Habit? What Translators Usually Do with Words. In: *Meta: Special Issue on the Corpus-based Approach* 43(4), 515–523.
- Kilgarriff, A. (2001), Web as Corpus. In: *Proceedings of Corpus Linguistics 2001*. Lancaster, United Kingdom, 342–344.
- Koehn, P. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of MT Summit 2005*. Phuket, Thailand, 79–86. Available at: <http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf>.
- Lambert, S./Moser-Mercer, B. (eds.) (1994), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*. Amsterdam/Philadelphia: John Benjamins.
- Laviosa-Braithwaite, S. (1996), *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. PhD Thesis. Manchester: UMIST.
- Lörscher, W. (1991), *Translation Performance, Translation Process, and Translation Strategies. A Psycholinguistic Investigation*. Tübingen: Gunter Narr.
- Maia, B. (2003), “Some Languages are More Equal than Others”: Training Translators in Terminology and Information Retrieval Using Comparable and Parallel Corpora. In: Zanettin/Bernardini/Stewart 2003, 43–53.
- Malmkjær, K. (2003), On a Pseudo-subversive Use of Corpora in Translator Training. In: Zanettin/Bernardini/Stewart 2003, 119–134.
- Mauranen, A. (1997), Hedging in Language Revisers’ Hands. In: Markkanen, R./Schröder, H. (eds.), *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Berlin: Walter de Gruyter, 115–133.
- Mauranen, A. (2004), Corpora, Universals and Interference. In: Mauranen/Kujamäki 2004, 65–82.
- Mauranen, A./Kujamäki, P. (eds.) (2004), *Translation Universals: Do they Exist?* Amsterdam/Philadelphia: John Benjamins.
- Meyer, B. (1998), What Transcripts of Authentic Discourse Can Reveal about Interpreting. In: *Interpreting* 3(1), 65–83.
- Meyer, B./Apfelbaum, B./Bischoff, A./Pöchhacker, F. (2003), Analyzing Interpreted Doctor-patient Communication from the Perspectives of Linguistics, Interpreting Studies and Health Sciences. In: Brunette, L./Bastin, G./Hemlin, I./Clarke, H. (eds.), *Interpreters in the Community. Selected Papers from the Third International Conference on Interpreting in Legal, Health and Social Service, Montreal, March 2001*. Amsterdam: John Benjamins, 67–79.
- Mihailov, M. (2001), Two Approaches to Automated Text Aligning of Parallel Fiction Texts. In: *Across Languages and Cultures* 2(1). Budapest: Akadémiai Kiadó, 87–96.
- Neumann, S. (2003), *Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer*. Frankfurt: Peter Lang.

- Neumann, S./Hansen-Schirra, S. (eds.) (2003), *Proceedings of Multilingual Corpora: Linguistic Requirements and Technical Perspectives*. Lancaster, United Kingdom.
- Nida, E. (1964), *Toward a Science of Translating*. Leiden: E. J. Brill.
- Och, F. J./Ney, H. (2003), A Systematic Comparison of Various Statistical Alignment Models. In: *Computational Linguistics* 29(1), 19–51.
- Olohan, M. (2003), How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus. In: *Target* 15(1), 59–89.
- Olohan, M. (2004), *Introducing Corpora in Translation Studies*. London/New York: Routledge.
- Olohan, M./Baker, M. (2000), Reporting that in Translated English: Evidence for Subconscious Processes of Explicitation? In: *Across Languages and Cultures* 1(2), 141–172.
- Pearson, J. (2000), Teaching Terminology Using Electronic Resources. In: Botley, S. P./McEnergy, A. M./Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 92–115.
- Pearson, J. (2003), Using Parallel Texts in the Translator Training Environment. In: Zanettin/Bernardini/Stewart 2003, 15–24.
- Pöchhacker, F. (2000), *Dolmetschen: Konzeptuelle Grundlagen und deskriptive Untersuchungen*. Tübingen: Stauffenburg 2000.
- Sager, J. C. (1984), Reflections on the Didactic Implications of an Extended Theory of Translation. In: Wilss, W./Thome, G. (eds.), *Translation Theory and its Implementation in the Teaching of Translating and Interpreting*. Tübingen: Gunter Narr, 333–343.
- Sager, J. C. (1994), *Language Engineering and Translation: Consequences of Automation*. Amsterdam/Philadelphia: John Benjamins.
- Samuelsson, Y./Volk, M. (2005), Presentation and Representation of Parallel Treebanks. In: *Proceedings of the Nodalida Workshop on Treebanks for Spoken Language and Discourse*. Joensuu, Finland. Available at: http://ling16.ling.su.se:8080/new_PubDB/doc_repository/214_Samuelsson_Volk_2005.pdf.
- Shlesinger, M. (1998), Corpus-based Interpreting Studies. In: *Meta: Special Issue on the Corpus-based Approach* 43(4), 468–493.
- Snell-Hornby, M. (1988), *Translation Studies – an Integrated Approach*. Amsterdam/Philadelphia: John Benjamins.
- Steiner, E. (2002), Grammatical Metaphor in Translation – Some Methods for Corpus-based Investigations. In: Hasselgård, H./Johansson, S./Behrens, B./Fabricius-Hansen, C. (eds.), *Information Structure in a Cross-linguistic Perspective*. Amsterdam: Rodopi, 213–228.
- Steiner, E. (2004), Ideational Grammatical Metaphor: Exploring Some Implications for the Overall Model. In: *Languages in Contrast* 4(1), 139–166.
- Teich, E. (2003), *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin and New York: Walter de Gruyter.
- Teich, E./Hansen, S./Fankhauser, P. (2001), Representing and Querying Multi-layer Annotated Corpora. In: *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia, PA, 228–237.
- Teubert, W. (ed.) (2001), *Text Corpora and Multilingual Lexicography. International Journal of Corpus Linguistics* 6 (Special Issue).
- Teubert, W. (2003), Collocations, Parallel Corpora and Language Teaching. In: *Selected Papers from the Twelfth International Symposium on English Teaching*. Taipei: English Teacher's Association, 143–156.
- Tirkkonen-Condit, S. (2004), Unique Items – Over- or Under-represented in Translated Language? In: Mauranen/Kujamäki 2004, 177–184.
- Toury, G. (1995), *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins.
- Vela, M./Hansen-Schirra, S. (2006), The Use of Multi-level Annotation and Alignment for the Translator. In: *Proceedings of the Translating and the Computer 28 Conference*. London, United Kingdom. Available at: http://fr46.uni-saarland.de/croco/vela_hansenschirra_aslib2006.pdf.

- Vinay, J.-P./Darbelnet, J. (1995), *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam/Philadelphia: John Benjamins.
- Vintar, S./Hansen, S. (2005), Cognates – Free Rides, False Friends or Stylistic Devices: A Corpus-based Comparative Study. In: Barnbrook, G./Danielsson, P./Mahlberg, M. (eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: Birmingham University Press, 208–221.
- Volk, M./Gustafson-Capková, S./Lundborg, J./Marek, T./Samuelsson, Y./Tidström, F. (2006), XML-based Phrase Alignment in Parallel Treebanks. In: *Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*. Trento, Italy, 93–96. Available from: <http://acl.ldc.upenn.edu/W/W06/W06-2717.pdf>.
- Zanettin, F./Bernardini, S./Stewart, D. (2003), *Corpora in Translator Education*. Manchester: St. Jerome.

Silvia Hansen-Schirra, Germersheim and Elke Teich, Darmstadt (Germany)

56. Corpora and machine translation

1. Introduction
2. Corpus-based tools for translators
3. Example-based MT (EBMT)
4. Statistical MT (SMT)
5. Variants of SMT
6. Rapid development of MT for less-studied languages
7. Conclusions
8. Acknowledgments
9. Literature

1. Introduction

This article concerns the use of corpora in Machine Translation (MT), and, to a lesser extent, the contribution of corpus linguistics to MT and vice versa. MT is of course perhaps the oldest non-numeric application of computers, and certainly one of the first applications of what later became known as natural language processing. However, the early history of MT is marked at first (between roughly 1948 and the early 1960s) by fairly ad hoc approaches as dictated by the relatively unsophisticated computers available, and the minimal impact of linguistic theory. Then, with the emergence of more formal approaches to linguistics, MT warmly embraced – if not exactly a Chomskyan approach – the use of linguistic rule-based approaches which owed a lot to transformational generative grammar. Before this, Gil King (1956) proposed some “stochastic” methods for MT, foreseeing the use of collocation information to help in word-sense disambiguation, and suggesting that distribution statistics should be collected so that, lacking any other information, the most common translation of an ambiguous word

- Vinay, J.-P./Darbelnet, J. (1995), *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam/Philadelphia: John Benjamins.
- Vintar, S./Hansen, S. (2005), Cognates – Free Rides, False Friends or Stylistic Devices: A Corpus-based Comparative Study. In: Barnbrook, G./Danielsson, P./Mahlberg, M. (eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: Birmingham University Press, 208–221.
- Volk, M./Gustafson-Capková, S./Lundborg, J./Marek, T./Samuelsson, Y./Tidström, F. (2006), XML-based Phrase Alignment in Parallel Treebanks. In: *Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*. Trento, Italy, 93–96. Available from: <http://acl.ldc.upenn.edu/W/W06/W06-2717.pdf>.
- Zanettin, F./Bernardini, S./Stewart, D. (2003), *Corpora in Translator Education*. Manchester: St. Jerome.

Silvia Hansen-Schirra, Germersheim and Elke Teich, Darmstadt (Germany)

56. Corpora and machine translation

1. Introduction
2. Corpus-based tools for translators
3. Example-based MT (EBMT)
4. Statistical MT (SMT)
5. Variants of SMT
6. Rapid development of MT for less-studied languages
7. Conclusions
8. Acknowledgments
9. Literature

1. Introduction

This article concerns the use of corpora in Machine Translation (MT), and, to a lesser extent, the contribution of corpus linguistics to MT and vice versa. MT is of course perhaps the oldest non-numeric application of computers, and certainly one of the first applications of what later became known as natural language processing. However, the early history of MT is marked at first (between roughly 1948 and the early 1960s) by fairly ad hoc approaches as dictated by the relatively unsophisticated computers available, and the minimal impact of linguistic theory. Then, with the emergence of more formal approaches to linguistics, MT warmly embraced – if not exactly a Chomskyan approach – the use of linguistic rule-based approaches which owed a lot to transformational generative grammar. Before this, Gil King (1956) proposed some “stochastic” methods for MT, foreseeing the use of collocation information to help in word-sense disambiguation, and suggesting that distribution statistics should be collected so that, lacking any other information, the most common translation of an ambiguous word

should be output (of course he did not use these terms). Such ideas did not resurface for a further 30 years however.

In parallel with the history of corpus linguistics, little reference is made to “corpora” in the MT literature until the 1990s, except in the fairly informal sense of “a collection of texts”. So for example, researchers at the TAUM group (Traduction Automatique Université de Montréal) developed their notion of sublanguage-based MT on the idea that a sublanguage might be defined with reference to a “corpus”: “Researchers at TAUM [...] have made a detailed study of the properties of texts consisting of instructions for aircraft maintenance. The study was based on *a corpus of 70,000 words* of running text in English” (Lehrberger 1982, 207; emphasis added). And in the Eurotra MT project (1983–1990), involving 15 or more groups working more or less independently, a multilingual parallel text in all (at the time) nine languages of the European Communities was used as a “reference corpus” to delimit lexical and grammatical coverage of the system. Apart from this, developers of MT systems worked in a largely theory-driven (rather than data-driven) manner, as characterised by Isabelle (1992a, iii) in his Preface to the proceedings of the landmark TMI Conference of that year: “On the one hand, the “rationalist” methodology, which has dominated MT for several decades, stresses the importance of basing MT development on better theories of natural language On the other hand, there has been renewed interest recently in more “empirical” methods, which give priority to the analysis of large corpora of existing translations”

The link between MT and corpora really first became established with the emergence of statistics-based MT (SMT) from 1988 onwards. The IBM group at Yorktown Heights, NY had got the idea of doing SMT based on their success with speech recognition, and then had to look round for a suitable corpus (Fred Jelinek, personal communication). Fortunately, the Canadian parliament had in 1986 started to make its bilingual (English and French) proceedings (Hansard) available in machine-readable form. However, again, the “corpus” in question was really just a collection of raw text, and the MT methodology had no need in the first instance of any sort of mark-up or annotation (cf. articles 16 and 32). In section 4 we will explain how SMT works and how it uses techniques of interest to corpus linguists.

The availability of large-scale parallel texts gave rise to a number of developments in the MT world, notably the emergence of various tools for translators based on them, the “translation memory” (TM) being the one that has had the greatest impact, though parallel concordancing also promises to be of great benefit to translators (see section 2). Both of these applications rely on the parallel text having been *aligned*, techniques for which are described in articles 16 and 32. Not all TMs are corpus-based however, as will be discussed in section 2.2.

Related to, but significantly different from TMs, is an approach to MT termed “Example-Based MT” (EBMT). Like TMs, this takes the idea that new translations can use existing translations as a model, the difference being that in EBMT it is the computer rather than the translator that decides how to manipulate the existing example. As with TMs, not all EBMT systems are corpus-based, and indeed the provenance of the examples that are used to populate the TM or the example-base is an aspect of the approach that is open to discussion. Early EBMT systems tended to use hand-picked examples, whereas the latest developments in EBMT tend to be based more explicitly on the use of naturally occurring parallel corpora also making use in some cases of mark-up and

annotations, this extending in one particular approach, to tree banks (cf. articles 13 and 28). All these issues are discussed in section 3. Recent developments in EBMT and SMT have seen the two paradigms coming closer together, to such an extent that some commentators doubt there is a significant difference.

One activity that sees particular advantage in corpus-based approaches to MT, whether SMT or EBMT, is the rapid development of MT for less-studied (or “low density”) languages (cf. article 21). The essential element of corpus-based approaches to MT is that they allow systems to be developed automatically, in theory without the involvement of language experts or native speakers. The MT systems are built by programs which “learn” the translation relationships from pre-existing translated texts, or apply methods of “analogical processing” to infer new translations from old. This learning process may be helped by some linguistically-aware input (for example, it may be useful to know what sort of linguistic features characterise the language pair in question) but in essence the idea is that an MT system for a new language pair can be built just on the basis of (a sufficient amount of) parallel text. This is of course very attractive for “minority” languages where typically parallel texts such as legislation or community information in both the major and minor languages exist. Most of the work in this area has been using the SMT model, and we discuss these developments in section 6.

2. Corpus-based tools for translators

Since the mid-1980s, parallel texts in (usually) two languages have become increasingly available in machine-readable form. Probably the first such “bitext” of significant size, to use the term coined by Harris (1988), was the Canadian Hansard mentioned above. The Hong Kong parliament, with proceedings at that time in English and Cantonese, soon followed suit, and the parallel multilingual proceedings of the European Parliament are a rich source of data; but with the explosion of the World Wide Web, parallel texts, sometimes in several languages, and of varying size and quality, soon became easily available.

Isabelle (1992b, 8) stated that “*Existing translations contain more solutions to more translation problems than any other existing resource*” [emphasis original], reflecting the idea, first proposed independently by Arthurn (1978), Kay (1980) and Melby (1981), that a store of past translations together with software to access it could be a useful tool for translators. The realisation of this idea had to wait some 15 years for adequate technology, but is now found in two main forms, parallel concordances, and TMs.

2.1. Parallel concordances

Parallel concordances have been proposed for use by translators and language learners, as well as for comparative linguistics and literary studies where translation is an issue (e.g. with biblical and quranic texts). An early implementation is reported by Church/Gale (1991), who suggest that parallel concordancing can be of interest to lexicographers, illustrated by the ability of a parallel concordance to separate the two French translations of *drug* (*médicament* ‘medical drug’ vs. *drogue* ‘narcotic’). An implementa-

tion specifically aimed at translators is *TransSearch*, developed since 1993 by RALI in Montreal (Simard/Foster/Perrault 1993), initially using the Canadian Hansard, but now available with other parallel texts. Part of a suite of *Trans-* tools, *TransSearch* was always thought of as a translation aid, unlike *ParaConc* (Barlow 1995) which was designed for the purpose of comparative linguistic study of translations, and *MultiConcord* (Romary/Mehl/Woolls 1995), aimed at language teachers. More recently, many articles dealing with various language combinations have appeared. In each case, the idea is that one can search for a word or phrase in one language, and retrieve examples of its use in the normal manner of a (monolingual) concordance, but in this case linked (usually on a sentence-by-sentence basis) to their translations. Apart from its use as a kind of lexical look-up, the concordance can also show contexts which might help differentiate the usage of alternate translations or near synonyms. Most systems also allow the use of wildcards, but also parallel search, so that the user can retrieve examples of a given source phrase coupled with a target word. This device can be used, among other things, to check for false-friend translations (e.g. French *librairie* as *library* rather than *bookshop*), or to distinguish, as above, different word senses.

A further use of a parallel corpus as a translator's aid is the RALI group's *TransType* (Foster/Langlais/Lapalme 2002), which offers translators text completion on the basis of the parallel corpus. With the source text open in one window, the translator starts typing the translation, and on the basis of the first few characters typed, the system tries to predict from the target-language side of the corpus what the translator wants to type. This predication capability is enhanced by Maximum Entropy, word- and phrase-based models of the target language and some techniques from Machine Learning. Part of the functionality of *TransType* is like a sophisticated TM, the increasingly popular translator's aid that we will discuss in the next section.

2.2. Translation memories (TMs)

The TM is one of the most significant computer-based aids for translators. First proposed independently by Arthern (1978), Kay (1980) and Melby (1981), but not generally available until the mid 1990s (see Somers/Fernández Díaz 2004, 6–8 for more detailed history), the idea is that the translator can consult a database of previous translations, usually on a sentence-by-sentence basis, looking for anything similar enough to the current sentence to be translated, and can then use the retrieved example as a model. If an exact match is found, it can be simply cut and pasted into the target text, assuming the context is similar. Otherwise, the translator can use it as a suggestion for how the new sentence should be translated. The TM will highlight the parts of the example(s) that differ from the given sentence, but it is up to the translator to decide which parts of the target text need to be changed.

One of the issues for TM systems is where the examples come from: originally, it was thought that translators would build up their TMs by storing their translations as they went along. More recently, it has been recognised that a pre-existing bilingual parallel text could be used as a ready-made TM, and many TM systems now include software for aligning such data (see article 16).

Although a TM is not necessarily a “corpus”, strictly speaking, it may still be of interest to discuss briefly how TMs work and what their benefits and limitations are. For a more detailed discussion, see Somers (2003).

Matching and equivalence

Apart from the question of where the data comes from, the main issue for TM systems is the problem of matching the text to be translated against the database so as to extract all and only the most useful cases to help and guide the translator. Most current commercial TM systems offer a quantitative evaluation of the match in the form of a “score”, often expressed as a percentage, and sometimes called a “fuzzy match score” or similar. How this score is arrived at can be quite complex, and is not usually made explicit in commercial systems, for proprietary reasons. In all systems, matching is essentially based on character-string similarity, but many systems allow the user to indicate weightings for other factors, such as the source of the example, formatting differences, and even significance of certain words. Particularly important in this respect are strings referred to as “placeables” (Bowker 2002, 98), “transwords” (Gaussier/Langé/Meunier 1992, 121), “named entities” (using the term found in information extraction) (Macklovitch/Russell 2000, 143), or, more transparently perhaps, “non-translatables” (*ibid.*, 138), i. e. strings which remain unchanged in translation, especially alphanumerics and proper names: where these are the only difference between the sentence to be translated and the matched example, translation can be done automatically. The character-string similarity calculation uses the well-established concept of “sequence comparison”, also known as the “string-edit distance” because of its use in spell-checkers, or more formally the “Levenshtein distance” after the Russian mathematician who discovered the most efficient way to calculate it. A drawback with this simplistic string-edit distance is that it does not take other factors into account. For example, consider the four sentences in (1).

- (1) a. Select ‘Symbol’ in the Insert menu.
- b. Select ‘Symbol’ in the Insert menu to enter a character from the symbol set.
- c. Select ‘Paste’ in the Edit menu.
- d. Select ‘Paste’ in the Edit menu to enter some text from the clip board.

Given (1a) as input, most character-based similarity metrics would choose (1c) as the best match, since it differs in only two words, whereas (1b) has eight additional words. But intuitively (1b) is a better match since it entirely includes the text of (1a). Furthermore (1b) and (1d) are more similar than (1a) and (1c): the latter pair may have fewer words different (2 vs. 6), but the former pair have more words in common (8 vs. 4), so the distance measure should count not only differences but also similarities.

The similarity measure in the TM system may be based on individual characters or whole words, or may take both into consideration. Although more sophisticated methods of matching have been suggested, incorporating linguistic “knowledge” of inflection paradigms, synonyms and even grammatical alternations (Cranias/Papageorgiou/Piperidis 1997; Planas/Furuse 1999; Macklovitch/Russell 2000; Rapp 2002), it is unclear whether any existing commercial systems go this far. To exemplify, consider (2a). The example (2b) differs only in a few characters, and would be picked up by any currently available TM matcher. (2c) is superficially quite dissimilar, but is made up of words which are related to the words in (2a) either as grammatical alternatives or near synonyms. (2d) is very similar in meaning to (2a), but quite different in structure. Arguably, any of (2b–d) should be picked up by a sophisticated TM matcher, but it is unlikely that any commercial TM system would have this capability.

- (2) a. When the paper tray is empty, remove it and refill it with paper of the appropriate size.
- b. When the tray is empty, remove it and fill it with the appropriate paper.
- c. When the bulb remains unlit, remove it and replace it with a new bulb
- d. You have to remove the paper tray in order to refill it when it is empty.

The reason for this is that the matcher uses a quite generic algorithm, as mentioned above. If we wanted it to make more sophisticated *linguistically*-motivated distinctions, the matcher would have to have some language-specific “knowledge”, and would therefore have to be different for different languages. It is doubtful whether the gain in accuracy would merit the extra effort required by the developers. As it stands, TM systems remain largely independent of the source language and of course wholly independent of the target language.

Nearly all TM systems work exclusively at the level of sentence matching. But consider the case where an input such as (3) results in matches like those in (4).

- (3) Select ‘Symbol’ in the Insert menu to enter a character from the symbol set.
- (4) a. Select ‘Paste’ in the Edit menu.
b. To enter a symbol character, choose the Insert menu and select ‘Symbol’.

Neither match covers the input sentence sufficiently, but between them they contain the answer. It would clearly be of great help to the translator if TM systems could present partial matches and allow the user to cut and paste fragments from each of the matches. This is being worked on by most of the companies offering TM products, and, in a simplified form, is currently offered by at least one of them, but in practice works only in a limited way, for example requiring the fragments to be of roughly equal length (see Somers/Fernández Díaz 2004).

Suitability of naturally occurring text

As mentioned above, there are two possible sources of the examples in the TM database: either it can be built up by the user (called “interactive translation” by Bowker 2002, 108), or else a naturally occurring parallel text can be aligned and used as a TM (“post-translation alignment”, ibid., 109). Both methods are of relevance to corpus linguists, although the former only in the sense that a TM collected in this way could be seen as a special case of a planned corpus. The latter method is certainly quicker, though not necessarily straightforward (cf. Macdonald 2001), but has a number of shortcomings, since a naturally occurring parallel text will not necessarily function optimally as a TM database.

The first problem is that it may contain repetitions, so that a given input sentence may apparently have multiple matches, but they might turn out to be the same. This of course could be turned into a good thing, if the software could recognize that the same phrase was being consistently translated in the same way, and this could bolster any kind of confidence score that the system might calculate for the different matches.

More likely though is that naturally occurring parallel text will be internally *inconsistent*: a given phrase may have multiple translations either because different translations are appropriate in different contexts, or because the phrase has been translated in different ways for no reason other than that translators have different ideas or like to introduce variety into their translations. Where different contexts call for different trans-

lations, then the parallel corpus is of value assuming that it can show the different contexts, as discussed in the previous section. For example, the simple phrase *OK* in a conversation may be translated into Japanese as *wakarimashita* ‘I understand’, *iidesu yo* ‘I agree’ or *ijō desu* ‘let’s change the subject’, depending on the context (example from Somers/Tsujii/Jones 1990, 403). However, this is not such a big problem because the TM is a translator’s *tool*, and in the end the responsibility for choosing the translation is the user’s. The problem of suitability of examples is more serious in EBMT, as we will discuss below.

3. Example-based MT (EBMT)

EBMT is often thought of as a sophisticated type of TM, although in fact this approach to MT initially developed somewhat independently of the TM idea, albeit around the same time. In this section we will explain briefly how it works, and clarify some important differences between TMs and EBMT.

The idea for EBMT surfaced in the early 1980s (the seminal paper presented by Makoto Nagao at a 1981 conference was not published until three years later – Nagao 1984), but the main developments were reported from about 1990 onwards, and it has slowly become established within the mainstream of MT research (cf. Carl/Way 2003, 2006–2007). Pioneers were mainly in Japan, including Sato/Nagao (1990) and Sumita/Iida/Kohyama (1990). As in a TM, the basic idea is to use a database of previous translations, the “example-base”, and the essential first step, given a piece of text to translate, is to find the best match(es) for that text. Much of what was said above regarding matching in TM systems also applies to EBMT, though it should be said that earlier implementations of EBMT often had much more complex matching procedures, linked to the fact that examples were often stored not just as plain text but as annotated tree or other structures, often explicitly aligned.

Once the match has been found, the two techniques begin to diverge. While the work of the TM system is over (the translator decides what to do with the matches), in EBMT the system now has to manipulate the example so as to produce a translation. This is done in three steps: first, the source text and the examples are aligned so as to highlight which parts of the examples correspond to text in the sentence to be translated. Next, and crucially, the corresponding target-language fragments of text must be identified in the translations associated with the matches. Finally, the target translation is composed from the fragments so identified.

We can illustrate the process with a simple example. Suppose the input sentence is (5), and the matching algorithm identifies as relevant to its translation the examples in (6) with their French translations. The fragments of text in the examples that match the input are underlined.

- (5) The operation was interrupted because the file was hidden.
- (6)
 - a. The operation was interrupted because the Ctrl-c key was pressed.
L'opération a été interrompue car la touche Ctrl-c a été enfoncée.
 - b. The specified method failed because the file is hidden.
La méthode spécifiée a échoué car le fichier est masqué.

The EBMT process must now pick out from the French examples in (6) which words correspond to the underlined English words, and then combine them to give the proposed translation. These two operations are known in the EBMT literature as “alignment” and “recombination”.

3.1. Alignment in EBMT

Alignment, similar to but not to be confused with the notion of aligning parallel corpora in general, involves identifying which words in the target-language sentences correspond to the source-language words that we have identified as being of interest. An obvious way to do this might be with the help of a bilingual dictionary, and indeed some EBMT systems do work this way. However, one of the attractions of EBMT is the idea that an MT system can be built up on the basis only of large amounts of parallel data, with lexical alignments extracted from the examples automatically by analogy. This idea is of interest to corpus linguists, and indeed there is a literature around this topic (cf. article 32). In particular, techniques relying on simple probabilities using contingency tables and measures such as Dice’s coefficient, are well explored.

Within EBMT, there is a strand of research which seeks to generalize similar examples and thereby extract lexical correspondences, as follows: suppose that in the example base we have the sentences in (7), with their corresponding Japanese translations.

- (7) a. The monkey ate a peach. \leftrightarrow *Saru wa momo o tabeta*.
- b. The man ate a peach. \leftrightarrow *Hito wa momo o tabeta*.

From the sentence pairs in (7) we can assume that the difference between the two English sentences, *monkey* vs. *man*, corresponds to the only difference between the two Japanese translations, *saru* vs. *hito*. Furthermore we can assume that the remaining parts which the two sentences have in common also represent a partial translation pair (8).

- (8) The X ate a peach. \leftrightarrow *X' wa momo o tabeta*.

Comparison with further examples which are minimally different will allow us to build up both a lexicon of individual word pairs, and a “grammar” of transfer template pairs. Ideas along these lines have been explored for example by Cicekli/Güvenir (1996), Cicekli (2006), Brown (2000, 2001) and by McTait/Trujillo (1999).

3.2. Recombination

Once the appropriate target-language words and fragments have been identified, it should be just a matter of sticking them together. At this stage, however, a further problem arises, generally referred to in the literature as “boundary friction” (Nirenburg/ Domashnev/Grannes 1993, 48; Collins 1998, 22): fragments taken from one context may not fit neatly into another slightly different context. For example, if we have the translation pair in (9) and replace *man* with *woman*, the resulting translation, with *homme*

replaced by *femme* is quite ungrammatical, because French requires gender agreement between the determiner, adjective and noun.

- (9) The old man is dead. \leftrightarrow *Le vieil homme est mort.*

Another problem is that the fragments to be pasted together sometimes overlap: if we look again at examples (5) and (6), the fragments we have to recombine are the French equivalents of the templates shown in (10a, b) from (6a, b) respectively.

- (10) a. The operation was interrupted because the ... was
b. The ... because the file ... hidden.

A number of solutions to these two difficulties have been suggested, including the incorporation of target-language grammatical information which itself might be derived from a parallel corpus (Wu 1997), or, of more interest to corpus linguists, a model of target-language word sequences, or matching the proposed target sentence against the target side of the bilingual corpus.

4. Statistical MT (SMT)

SMT in its various forms is probably the approach to MT whose techniques and methodologies are most familiar to corpus linguists. In this section, we will discuss briefly the main ideas behind SMT, and some of the latest developments.

In its pure form, the statistics-based approach to MT makes use of no traditional linguistic data. The essence of the method is first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language. An essential feature is the availability of a suitable large bilingual corpus of reliable (authoritative) translations.

The “empirical” approach to MT was pioneered by the IBM research group at Yorktown Heights, NY, who had had some success with non-linguistic approaches to speech recognition, and turned their attention to MT in the early 1990s (Brown et al. 1990). Perhaps because of the promise this approach showed – systems could be built in a matter of weeks which came fairly close to the quality of rule-based systems which had taken many person-years to build – or simply because of the attraction of a rather new slant on an old problem, an SMT approach was taken up by a number of groups.

As already mentioned, the idea is to “model” the translation process in terms of statistical probabilities. For a given source-language sentence S , there are an infinite number of “translations” T of varying probability. The idea of SMT is to find just the T that maximizes the probability $P(T|S)$. This probability is seen as a function of two elements: a set $\{t_1, t_2, \dots, t_m\}$ of most probable target-language words given the set of source-language words $\{s_1, s_2, \dots, s_n\}$ which make up S , and the most probable order in which that given set of target-language words might occur. These two elements are referred to as the “translation model” and the “language model” respectively. Both are computed on the basis of the bilingual corpus.

The translation process in SMT therefore consists of applying the translation model to a given source sentence S to produce a set of probable words, and then applying the language model to those words to produce the target sentence T . However, since there are different probabilities involved, this is not a straightforward calculation, because the different probabilities interact. In effect, we start with the target-language words which look most likely to be part of the solution, see how these choices fit with the language model, and, in a systematic way, keep trying different combinations until we cannot improve the overall “score” any more. This so-called “decoding” stage of SMT is further discussed below.

4.1. The translation model

The translation model is the set of probabilities for each word on the source-language side of the corpus that it corresponds to or gives rise to each word on the target-language side of the corpus. Of course for many of these word pairs, the probability will be close to 0. The hope is that for words which are translational equivalents, the probabilities will be suitably high. One problem for this approach is that, as all linguists know, there is generally not a 1:1 correspondence between the words of one language and another. For example, French translations of adjectives in English have different forms depending on gender and number agreement. Homonyms in one language will have different translations in the target language. Importantly also some single words in one language may be translated by a string of words in the other language: for example, the single word *implemented* in English may be rendered in French as *mis en application*. This is referred to as the “fertility” of the source-language word. For this reason, the translation model includes not just word-pair translation probabilities, but a second set of parameters measuring the probability of different fertilities.

For practical reasons, these may be restricted to a small given range, for example 0–2 (0, because a word on the source side may “disappear” in translation, for example the two English words *may have* give rise to just one French word *aurait*). Fertility is nicely illustrated in the original IBM work (Brown et al. 1990), with data from the bilingual Canadian Hansards. The English word *the* translates as *le* with $P = .610$, *la* with $P = .178$, and some other words with much smaller values. The fertility probability is .817 for $f = 1$. The word *not* on the other hand translates as *pas* with $P = .469$, *ne* with $P = .460$, that is, with roughly equal probability. The fertility probabilities are .758 for $f = 2$, .133 for $f = 0$ and .106 for $f = 1$. In other words, the French for *not* is very likely to be *ne ... pas*. One last example is particular to the Hansard corpus. Very frequent in this corpus is the English phrase *hear hear*. The English word *hear* is coupled with the French *bravo* with $P = .992$ (and with much lower probabilities to various forms of the French verb *entendre*); the fertility probabilities are almost evenly split between $f = 0$ ($P = .584$) and $f = 1$ ($P = .416$). In other words, *hear* is almost certain to be translated as *bravo*, when it is translated at all, but half the time it should be simply omitted.

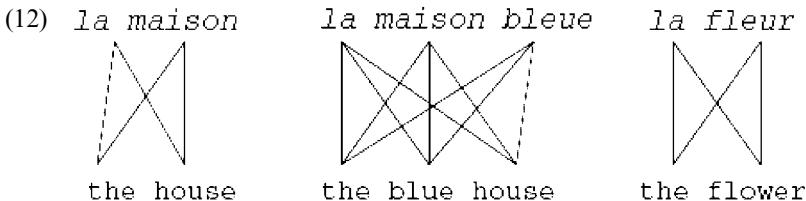
One can imagine various different methods of computing these probabilities. Assuming that the bilingual corpus is sentence-aligned, and based on their experience with speech recognition, Brown et al. (1990) use the Expectation-Maximization (EM) algorithm to compute the most likely word alignments, allowing only 1:0, 1:1 and 1:2 couplings (notably not 0:n, or many : many).

4.2. Word alignment with the EM algorithm

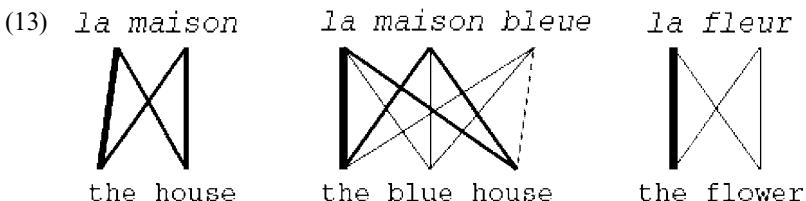
The EM algorithm (Dempster/Laird/Rubin 1977) is widely used in a variety of tasks involving incomplete data, where missing parameters are estimated, then these estimates are used to retrain the model, the process iterating until the best results are obtained. We can illustrate its use in word alignment for translation modelling by considering the examples in (11) (from Knight/Koehn 2004), which we assume to have been extracted from a larger corpus containing many other translation pairs.

- (11) *la maison* ↔ the house
la maison bleue ↔ the blue house
la fleur ↔ the flower

Initially, we assume that all word alignments are equally likely, as in (12), but a first pass shows that 3 of the 17 connections link *la* with *the*, and 2 out of 17 link *la* with *house*, *maison* with *house*, and *maison* with *the*.



A first iteration strengthens these more likely connections and at the same time weakens connections that are in conflict with them (13).



Further iterations strengthen connections such as the one between *fleur* and *flower*: because *la* is linked with *the*, *flower* is the only link open for *fleur* (14).



Eventually, there is convergence, and the inherent structure (15) is arrived at.



Obviously, the perfect alignment as seen in (15) is an ideal result: in practice, the resulting alignments are a set of probabilities, which reflect the alignments over the corpus. For example, besides the alignment of *la* with *the*, one could expect in a corpus that there would also be evidence for aligning *le* and *les* with *the*, and probabilities would reflect the relative strengths of these pieces of evidence.

4.3. The IBM Models

Brown et al. (1993) suggested a number of different ways in which their original (1990) basic model could be enhanced, in what have become known as “IBM Models” 1–5. In what follows we give a necessarily brief overview of the five models: for mathematical details readers are referred to the original source. The simplest, Model 1, assumes a uniform distribution, i.e. that the target-language word should occur in the place in the sequence corresponding to its place in the source-language sequence. Model 2 tries to model relative position in the word stream by calculating the probabilities of a certain position in the target string for each word given its position in the source string, and the lengths of the two strings: a word near the beginning of the source sentence is more likely to correspond to a word near the beginning of the target sentence, especially if the sentence is long. Model 3 includes fertility probabilities, as described above, and models distortion better. Model 4 additionally takes into account the fact that often words in the source language constitute a phrase which is translated as a unit in the target language. For example, in the translation pair in (16), *nodding* is associated with the phrase *faire signe que oui* in Model 4, while in Model 3 it is connected only to the words *signe* and *oui*.

- (16) *Il me semble faire signe que oui.*
It seems to me that he is nodding.

Finally, Model 5 rectifies a deficiency in Models 3 and 4 whereby words can be assigned to the same position in the target string, or to positions before or after the start or end of the target string.

Other researchers have typically taken one of models 1–3 as a starting point, and tried to develop strategies from there (see Och/Ney 2003).

Some alternatives to the word-based IBM models have emerged more recently: we will discuss these approaches in section 5.

4.4. The target language model

As mentioned above, the aim of the language model is to predict the most likely sequence of target-language words chosen by the translation model. To some extent, word-

sequence is determined by the translation model (especially the higher-numbered IBM models, and also more recent approaches to be discussed in section 5), but the language model is necessary to ensure that target-language-specific features, such as agreement and long-distance dependencies, not easily captured by the translation model, are covered.

The target-language model essentially models the probability of sequences of words. In principle, we could model the probability of a sequence of words w_1, w_2, \dots, w_m , by modelling the probability of each successive word given the preceding sequence $P(w_i | w_1, \dots, w_{i-1})$, so that the probability of the entire sequence would be (17).

$$(17) \quad P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1})$$

Unfortunately, this is impractical or even impossible, given the infinite nature of language and the problem of sparse data (see article 37). In practice, it turns out that the trade-off between practicality and usability comes if we look only at sequences of 3 or 4 words, referred to as n -grams with $n = 3$ or $n = 4$. The probability of a given string of words using a trigram model is given by (18).

$$(18) \quad P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_{i-2}, w_{i-1})$$

Probabilities for $i < 3$ are catered for by considering start-of-sentence as a pseudo-word. One problem with this model is again that of sparse data: if any of the trigrams happen not to occur in the training data, as is very likely, they will receive a 0 probability score, which will of course result in the product being 0. This is overcome in two ways: smoothing, and back-off. “Smoothing” consists of adjusting all the parameters so that none of them are 0. Crudely, one could add a tiny value to each parameter, but there are numerous other better motivated methods of smoothing (see Manning/Schütze 1999, 199 ff.; Jurafsky/Martin 2000, 206 ff.). “Back off” involves looking at $(n - 1)$ -gram models if the n -gram is unseen. So a trigram model would back off to include bigram and if necessary unigram statistics, as in (19),

$$(19) \quad \hat{P}(w_i | w_{i-2}, w_{i-1}) = \begin{cases} P(w_i | w_{i-2}, w_{i-1}) & \text{if } f(w_{i-2} w_{i-1} w) > 0 \\ \alpha P(w_i | w_{i-1}) & \text{if } f(w_{i-2} w_{i-1} w) = 0 \\ & \text{and } f(w_{i-1} w) > 0 \\ \beta P(w_i) & \text{otherwise} \end{cases}$$

where f is the frequency count of the n -gram sequence, and α, β are weights (see Manning/Schütze 1999, 219 ff.; Jurafsky/Martin 2000, 216ff.).

4.5. The decoder

To complete the SMT system we need a program which can apply the translation and language models to a given input text, that is to search for the target text which maxi-

mizes the probability equations. This part of the process has come to be known as the “decoder”. Evidently, given the number of statistical parameters, an exhaustive search of all possible combinations is impractical. Knight (1999) demonstrated that the problem was NP-complete. Various alternatives have been proposed for this problem.

The simplest is perhaps the “stack search” which basically starts by proposing a simple hypothesis, for example take the most probable word-by-word translation in the order of the source text, and explore the surrounding search space in a motivated way until the “score” cannot be further improved (Wang/Waibel 1997; Germann et al. 2001). A variant of this is a “beam search”. In this case the target text is built up left-to-right by expanding the translation hypotheses. Since this method is exponential in the length of the sentence, various tactics are needed to make the search more tractable. Pruning obviously weak hypotheses is a good start, for example eliminating texts where the number of words in the source and target texts are vastly different. Ueffing/Och/Ney (2002) used word graphs to maximise the efficiency of the beam search. Decoding viewed as state-space search to be tackled using methods based on Dynamic Programming is an approach taken by García-Varea/Casacuberta/Ney (1998), Niessen et al. (1998), Och/Ueffing/Ney (2001) and Tillman/Ney (2003). Tillmann et al. (1997) use an approach based on Hidden Markov Model alignments. Watanabe/Sumita (2003) present a method which uses some techniques borrowed from EBMT.

5. Variants of SMT

Early on in the history of SMT it was recognised that simple word-based models would only go so far in achieving a reasonable quality of translation. In particular, cases where single words in one language are translated as multi-word phrases in the other, and cases where the target-language syntax is significantly distorted with respect to the source language often cause bad translations in simple SMT models. Examples of these two phenomena are to be found when translating between German and English, as seen in (20)–(21) (from Knight/Koehn 2004).

- (20) a. *Zeitmangel erschwert das Problem.*
lit. Lack-of-time makes-more-difficult the problem
'Lack of time makes the problem more difficult.'
 - b. *Eine Diskussion erübriggt sich demnach.*
lit. A discussion makes-unnecessary itself therefore
'Therefore there is no point in discussion.'
- (21) a. *Das ist der Sache nicht angemessen.*
lit. That is to-the matter not appropriate
'That is not appropriate for this matter.'
 - b. *Den Vorschlag lehnt die Kommission ab.*
lit. The proposal rejects the Commission off
'The Commission rejects the proposal.'

To address these problems, variations of the SMT model have emerged which try to work with phrases rather than words, and with structure rather than strings. These approaches are described in the next two sections.

5.1. Phrase-based SMT

The idea behind “phrase-based SMT” is to enhance the conditional probabilities seen in the basic models with joint probabilities, i. e. “phrases”. Because the alignment is again purely statistical, the resulting phrases need not necessarily correspond to groupings that a linguist would identify as constituents.

Wang/Waibel (1998) proposed an alignment model based on shallow model structures. Since their translation model reordered phrases directly, it achieved higher accuracy for translation between languages with different word orders. Other researchers have explored the idea further (Och/Tillmann/Ney 1999; Marcu/Wong 2002; Koehn/Knight 2003; Koehn/Och/Marcu 2003).

Och/Ney's (2004) alignment template approach takes the context of words into account in the translation model, and local changes in word order from source to target language are learned explicitly. The model is described using a log-linear modelling approach, which is a generalization of the often used source–channel approach. This makes the model easier to extend than classical SMT systems. The system has performed well in evaluations.

To illustrate the general idea more exactly, let us consider (22) as an example (from Knight/Koehn 2004).

- (22) *Maria no daba una bofetada a la bruja verde.*

lit. Maria not gave a slap to the witch green

‘Maria did not slap the green witch.’

First, the word alignments are calculated in the usual way. Then potential phrases are extracted by taking word sequences which line up in both the English and Spanish, as in Figure 56.1.

If we take all sequences of contiguous alignments, this gives us possible phrase alignments as in (23) for which probabilities can be calculated based on the relative co-occurrence frequency of the pairings in the rest of the corpus.

Maria no daba una bofetada a la bruja verde								
Maria								
did		■						
not			■					
slap				■	■			
the						■	■	
green								■
witch							■	

Fig. 56.1: Initial phrasal alignment for example (22)

- (23) (Maria, *Maria*)
 (did not, *no*)
 (slap, *daba una bofetada*)
 (the, *a la*)
 (green, *verda*)
 (witch, *bruja*)

By the same principle, a further iteration can identify larger phrases, as long as the sequences are contiguous, as in Figure 56.2.

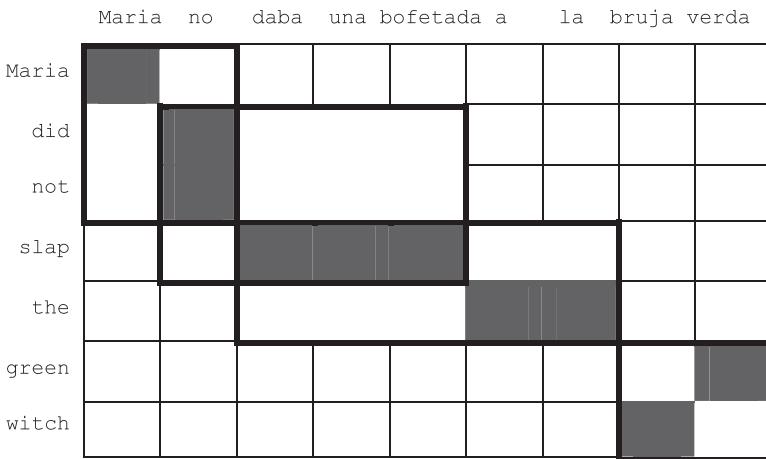


Fig. 56.2: Further phrasal identification

- (24) (Maria did not, *Maria no*)
 (did not slap, *no daba una bofetada*)
 (slap the, *daba una bofetada a la*)
 (green witch, *bruja verda*)

The process continues, each time combining contiguous sequences giving the phrases in (25), (26) and finally (27), the whole sentence.

- (25) (Maria did not slap, *Maria no daba una bofetada*)
 (did not slap the, *no daba una bofetada a la*)
 (the green witch, *a la bruja verda*)
- (26) (Maria did not slap the, *Maria no daba una bofetada a la*)
 (slap the green witch, *daba una bofetada a la bruja verda*)
- (27) (Maria did not slap the green witch, *Maria no daba una bofetada a la bruja verda*)

Of course, as the phrases get longer, the probabilities get smaller, as their frequency in the corpus diminishes.

Koehn/Och/Marcu (2003) evaluated a number of variants of the phrase-based SMT approach, and found that they all represented an improvement over the original word-

based approaches. Furthermore, increased corpus size had a more marked positive effect than it did with word-based models. The best results were obtained when the probabilities for the phrases were weighted to reflect lexical probabilities, i. e. scores for individual word-alignments. And, most interestingly, if phrases not corresponding to constituents in a traditional linguistic view were excluded, the results were not as good.

5.2. Structure-based SMT

Despite the improvements, a number of linguistic phenomena still prove troublesome, notably discontinuous phrases and long-distance reordering, as in (21). To try to handle these, the idea of “syntax-based SMT” or “structure-based SMT” has developed, benefiting from ideas from stochastic parsing and the use of treebanks (see articles 5, 13, 28).

Wu (1997) introduced Inversion Transduction Grammars as a grammar formalism to provide structural descriptions of two languages simultaneously, and thereby a mapping between them: crucially, his grammars of English and Cantonese were derived from the bilingual Hong Kong Hansard corpus. The development of an efficient decoder based on Dynamic Programming permits the formalism to be used for SMT (Wu/Wong 1998). Alshawi/Srinivas/Douglas (1998) developed a hierarchical transduction model based on finite-state transducers: using an automatically induced dependency structure, an initial head-word pair is chosen, and the sentence is then expanded by translating the dependent structures. In Yamada/Knight’s (2001) “tree-to-string” model a parser is used on the source text only. The tree is then subject to reordering, insertion and translation operations, all based on stochastic operations. Charniak/Knight/Yamada (2003) adapted this model with an entropy-based parser which enhanced the use made of syntactic information available to it. Gildea (2003) proposed a tree-to-tree alignment model in which subtree cloning was used to handle more reordering in parse trees. Dependency treebanks have been used for Czech–English SMT by Čmejrek/Cuřín/Havelka (2003). Och et al. (2004) present and evaluate a wide variety of add-ons to a basic SMT system.

Another treebank-based approach to MT is the Data-Oriented Translation approach of Poutsma (2000) and Hearne/Way (2003). The authors consider this approach to be EBMT rather than SMT, and one could argue that with SMT taking on a more phrase-based and syntax-based approach, while EBMT incorporates statistical measures of collocation and probability, the two approaches are quickly merging, a position argued by Way/Gough (2005).

5. Rapid development of MT for less-studied languages

An important attraction of corpus-based MT techniques is the possibility that they can be used to quickly develop MT systems for less-studied languages (cf. article 21), inasmuch as these MT techniques require only bilingual corpora and appropriate tools for alignment, extraction of linguistic data and so on. It must be said that some of the latest ideas, particularly in SMT, requiring treebanks and parsers make this less relevant. Nevertheless, empirical methods do seem to embody the best hope for resourcing under-resourced languages.

The first such attempt to demonstrate the feasibility of this was at the Johns Hopkins Summer Workshop in 1999, when students built a Chinese–English SMT system in one day (Al-Onizan et al. 1999). Although Chinese is not a less-studied language as such, it is of interest because English and Chinese are typologically quite dissimilar. The corpus used was the 7-million-word “Hong Kong Laws” corpus and the system was built using the EGYPT SMT toolkit developed at the same workshop and now generally available online.

Germann (2001) tried similar techniques with rapidly developed resources, building a Tamil–English MT system by manually translating 24,000 words of Tamil into English in a six-week period. Weerasinghe (2002) worked on Sinhala–English using a 50,000-word corpus from the World Socialist Web Site. Oard/Och (2003) built a system to translate between English and the Philippine language Cebuano, based on 1.3 m words of parallel text collected from five sources (including Bible translations and on-line and hard-copy newsletters). Foster et al. (2003) describe a number of difficulties in their attempt to build a Chinese–English MT system in this way.

7. Conclusions

MT is often described as the historically original task of Natural Language Processing, as well as the archetypical task in that it has a bit of everything, indeed in several languages; so it is no surprise that corpora – or at least collections of texts – have played a significant role in the history of MT. However, it is only in the last 10–15 years that they have really come to the fore with the emergence and now predominance of corpus-based techniques for MT. This article has reviewed that history, from “reference corpora” in the days of rule-based MT via corpus-based translators’ tools, to MT methods based exclusively on corpus information. Many of the tools developed for corpus exploitation and described in other articles in this book have had their genesis in MT, and research in corpus-based MT is certainly at the forefront of computational linguistics at the moment.

8. Acknowledgments

I would like to thank an anonymous reviewer for their very helpful comments on earlier drafts of this article. I would also like to thank Andy Way for his advice and suggestions on several sections of this article. All errors and infelicities remain of course my own responsibility.

9. Literature

- Al-Onizan, Y./Curin, J./Jahr, M./Knight, K./Lafferty, J./Melamed, D./Och, F. J./Purdy, D./Smith, N. A./Yarowsky, D. (1999), *Statistical Machine Translation: Final Report, JHU Workshop 1999*. Technical report, Johns Hopkins University, Baltimore, MD. Available at http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps [accessed 7 June 2005].

- Alshawi, H./Srinivas, B./Douglas, S. (1998), Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Montreal, Canada, 41–47.
- Arthern, P. J. (1979), Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint. In: Snell, B. M. (ed.), *Translating and the Computer: Proceedings of a Seminar, London, 14th November 1978*. Amsterdam: North Holland, 77–108.
- Barlow, M. (1995), ParaConc: A Concordancer for Parallel Texts. In: *Computers and Texts* 10, 14–16.
- Bowker, L. (2002), *Computer-aided Translation Technology. A Practical Introduction*. Ottawa: University of Ottawa Press.
- Brown, P. F./Cocke, J./Della Pietra, S. A./Della Pietra, V. J./Jelinek, F./Lafferty, J. D./Mercer, R. L./Roossin P. S. (1990), A Statistical Approach to Machine Translation. In: *Computational Linguistics* 16, 79–85. Reprinted in Nirenburg/Somers/Wilks 2003, 355–362.
- Brown, P. F./Della Pietra, S. A./Della Pietra, V. J./Mercer, R. L. (1993), The Mathematics of Statistical Machine Translation: Parameter Estimation. In: *Computational Linguistics* 19, 263–311.
- Brown, R. D. (2000), Automated Generalization of Translation Examples. In: *Proceedings of the 18th International Conference on Computational Linguistics, Coling 2000 in Europe*. Saarbrücken, Germany, 125–131.
- Brown, R. D. (2001), Transfer-rule Induction for Example-based Translation. In: *MT Summit VIII Workshop on Example-based Machine Translation*. Santiago de Compostela, Spain, 1–11.
- Carl, M./Way, A. (eds.) (2003), *Recent Advances in Example-based Machine Translation*. Dordrecht: Kluwer Academic Press.
- Carl, M./Way, A. (eds.) (2006–2007), *Example-based Machine Translation. Machine Translation* 19(3–4) and 20(1) (Special Issue).
- Charniak, E./Knight, K./Yamada, K. (2003), Syntax-based Language Models for Statistical Machine Translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, 40–46.
- Church, K. W./Gale, W. A. (1991), Concordances for Parallel Texts. In: *Using Corpora, Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research*. Oxford, United Kingdom, 40–62.
- Cicekli, I. (2006), Inducing Translation Templates with Type Constraints. In: *Machine Translation* 19, 281–297.
- Cicekli, I./Güvenir, H. A. (1996), Learning Translation Rules from a Bilingual Corpus. In: *NeM-Lap-2: Proceedings of the Second International Conference on New Methods in Language Processing*. Ankara, Turkey, 90–97.
- Cicekli, I./Güvenir, H. A. (2003), Learning Translation Templates from Bilingual Translation Examples. In: Carl/Way 2003, 255–286.
- Čmejrek, M./Cuřín, J./Havelka, J. (2003), Treebanks in Machine Translation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Växjö, Sweden, 209–212.
- Collins, B. (1998), Example-based Machine Translation: An Adaptation-guided Retrieval Approach. PhD thesis, Trinity College, Dublin.
- Cranias, L./Papageorgiou, H./Piperidis, S. (1997), Example Retrieval from a Translation Memory. In: *Natural Language Engineering* 3, 255–277.
- Dempster, A. P./Laird, N. M./Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. In: *Journal of the Royal Statistical Society Series B* 39(1), 1–38.
- Foster, G./Gandrabur, S./Langlais, P./Plamondon, P./Russel, G./Simard, M. (2003), Statistical Machine Translation: Rapid Development with Limited Resources. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, 110–117.
- Foster, G./Langlais, P./Lapalme, G. (2002), User-friendly Text Prediction for Translators. In: *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Philadelphia, PA, 148–155.

- García-Varea, I./Casacuberta, F./Ney, H. (1998), An Iterative DP-based Search Algorithm for Statistical Machine Translation. In: *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP 98)*. Sydney, Australia, 1135–1139.
- Gaußier, E./Langé, J.-M./Meunier, F. (1992), Towards Bilingual Terminology. In: *Proceedings of the Joint ALLC/ACH Conference*. Oxford, United Kingdom, 121–124.
- Germann, U. (2001), Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect? In: *ACL-2001 Workshop on Data-driven Methods in Machine Translation*. Toulouse, France, 1–8.
- Germann, U./Jahr, M./Knight, K./Marcu, D./Yamada, K. (2001), Fast Decoding and Optimal Decoding for Machine Translation. In: *Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter*. Toulouse, France, 228–235.
- Gildea, D. (2003), Loosely Tree-based Alignment for Machine Translation. In: *41st Annual Conference of the Association for Computational Linguistics*. Sapporo, Japan, 80–87.
- Harris, B. (1988), Bi-text, a New Concept in Translation Theory. In: *Language Monthly* 54, 8–10.
- Hearne, M./Way, A. (2003), Seeing the Wood for the Trees: Data-oriented Translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, 165–172.
- Isabelle, P. (1992a) Préface – Preface. In: *Quatrième colloque international sur les aspects théoriques et méthodologiques de la traduction automatique. Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-92*. Montréal, Canada, iii.
- Isabelle, P. (1992b) Bi-textual Aids for Translators. In: *Screening Words: User Interfaces for Text. Proceedings of the 8th Annual Conference of the UW Centre for the New OED and Text Research*. Waterloo, Ont. Available at http://rali.iro.umontreal.ca/Publications/urls/bi_textual_aids.ps.
- Jurafsky, D./Martin, J. H. (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.
- Kay, M. (1980), The Proper Place of Men and Machines in Language Translation. Research Report CSL-80-11, Xerox PARC, Palo Alto, CA. Reprinted in: *Machine Translation* 12 (1997), 3–23; and in: Nirenburg/Somers/Wilks 2003, 221–232.
- King, G. W. (1956), Stochastic Methods of Mechanical Translation. In: *Mechanical Translation* 3(2), 38–39. Reprinted in: Nirenburg/Somers/Wilks 2003, 37–38.
- Knight, K. (1999), Decoding Complexity in Word-replacement Translation Models. In: *Computational Linguistics* 25, 607–615.
- Knight, K./Koehn, P. (2004), What's New in Statistical Machine Translation? Tutorial at *HLT-NAACL 2004, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Edmonton, Alberta, Canada. Available from: <http://www.iccs.inf.ed.ac.uk/~pkoechn/publications/tutorial2003.pdf>.
- Koehn, P./Knight, K. (2003), Feature-rich Statistical Translation of Noun Phrases. In: *41st Annual Conference of the Association for Computational Linguistics*. Sapporo, Japan, 311–318.
- Koehn, P./Och, F. J./Marcu, D. (2003), Statistical Phrase-based Translation. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Edmonton, Alberta, 127–133.
- Lehrberger, J. (1982), Automatic Translation and the Concept of Sublanguage. In: Kittredge, R. I./Lehrberger, J. (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: Mouton de Gruyter, 81–106. Reprinted in Nirenburg/Somers/Wilks 2003, 207–220.
- Macdonald, K. (2001), Improving Automatic Alignment for Translation Memory Creation. In: *Translating and the Computer 23: Proceedings from the Aslib Conference*. London [pages not numbered].
- Macklovitch, E./Russell, G. (2000), What's been Forgotten in Translation Memory. In: White, J. S. (ed.), *Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico*. Berlin: Springer, 137–146.
- Manning, C. D./Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

- Marcu, D./Wong, W. (2002), A Phrase-based, Joint Probability Model for Statistical Machine Translation. In: *Conference on Empirical Methods for Natural Language Processing (EMNLP 2002)*. Philadelphia, PA, 133–139.
- McTait, K./Trujillo, A. (1999), A Language-neutral Sparse-data Algorithm for Extracting Translation Patterns. In: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*. Chester, England, 98–108.
- Melby, A. (1982), A Bilingual Concordance System and its Use in Linguistic Studies. In: Gutwinski, W./Jolly, G. (eds.), *LACUS 8: The 8th Lacus Forum, Glendon College, York University, Canada, August 1981*. Columbia, SC: Hornbeam Press, 541–554.
- Nagao, M. (1984), A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In: Elithorn, A./Banerji, R. (eds.), *Artificial and Human Intelligence*. Amsterdam: North-Holland Publishing Company, 173–180. Reprinted in Nirenburg/Somers/Wilks 2003, 351–354.
- Niessen, S./Vogel, S./Ney, H./Tillmann, C. (1998), A DP-based Search Algorithm for Statistical Machine Translation. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Montreal, Canada, 960–967.
- Nirenburg, S./Domashnev, C./Grannes, D. J. (1993), Two Approaches to Matching in Example-based Machine Translation. In: *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation TMI '93: MT in the Next Generation*. Kyoto, Japan, 47–57.
- Nirenburg, S./Somers, H./Wilks, Y. (eds.) (2003), *Readings in Machine Translation*. Cambridge, MA: MIT Press.
- Oard, D. W./Och, F. J. (2003), Rapid-response Machine Translation for Unexpected Languages. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, 277–283.
- Och, F. J./Gildea, D./Khudanpur, S./Sarkar, A./Yamada, K./Fraser, A./Kumar, S./Shen, L./Smith, D./Eng, K./Jain, V./Jin, Z./Radev, D. (2004), A Smorgasbord of Features for Statistical Machine Translation. In: *Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Boston, MA, 161–168.
- Och, F. J./Ney, H. (2003), A Systematic Comparison of Various Statistical Alignment Models. In: *Computational Linguistics* 29, 19–51.
- Och, F. J./Ney, H. (2004), The Alignment Template Approach to Statistical Machine Translation. In: *Computational Linguistics* 30, 417–449.
- Och, F. J./Tillmann, C./Ney, H. (1999), Improved Alignment Models for Statistical Machine Translation. In: *Proceedings of the 1999 Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, MD, 20–28.
- Och, F. J./Ueffing, N./Ney, H. (2001), An Efficient A* Search Algorithm for Statistical Machine Translation. In: *Proceedings of the Data-driven Machine Translation Workshop, 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 55–62.
- Planas, E./Furuse, O. (1999), Formalizing Translation Memories. In: *Machine Translation Summit VII*, Singapore, 331–330. Reprinted in Carl/Way 2003, 157–188.
- Poutsma, A. (2000), Data-oriented Parsing. In: *COLING 2000 in Europe: The 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, 635–641.
- Rapp, R. (2002), A Part-of-speech-based Search Algorithm for Translation Memories. In: *LREC 2002, Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain, 466–472.
- Romary, L./Mehl, N./Woolfs, D. (1995), The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purposes. In: *Text Technology* 5, 206–220.
- Sato, S./Nagao, M. (1990), Toward Memory-based Translation. In: *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*. Volume 3. Helsinki, Finland, 247–252.

- Simard, M./Foster, G./Perrault, F. (1993), TransSearch: A Bilingual Concordance Tool. Industry Canada Centre for Information Technology Innovation (CITI), Laval, Canada, October 1993. Available at <http://rali.iro.umontreal.ca/Publications/urls/sfpTS93e.ps>.
- Somers, H. (2003), Translation Memory Systems. In: Somers, H. (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam: Benjamins, 31–47.
- Somers, H./Fernández Díaz, G. (2004), Translation Memory vs. Example-based MT: What is the Difference? In: *International Journal of Translation* 16(2), 5–33. Based on: Diferencias e interconexiones existentes entre los sistemas de memorias de traducción y la EBMT. In: Corpas Pastor, G./Varela Salinas, M.-J. (eds.), *Entornos informáticos de la traducción profesional: las memorias de traducción*. Granada: Editorial Atrio (2003), 167–192.
- Somers, H./Tsujii, J./Jones, D. (1990), Machine Translation without a Source Text. In: *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*. Volume 3. Helsinki, Finland, 271–276. Reprinted in Nirenburg/Somers/Wilks 2003, 401–406.
- Sumita, E./Iida, H./Kohyama, H. (1990), Translating with Examples: A New Approach to Machine Translation. In: *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*. Austin, Texas, 203–212.
- Tillmann, C./Ney, H. (2003), Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. In: *Computational Linguistics* 29, 97–133.
- Tillmann, C./Vogel S./Ney, H./Sawaf, H. (2000), Statistical Translation of Text and Speech: First Results with the RWTH System. In: *Machine Translation* 15, 43–74.
- Tillmann, C./Vogel S./Ney, H./Zubiaga, A. (1997), A DP-based Search using Monotone Alignments in Statistical Translation. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 289–296.
- Ueffing, N./Och, F. J./Ney, H. (2002), Generation of Word Graphs in Statistical Machine Translation. In: *Conference on Empirical Methods for Natural Language Processing (EMNLP 2002)*. Philadelphia, PA, 156–163.
- Wang, Y.-Y./Waibel, A. (1997), Decoding Algorithm in Statistical Machine Translation. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 366–372.
- Wang, Y.-Y./Waibel, A. (1998), Modeling with Structures in Statistical Machine Translation. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Montreal, Canada, 1357–1363.
- Watanabe, T./Sumita, E. (2003), Example-based Decoding for Statistical Machine Translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, 410–417.
- Way, A./Gough, N. (2005), Comparing Example-based and Statistical Machine Translation. In: *Journal of Natural Language Engineering* 11, 295–309.
- Weerasinghe, R. (2002), Bootstrapping the Lexicon Building Process for Machine Translation between ‘New’ Languages. In: Richardson, S. D. (ed.), *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA2002, Tiburon, CA*. Berlin: Springer, 177–186.
- Wu, D. (1997), Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. In: *Computational Linguistics* 23, 377–403.
- Wu, D./Wong, H. (1998), Machine Translation with a Stochastic Grammatical Channel. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Montreal, Canada, 1408–1414.
- Yamada, K./Knight, K. (2001), A Syntax-based Statistical Translation Model. In: *Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter*. Toulouse, France, 523–530.

57. Corpus linguistics and first language acquisition

1. Introduction
2. The Child Language Exchange System
3. Sample studies
4. Conclusion
5. Literature

1. Introduction

Child language looks very different from adult language. The first meaningful utterances children produce appear a few months after the first birthday; they consist of single words and holophrases, i. e. unanalyzed chunks of frequent word combinations such as *What's that* and *All gone*. In the course of the development, children's utterances become increasingly more complex and variable, but the process of language acquisition continues well into the school years (cf. Clark 2003; for an overview of the field). In order to study this process, child language researchers use two major methods: They test children's linguistic knowledge in experiments and examine naturally occurring child language (cf. McDaniel/McKee/Smith Cairns 1996). When child language is transcribed and compiled in a computerized database, it can be seen as a linguistic corpus. Child language researchers of all theoretical persuasions make common use of corpus data to investigate the development of children's linguistic knowledge.

This paper provides an overview of corpus-based approaches to the study of language acquisition. It is divided into three parts. The first part describes the particular properties of a child language corpus and provides a short summary of the history of corpus-based approaches to the study of language acquisition; the second part describes the Child Language Exchange System, short CHILDES, which includes the largest and most widely used collection of child language corpora and a set of research tools to work with these corpora; and the third part discusses some recent sample studies on syntactic development that have drawn on data from the CHILDES database.

1.1. Features of a child language corpus

The most dramatic developments of language acquisition occur during the early pre-school years, which have been in the focus of child language research. Since preschool children are not able to write, child language corpora are usually limited to spoken data, taken from spontaneous conversations between the child and an adult speaker (see article 11). Since the parents are the most important speech partners of a preschool child, most child language corpora consist of child-parent conversations recorded in natural settings at home.

Compared to adult corpora, child language corpora are relatively homogeneous. In contrast to adult speakers, who master a wide variety of different genres and speech registers, children use language in a limited range of social settings. However, there are two dimensions of variation that play an important role in the study of language acquisi-

tion and the organization of a child language corpus. First, since children's linguistic performance changes with age, most child language studies use data from children at different ages. Second, since there can be significant differences in the development of individual children (cf. Lieven/Pine/Barnes 1992; Bates/Dale/Thal 1994), most child language studies draw on data from several children. The data can be collected in two different ways (see article 9). In longitudinal studies researchers track the development of individual children over an extended period of time, and in cross-sectional studies they collect data from multiple children of different ages all at one time.

Apart from individual differences and age-related changes, there are various other factors that must be taken into account in a corpus-based study of language acquisition. To begin with, one factor that can influence the outcome of a child language study is the socio-economic background of the child. There are often important linguistic differences between children of different economic classes, different ethnic backgrounds, and different cultural communities (cf. Hoff-Ginsberg 1991; Lawrence/Shipley 1996). Another important factor is the sex of the child; there is evidence that boys and girls acquire language somewhat differently (cf. Karrass et al. 2002). Furthermore, the overall amount of the data and the time intervals at which the data are collected are essential to a child language study. How much data do we need to investigate the development of a particular phenomenon? Specifically, how much data do we need to determine the age of appearance, the order of acquisition, and the developmental pathway?

Until recently, most researchers have collected their data in one-hour recordings prepared at regular intervals of one to three weeks. Using this sampling method, the transcriptions represent between 1.0% and 1.5% of the language children hear and produce during the time of the study (cf. Tomasello/Stahl 2004). While 1.5% of the data may provide enough information to trace the development of frequently occurring structures and expressions, it is not sufficient to study the development of less frequent phenomena. As Tomasello/Stahl (2004) have shown, in order to trace the development of infrequent grammatical patterns one needs to examine a much larger proportion of children's speech. In particular, the age of appearance and the order of acquisition are not reliable if we use the standard language samples of one-hour recordings collected once or twice a month; rather, what we need are much denser data collections as they are currently prepared for the first time at the Max Planck Institute for Evolutionary Anthropology in Leipzig (Tomasello/Stahl 2004). These 'dense corpora' consist of one-hour recordings collected five times a week during the early preschool years. They contain about 10% of the language a child produces and hears, making it possible to trace even the development of infrequent linguistic phenomena in detail.

1.2. From diary studies to the study of large computerized corpora

The study of spontaneous child language has a long history. The first investigations appeared more than a hundred years ago; they were parental diary studies tracing the linguistic development of a particular child based on the parents' observations. One of the earliest and most frequently cited diary studies was a study by Clara and William Stern (1907). The Sterns carefully documented the speech of their two German-speaking children, Hilde and Gunter, and proposed the first widely accepted stages of child lan-

guage development. Another early diary study that had an important impact on the field was an investigation by Leopold (1939–1949), who described the linguistic development of his daughter Hildegard from birth to age two.

In the 1940s and 1950s, behaviourist psychologists began to collect child language data more systematically. In contrast to the early diary studies, the behaviourists worked primarily with cross-sectional data; that is, rather than studying the development of one child in detail, they collected data from a large number of children using small language samples of about an hour of speech. Some of these cross-sectional investigations included data from more than 100 hundred children carefully selected based on their ages, sexes, and socio-economic backgrounds.

Starting in the 1960s, the first systematic longitudinal studies appeared (cf. Braine 1963; Brown 1968; Bellugi 1967; Bloom 1970). Like the early diary studies, these studies traced the development of a particular child (or a few children) over an extended period of time. However, in contrast to the early diary studies, these studies did not use parental observations as their primary source of data; rather, they used a new technology to collect data. For the first time in history, child language was first audio-taped and then carefully transcribed and annotated by the researcher. While parental diary data is fairly reliable for investigating children's early one- and two-word utterances, it is impossible to keep a systematic diary of more advanced children. The use of audio tapes allowed the researcher to systematically document and analyze more complex linguistic phenomena that emerge only after the two-word stage. The studies that appeared at this time include the pioneering works by Martin Braine (1963, 1976), Lois Bloom (1970), and Roger Brown (1973), which have become classics of child language research.

During the 1970s and 1980s, the number of longitudinal studies increased steadily; however, the data collected were only available to a small number of researchers, usually the researchers who had collected the data and their colleagues. Advances in computer technology made it possible to share child language data more easily. Starting in the early 1980s, Brian MacWhinney and Catherine Snow established a large computerized database, which they called the Child Language Exchange System, or short CHILDES (cf. MacWhinney/Snow 1985). The CHILDES database is the largest and most widely used collection of child language to date; it includes data from more than a hundred research projects from a variety of languages. MacWhinney and Snow collected these data and developed a number of research tools to facilitate their analysis. Among other things, they developed a coding system to standardize the transcriptions and a set of computer programs to search and manipulate the database. Both the database and the computer programs are freely available on the internet. Over the years, CHILDES has grown to a whole research system, providing not only a comprehensive database and computer programs but also access to an electronic bibliography, a mailing list, and various other research tools (<http://childe.s.psy.cmu.edu/>). Since CHILDES has become the standard corpus environment for child language data, the next section describes the CHILDES database and some of the CHILDES research tools in more detail.

2. The Child Language Exchange System

CHILDES consists of three major components: (1) the database, (2) a system of notations and codings called CHAT, and (3) a set of computer programs called CLAN. The three components will be discussed in turn (cf. MacWhinney 1995, 1996, 2000).

2.1. The database

The database includes transcripts from over a hundred research projects from more than 20 languages (cf. MacWhinney 2000, 280–420). Apart from English, there are at present data from Cantonese, Danish, Dutch, French, German, Greek, Hebrew, Hungarian, Italian, Japanese, Mambila, Mandarin, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Turkish, and Ukrainian. For some of these languages, the database includes only a few transcripts from one child, but for other languages it includes comprehensive speech samples from several children. However, the bulk of the data come from English, both American English and British English; together they account for more than 50% of the data included in the CHILDES database. Most of the data are based on longitudinal studies of normally developing monolingual children, but there are also data from cross-sectional studies and data from children growing up bilingual or with language disorders (e.g. Down syndrome, autism, SLI = Specific Language Impairment) (Conti-Ramsden/Dykins 1991; Chapman 1994; Joseph/Serratrice/Conti-Ramsden 2002). In addition, there are some data from adult aphasics and second language learners (de Houwer 1990; Yip/Matthews 2000). The longitudinal data consist of a set of transcripts of naturally occurring speech of individual children that were prepared at regular intervals over an extended period of time. Most studies are based on data of one-hour recordings that were collected once or twice a month during the preschool years. Table 57.1 shows the subdirectory of the data files of one child from the English corpus of the CHILDES database. The subdirectory includes 20 files of child-parent conversations that were recorded over a period of 14 months between the ages of 1;9 and 3;1.

Tab. 57.1: Subdirectory of Peter's data files (MacWhinney 2000, 292)

File	Age	File	Age	File	Age	File	Age
01	1; 9.7	06	2;0.7	11	2;3.21	16	2; 7.14
02	1; 9.21	07	2;0.7	12	2;4.14	17	2; 8.14
03	1;10.15	08	2;1.21	13	2;5.0	18	2; 9.14
04	1;11.7	09	2;2.14	14	2;5.21	19	2;10.21
05	1;11.21	10	2;3.0	15	2;6.14	20	3; 1.21

Some of the recent CHILDES transcripts are accompanied by digitized audio and video files, but the majority of the corpora, notably the older corpora, consist of transcripts without audio or video (see articles 12, 31). Some of the text files are based on data that were collected in the 1960s and 1970s and only later computerized to be included in the CHILDES database, but most of the data come from more recent studies. Since MacWhinney and Snow launched the CHILDES project, it has become common practice that researchers who have collected new child language data contribute their data to the CHILDES database after a few years. As a result, the CHILDES database has been continuously growing over the past 20 years and is likely to grow in the future.

2.2. CHAT

In order to standardize the transcripts of the CHILDES database, MacWhinney/Snow (1985) developed a system of notations called CHAT (cf. MacWhinney 2000, 5–139).

This system accommodates a standard level of analysis, but is flexible enough that researchers can extend it to investigate particular aspects of children's speech. Every CHAT file includes a header providing general information about the recording and the transcript (cf. 1).

(1) @Begin
 @Languages: en
 @Participants: CHI Nina Target_Child, MOT Mother, LIN Linda Aunt, DAV David Uncle
 @ID:
 english|suppes|CHI|2;0.17||normal||Target_Child||
 @ID:
 english|suppes|MOT||||Mother||
 @ID:
 english|suppes|LIN||||Aunt||
 @ID:
 english|suppes|DAV||||Uncle||
 @Date: 06-DEC-1970
 @Time Duration: 11:00–12:00
 @Situation: linda and david are visiting nina and her mother. linda is nina's mother's younger sister. nina was very excited about linda's visit. linda and david arrived the previous evening. the last part of this tape was erased.

As can be seen in this example, the header includes information about the language of the conversation, the people who participated in the conversation, the researcher who conducted the study, the age of the child, the date of the recording, the duration of the recording, and some general information about the situational context. Except for the participants, all of the information given in the header is optional, but in most transcripts the header provides detailed information about the recording, the situational context, and the child.

The transcriptions are written in standard orthography but include diacritics for various aspects of the conversation. For instance, there are diacritics for incomplete words (cf. 2), retractions (cf. 3), false starts (cf. 4), and many other aspects that are relevant to the interpretation of an utterance (e.g. omissions, corrections, interruptions, repairs, errors, overlapping speech).

- (2) *CHI: I been sit(ting) all day .
- (3) *CHI: <I wanted> / I wanted to invite Margie .
- (4) *CHI: <I wanted> [/–] uh when is Margie coming ?

Moreover, the transcript may include diacritics for particular prosodic features such as syllable length (':', cf. 5), pauses ('#' cf. 6), and intonation contours ('–', cf. 7).

- (5) *CHI: baby want bana:nas ?
- (6) *CHI: I don't # know .
- (7) *CHI: well – .

While these diacritics are directly inserted into the utterance, CHAT also provides the possibility to annotate an utterance on several ‘dependent tiers’, which accompany the ‘utterance tier’. The dependent tiers provide information about parts of speech, morphological structures, intonation, speech errors, speech acts, paralinguistic behaviours, the situational context, and various other aspects of the utterance or the communicative context. The dependent tiers are optional, but at least all of the English data are accompanied by the %mor tier providing information about parts of speech and morphological structure (cf. 8) (see article 24).

- (8) *CHI: Mommy # we have to do something # lock him up .
 %mor: n:prop|Mommy pro|we v:aux|have inf|to v|do pro:indef|something
 v|lock pro|him prep|up .
 *MOT: he won't bother you .
 %mor: pro|he v:aux|will~neg|not v|bother pro|you .
 *MOT: he's all tired out .
 %mor: pro|he~v|be&3S qn|all adj|tired prep|out .
 *CHI: Mommy +...
 %mor: n:prop|Mommy +...
 *MOT: his name is Rinny from Rin_tin_tin .
 %mor: pro:poss:det|his n|name v|be&3S n:prop|Rinny prep|from
 n:prop|Rin_tin_tin

For every word on the utterance tier, the %mor tier includes a speech tag consisting of the word’s stem and a label for its categorical status. For instance, in example (8) the first utterance includes the word *have*, which occurs with the speech tag ‘v:aux|have’ indicating that *have* is an auxiliary verb in this case. If the word is morphologically complex, affixes and clitics are indicated separately and marked by diacritics. For instance, in the third utterance *won’t* occurs with the following speech tag ‘v:aux|will~neg|not’, which indicates that *won’t* consists of the auxiliary *will* and a cliticized form of the negative particle *not*.

Apart from the %mor tier, there are various other dependent tiers that can accompany an utterance. For instance, some transcripts include a dependent tier for speech act coding. As can be seen in examples (9) and (10), the speech act tier may provide information about the illocutionary force (e.g. \$IMP = imperative), the occurrence of imitations (e.g. \$IMIT = imitation), and other pragmatic aspects of the utterance.

- (9) *CHI: write dat piece a paper .
 %spa: \$IMP
 (10) *FAT: there we go .
 *CHI: there we go # Cromer .
 %spa: \$IMIT

Another important dependent tier is the error tier providing information about children’s errors, which may concern all levels of an utterance: the pronunciation, the choice of lexical expressions, and the morphosyntactic structure. CHILDES provides a set of stan-

dardized error codes (e.g. \$LEX = lexical error, \$BLN = blend) to indicate the type of error. The precise location of the error is marked on the utterance tier by a star [*] after the erroneous element (cf. 11–12).

- (11) *CHI: it has a very nice flate [*].
%err: flaste /fleɪst/ = taste /teɪst/ \$=flavour, taste \$LEX \$BLN ;
- (12) *CHI: it tooked [*] a while.
%err: tooked = took \$MOR \$SUF \$NFL \$REG \$FUK ; ;

A few corpora include a phonetic transcription represented on the %pho tier (see article 30). The phonetic transcription is based on the International Phonetic Alphabet, short IPA, but is written in ASCII characters. The ASCII translation of the IPA system is called UNIBET. Instead of UNIBET, some transcripts include a phonetic transcription in PHONASCII providing more detailed phonetic information. UNIBET transcriptions of individual words may also be included on the utterance tier, but longer UNIBET transcriptions and all transcriptions in PHONASCII are represented on the %pho tier (cf. 13–16).

- | | | |
|------|--|-----------|
| (13) | *CHI: I don't wanna go . | UNIBET |
| | %pho: 'ai don't "wan6 #go . | |
| (14) | *CHI: you better believe it . | UNIBET |
| | %pho: #yu "bEt3r b6!liv It . | |
| (15) | *CHI: hammer . | PHONASCII |
| | %pho: h ae,n,> m_b, os 3r | |
| (16) | *CHI: don't believe him . | PHONASCII |
| | %pho: d o,n, - ? b 1,\$ i, : v i-, n m | |

2.3. CLAN

The CLAN computer programs have been specifically designed to analyze the transcripts in the CHILDES database. They are freely available on the internet and can be installed on both MS-DOS operating systems and on Macintosh. Some of the CLAN programs serve to build new corpora, but most of them have been developed to work with child language transcripts in the CHAT format. In this section, I concentrate on some of the latter (cf. MacWhinney 2000, 130–279).

One of the most powerful CLAN programs is COMBO. It allows the user to conduct Boolean searches on the utterance tier and on any of the dependent tiers. COMBO provides the possibility to search for individual words, word fragments, word strings, affixes, parts of speech, morphological categories, speech act types, particular types of errors, and various other types of information provided in a CHAT transcript. The output of a COMBO search consists of a list of utterances that can be manipulated in various ways; (17) shows the output of a COMBO search.

(17) combo +sbecause +t*CHI +d2 adam1*

Wed Sep 21 10:19:40 2005

combo (11-Aug-2005) is conducting analyses on:

ONLY speaker main tiers matching: *CHI;

From file <adam10.cha>
 From file <adam11.cha>
 From file <adam12.cha>
 *CHI: because nail in (th)ere .
 From file <adam13.cha>
 From file <adam14.cha>
 From file <adam15.cha>
 From file <adam16.cha>
 *CHI: (be)cause [?] up air .
 *CHI: because Bobo@f crying .
 From file <adam17.cha>
 *CHI: because # Shadow [/] Shadow_Gay # heard [?] me .
 *CHI: because # Urs(ula broken .
 From file <adam18.cha>
 *CHI: because # Shadow_Gay hug me .
 *CHI: because Mommy hug [?] me .
 From file <adam19.cha>

The first line shows the search command, the second line exhibits the day and time at which the search was conducted, and the two following lines indicate at what tiers COMBO searched for the target string. In this case, the target string was the word *because* in all child utterances that occurred in files 10 to 19. The rest of the output consists of the target utterances and the names of the files that have been searched. The target utterances can also be shown together with their dependent tiers and in their discourse context (cf. MacWhinney 2000, 180–186).

Another powerful CLAN program is FREQ, which conducts frequency counts of words in specified files. The output indicates the number of tokens of each word included in the search files and calculates the type-token ratio of words produced by selected speakers. The type-token ratio is a rough measure of lexical diversity; it is based on the total number of all word types divided by the total number of all instances, i. e. all tokens, of these words. An example of a FREQ frequency count is given in (18).

(18) freq adam01.cha +t*CHI +o

Wed Sep 21 11:18:20 2005

freq (11-Aug-2005) is conducting analyses on:

ONLY speaker main tiers matching: *CHI;

From file <adam01.cha>
 110 who
 95 dat
 81 my

76 mommy

74 no

...

1 window

1 wipe

1 work

1 yours

357 Total number of different word types used

2669 Total number of words (tokens)

0.134 Type/Token ratio

A CLAN program that has been specifically designed to investigate child language data is the MLU program. MLU stands for Mean Length of Utterance, which is a measure, suggested by Roger Brown (1973), to determine the child's linguistic knowledge. The MLU program computes the average number of morphemes that a child produces in an utterance at a particular age. Since there can be significant differences in the linguistic development of individual children, the MLU is a more reliable measure for the child's linguistic proficiency than age. In example (19) the MLU program calculated an average number of 2.098 morphemes per utterance in the first file of a 2;3 year old English-speaking child.

(19) mlu adam01.cha +t*CHI

Wed Sep 21 11:33:57 2005

mlu (11-Aug-2005) is conducting analyses on:

ONLY speaker main tiers matching: *CHI;

From file t<adam01.cha>

MLU for Speaker: *CHI

MLU (xxx and yyy are EXCLUDED from the utterance and morpheme counts):

Number of: utterances = 1239, morphemes = 2599

Ratio of morphemes over utterances = 2.098

Standard deviation = 1.036

3. Sample studies

Observational studies have always played an important role in the study of language acquisition, but with the advance of new computer technologies and the development of the CHILDES system, observational studies, i.e. corpus-based studies, have become even more important. There are literally thousands of published research papers on first and second language acquisition, language disorders, and aphasia that have drawn on data from CHILDES (which is the only publicly available database of child language; cf. MacWhinney 1996, 5). Today, even many experimental studies begin with an explorative examination of corpus data to motivate the hypotheses tested in the experiments. To be

sure, corpus linguistics cannot replace the role of experimental studies in child language research, but it can supplement the experimental approach in areas that cannot be so easily studied with experimental methods. For instance, all frequency-related issues of language acquisition crucially rely on corpus data. It is thus not a coincidence that the development of modern corpus linguistics has led to a growing interest in questions concerning input frequency and frequency of occurrence in children's speech. Concluding this article, I discuss some recent corpus-based studies on the acquisition of English syntax (references to the acquisition of other linguistic phenomena and other languages are given in section 4).

3.1. The acquisition of grammatical categories

One of the most hotly debated issues in child language research is the acquisition of grammatical categories. In the 1970s and 1980s, many researchers assumed that grammatical categories, notably nouns and verbs, are learned based on semantically grounded word classes that children acquire prior to grammatical categories. In this view, the grammatical categories of nouns and verbs emerge from two semantically specified word classes, one for objects, persons, and other concrete entities, and the other for activities and events. The semantically specified word classes tend to correlate with particular morphosyntactic features. For instance, in many languages words denoting objects, persons, and concrete entities occur with case markers and determiners, and words denoting activities and events occur with tense and aspect markers. When the child recognizes such correlations he has learned the grammatical properties of nouns and verbs, which can then be extended to lexical expressions that deviate from the semantic prototypes (cf. Bates/MacWhinney 1979, 1989; Bowerman 1973; Schlesinger 1974). Building on this account, Pinker (1984) proposed a theory of category acquisition that has become known as semantic bootstrapping. In this theory, the semantic word classes do not directly evolve into grammatical categories, but help the child to 'hook up' lexical items to innate categories. Specifically, Pinker claims that the semantically specified word classes are needed to link the words of a particular language to the grammatical categories of innate universal grammar.

While these studies emphasize the importance of semantic factors for the acquisition of grammatical categories, other studies have emphasized the importance of distributional cues for linguistic category acquisition. Specifically, Maratsos/Chalkley (1980) have claimed that the distribution of bound morphemes and function words can provide important information about category membership. This hypothesis was at first rejected by Pinker (1984) who argued that distributional patterns are too complex and inconsistent to provide reliable cues for grammatical categories, but more than a decade later, a new generation of computationally oriented child-language researchers presented strong empirical evidence for Maratsos and Chalkley's hypothesis.

Investigating corpora from the CHILDES database, these researchers have shown that there is an enormous amount of distributional information in the ambient language that could help children to learn the grammatical word classes. In one of these studies, Redington/Chater/Finch (1998) examined the distribution of the 1000 most frequent words in the adult English corpus of the CHILDES database, using the 150 most fre-

quent words as context words; that is, for each of the 1000 target words, they determined how often it co-occurred with one of the 150 context words. This information was represented in a ‘context vector’, which they submitted to a hierarchical cluster analysis. The cluster analysis grouped the context vectors into classes (i.e. clusters) based on their similarity. The results of the cluster analysis were then compared to the categorization of the 1000 target words in the Collins Cobuild lexical database, which served as a benchmark. The comparison between the cluster analysis and the categorizations in the CC-database revealed a great deal of overlap, suggesting that distributional information may play an important role in grammatical category acquisition. When the full range of grammatical categories was taken into account, 72% of the target words were clustered correctly. When only nouns and verbs were taken into account, performance improved to 90% accuracy in the case of nouns and 72% accuracy in the case of verbs. Moreover, Redington/Chater/Finch found that context words preceding the target word are more informative for categorization than context words that follow it and that the distributional cues for content words are more reliable than the distributional cues for function words.

Using a somewhat different method, Mintz/Newport/Bever (2002) arrived at the same conclusion as Redington/Chater/Finch and investigated some additional aspects of distributional category acquisition. Among other things they found that information about phrasal boundaries can improve the categorization of nouns and verbs. Assuming that the beginning of a noun phrase or verb phrase correlates with the occurrence of a function word (e.g. a determiner or auxiliary), which even young infants can recognize (cf. Jusczyk 1997), Mintz et al. determined the context of each target word by the last function word preceding it. If we define the context in this way, i.e. if we take phrasal boundaries into account, the distributional analysis is even more powerful to solve the categorization task.

Extending this line of research, Monaghan/Chater/Christiansen (2005) conducted a corpus-based analysis in which they investigated the differential role of phonological and distributional cues in grammatical categorization. Using adult data from the CHILDES database which they transformed into phonological transcriptions, they have shown that there are strong correlations between certain phonological features and particular word classes in the ambient language. Specifically, they have demonstrated that the categorical distinction between nouns and verbs, and open class and closed class items correlates with several phonological features: the presence of stress, the position of stress, the number of syllables, the occurrence of reduced vowels, the complexity of the onset, the occurrence of final voicing, and the occurrence of certain speech sounds (cf. Kelly 1992; Shi/Morgan/Allopenna 1998). Interestingly, the phonological features do not just reinforce the information that children may extract from distributional regularities; rather, they seem to be especially powerful in lexical domains in which distributional information is not so easily available. Distributional cues are especially useful for the categorization of high frequency items, which children encounter many times in the same context; but they are less useful for the categorization of low frequency items, which are not frequent enough to be associated with a particular syntactic context. However, since low frequency items tend to be longer and phonologically more complex than high frequency items (the latter are often phonologically reduced; cf. Zipf 1935), they tend to provide more phonetic information that children can exploit to solve the categorization task.

3.2. The acquisition of complex sentences

Corpus-based methods are not only useful to test hypotheses about quantitative aspects of language acquisition, they are also important to explore the development of particular linguistic phenomena and to generate new hypotheses as to how the acquisition process may proceed. A recent explorative study that has drawn on data from the CHILDES database is Diessel's (2004) analysis of the development of complex sentences in English. Using data from five children between the ages of 2;0 and 5;0, he investigates the development of finite and non-finite complement clauses, finite and non-finite relative clauses, and adverbial and coordinate clauses (see also Diessel/Tomasello 1999, 2000, 2001).

The earliest subordinate clauses that English-speaking children produce are infinitival complement clauses that occur with the complement-taking verbs *want* (i. e. *wanna*) and *have* (i. e. *hafta*) (cf. 20a-b).

- (20) a. I wanna sit down.
- b. I hafta go to my room.

Although the complement-taking verbs behave grammatically like matrix verbs, semantically they function like modals. They do not denote an independent state of affairs, but indicate the child's desire or obligation to perform an activity denoted by the non-finite verb. The whole utterance contains a single proposition and thus functions semantically like a simple main clause. Other early infinitival and participial complement clauses occur with aspectual verbs such as *start* and *stop*, which specify the temporal (or aspectual) structure of the non-finite verb. Like the early quasi-modals, the aspectual verbs do not denote an independent situation, but elaborate the meaning of the activity denoted by the non-finite verb. As children grow older, these constructions gradually evolve into complex sentences in which main and subordinate clauses denote two independent states of affairs (e. g. *Sarah wants Mummy to cook a soup*).

The earliest finite complement clauses occur in sentences including a short and formulaic main clause such as *I think* or *I know* and a complement clause that does not include a complementizer (cf. 21).

- (21) a. I think he's gone.
- b. I know Daddy's there.

Although these utterances comprise two finite clauses, they function semantically like simple sentences. The main clause does not denote a state of affairs, but functions as an epistemic marker, an attention getter, or a marker of the illocutionary force (cf. Thompson 2002; Verhagen 2005). As children grow older, some of the early main clauses become semantically more substantial and new complement-taking verbs emerge that denote an independent state of affairs.

The first relative clauses are finite or non-finite subject-relatives (i. e. relative clauses in which the subject is gapped or relativized) that are attached to the predicate nominal of a copular main clause (cf. 22a–b):

- (22) a. Here's the tiger that's gonna scare him.
- b. That's the horse sleeping in the cradle.

Although these constructions are biclausal, they denote only a single state of affairs. The copular clause does not function as an independent assertion but serves to establish a referent in focus position, where it becomes available for the predication expressed in the relative clause (cf. Lambrecht 1988). As children grow older, they begin to produce more complex relative constructions in which the relative clause is attached to a nominal referent in a fully developed main clause.

Finally, the earliest adverbial and coordinate clauses are independent sentences that are pragmatically linked to a previous utterance: They are intonationally unbound, include a finite verb, and always follow the semantically associated clause. In fact, many of the earliest adverbial and coordinate clauses occur in discourse routines in which they are semantically linked to a previous adult utterance. For instance, children's early *because*-clauses occur in response to an adult *why*-question (cf. 23). As children grow older, these sentences are integrated in bi-clausal constructions that children plan and produce as one unit.

- (23) MOT: Why did you put them in the car?
CHI: Cause Jenny's gonna get the crayons.

The analysis of children's spontaneous use of complex sentences has given rise to new hypotheses, which can now be tested in experiments. Since previous experimental studies were not informed by corpus data, they have often focused on structures that are basically absent from naturally-occurring (child) language (cf. Sheldon 1974; Tavakolian 1977). As a result, children performed very poorly in most of these studies, suggesting that preschool children know very little about complex sentences. However, the corpus-based analyses have shown that the development of complex sentences begins very early and that preschool children make common use of a wide variety of complex sentence constructions. What we need are new experiments that take the corpus-based findings into account (e. g. Diessel/Tomasello 2005).

4. Conclusion

Linguistic corpora play an important role in child language research. The previous section has concentrated on the acquisition of English syntax (see also O'Grady 1997), but corpus-based methods have also been used in studies on the acquisition of phonology (cf. Vihman 1996), morphology (cf. Brown 1973; Marcus et al. 1992), discourse (cf. Hickman 2002), and lexical semantics (cf. Clark 1993, 2003). There are also corpus-based studies on the acquisition of other languages (see Slobin 1985, 1992, 1997 for a comprehensive overview of cross-linguistic research on language acquisition), but most child language corpora of other languages are small and untagged. There is thus an enormous discrepancy between the data available for English and the data for other languages. But even the English corpus data are not sufficient to address all research questions. In particular, the development of rare linguistic phenomena cannot be studied based on the corpora included in CHILDES. What we need, in addition to more data from other languages, are dense corpora that allow the child language researcher to track the development of particular linguistic phenomena in more detail (see section 1.1.).

5. Literature

- Bates, E./MacWhinney, B. (1979), The Functionalist Approach to the Acquisition of Grammar. In: Ochs, E./Schieffelin, B. B. (eds.), *Developmental Pragmatics*. New York: Academic Press, 167–211.
- Bates, E./MacWhinney, B. (1989), Functionalism and Competition Model. In: MacWhinney, B./Bates, E. (eds.), *Mechanisms of Language Acquisition*. Cambridge: Cambridge University Press, 3–73.
- Bates, E./Dale, P. S./Thal, D. (1994), Individual Differences and their Implications for Theories of Language Development. In: Fletcher, P./MacWhinney, B. (eds.), *The Handbook of Child Language*. Cambridge: Blackwell, 96–151.
- Bellugi, U. (1967), The Acquisition of Negation. Ph.D. dissertation, Harvard University.
- Bloom, L. (1970), *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA: MIT Press.
- Bloom, L. (1973), *One Word at a Time. The Use of Single Word Utterances*. The Hague: Mouton.
- Bowerman, M. (1973), Structural Relationships in Children's Utterances: Syntactic or Semantic? In: Moore, T. A. (ed.), *Cognitive Development and the Acquisition of Language*. New York: Academic Press, 197–213.
- Braine, M. (1963), The Ontogeny of English Phrase Structure: The First Phase. *Language* 39, 1–13.
- Braine, M. (1976), *Children's First Word Combinations*. (Monographs of the Society for Research in Child Development, 41.) Chicago: University of Chicago Press.
- Brent, M. R./Cartwright, T. A. (1996), Distributional Regularity and Phonotactic Constraints are Useful for Segmentation. In: *Cognition* 61, 93–125.
- Brown, R. (1968), The Development of *wh* Questions in Child Speech. In: *Journal of Abnormal and Social Psychology* 55, 1–5.
- Brown, R. (1973), *A First Language. The Early Stages*. Cambridge, MA: Harvard University Press.
- Cartwright, T. A./Brent, M. R. (1997), Syntactic Categorization in Early Language Acquisition: Formalizing the Role of Distributional Analysis. In: *Cognition* 63, 121–170.
- Chapman, R. S. (1994), Language Development in Children and Adolescents with Down Syndrom. In: Fletcher, P./MacWhinney, B. (eds.), *The Handbook of Child Language*. Cambridge: Blackwell, 641–663.
- Clark, E. V. (1993), *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.
- Clark, E. V. (2003), *First Language Acquisition*. Cambridge: Cambridge University Press.
- Conti-Ramsden, G./Dykins, J. (1991), Mother-Child Interactions with Language-impaired Children and their Siblings. In: *British Journal of Disorders of Communication* 26, 337–354.
- de Houwer, A. (1990), *The Acquisition of Two Languages: A Case Study*. New York: Cambridge University Press.
- Diessel, H. (2004), *The Acquisition of Complex Sentences*. Cambridge: Cambridge University Press.
- Diessel, H./Tomasello, M. (1999), Why Complement Clauses do not Include a *that*-complementizer in Early Child Language. In: *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA, 86–97.
- Diessel, H./Tomasello, M. (2000), The Development of Relative Clauses in English. In: *Cognitive Linguistics* 11, 131–151.
- Diessel, H./Tomasello, M. (2001), The Acquisition of Finite Complement Clauses in English: A Corpus-based analysis. In: *Cognitive Linguistics* 12, 1–45.
- Diessel, H./Tomasello, M. (2005), A New Look at the Acquisition of Relative Clauses. In: *Language* 81, 1–25.
- Hickman, M. (2002), *Children's Discourse: Person, Space and Time across Languages*. Cambridge: Cambridge University Press.
- Hoff-Ginsberg, E. (1991), Mother-Child Conversation in Different Social Classes and Communicative Social Settings. In: *Child Development* 62, 782–796.

- Joseph, K. L./Serrratrice, L./Conti-Ramsden, G. (2002), Development of Copula and Auxiliary BE in Children with Specific Language Impairment and Younger Unaffected Controls. In: *First Language* 22, 137–172.
- Jusczyk, P. W. (1997), *The Discovery of Spoken Language*. Cambridge: MIT Press.
- Karrass, J./Braungart-Rieker, J. M./Mullins, J./Lefever, J. B. (2002), Processes in Language Acquisition: The Roles of Gender, Attention, and Maternal Encouragement of Attention over Time. In: *Journal of Child Language* 29, 519–543.
- Kelly, M. H. (1992), Using Sound to Solve Syntactic Problems: The Role of Phonology in Grammatical Category Assignment. In: *Psychological Review* 99, 349–364.
- Lambrecht, K. (1988), There was a farmer had a dog: Syntactic Amalgams Revisited. In: *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA, 319–339.
- Lawrence V. W./Shipley, E. F. (1996), Parental Speech to Middle- and Working-class Children from Two Racial Groups and Three Settings. In: *Applied Psycholinguistics* 17, 233–255.
- Leopold, W. (1939–1949), *Speech Development of a Bilingual Child*, 4 vols. Evanston, IL: Northwestern University Press.
- Lieven, E./Pine, J./Barnes, H. (1992), Individual Differences in Early Vocabulary Development: Redefining the Referential-expressive Distinction. In: *Journal of Child Language* 19, 287–310.
- MacWhinney, B. (1995), *The CHILDES Project. Tools for Analyzing Talk*. Second edition. Hillsdale, NJ: Lawrence Erlbaum.
- MacWhinney, B. (1996), The CHILDES System. In: *American Journal of Speech-Language Pathology* 5, 5–14.
- MacWhinney, B. (2000), *The CHILDES Project. Tools for Analyzing Talk*. Vol. II: *The Database*. Third edition. Hillsdale, NJ: Lawrence Erlbaum.
- MacWhinney, B./Snow, C. (1985), The Child Language Data Exchange System. In: *Journal of Child Language* 12, 271–296.
- Maratsos, M./Chalkley, M. A. (1980), The Internal Language of Children's Syntax: The Ontogenesis and Representation of Syntactic Categories. In: K. Nelson (ed.), *Children's Language*, vol. 2. New York: Gardner Press, 127–214.
- Marcus, G. F./Pinker, S./Ullman, M./Hollander, M./Rosen, T. J./Xu, F. (1992), *Overregularization in Language Acquisition*. (Monographs of the Society for Research in Child Development 57 (Serial No. 228).) Chicago: University of Chicago Press.
- McDaniel, D./McKee, C./Smith Cairns, H. (eds.) (1996), *Methods for Assessing Children's Syntax*. Cambridge: MIT Press.
- Mintz, T. H./Newport, E. L./Bever, T. G. (2002), The Distributional Structure of Grammatical Categories in the Speech to Young Children. In: *Cognitive Science* 26, 393–424.
- Monaghan, P./Chater, N./Christiansen, M. H. (2005), The Differential Role of Phonological and Distributional Cues in Grammatical Categorization. In: *Cognition* 96, 143–182.
- O'Grady, W. (1997), *Syntactic Development*. Chicago: Chicago University Press.
- Pinker, S. (1984), *Language Learnability and Language Development*. Cambridge: Harvard University Press.
- Redington, M./Chater, N./Finch, S. (1998), Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. In: *Cognitive Science* 22, 435–469.
- Schlesinger, I. M. (1974), Relational Concepts Underlying Language Acquisition. In: Schiefelbusch, R. L./Lloyd, L. (eds.), *Language Perspectives: Acquisition, Retardation, and Intervention*. Baltimore: University Park Press, 129–151.
- Sheldon, A. (1974), The Role of Parallel Function in the Acquisition of Relative Clauses in English. In: *Journal of Verbal Learning and Verbal Behavior* 13, 272–281.
- Shi, R./Morgan, J./Alloppenna, P. (1998), Phonological and Acoustic Bases for Earliest Grammatical Category Assignment: A Cross-linguistic Perspective. In: *Journal of Child Language* 25, 169–201.

- Slobin, D. I. (ed.) (1985), *The Cross-linguistic Study of Language Acquisition*, vol. 1–2. Hillsdale: Lawrence Erlbaum.
- Slobin, D. I. (ed.) (1992), *The Cross-linguistic Study of Language Acquisition*, vol. 3. Hillsdale: Lawrence Erlbaum.
- Slobin, D. I. (ed.) (1997), *The Cross-linguistic Study of Language Acquisition*, vol. 4–5. Hillsdale: Lawrence Erlbaum.
- Stern, C./Stern, W. (1907), *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung*. Leipzig: Barth.
- Tavakolian, S. L. 1977. Structural Principles in the Acquisition of Complex Sentences. Ph.D. dissertation, University of Massachusetts.
- Thompson, S. (2002), ‘Object Complements’ and Conversation: Towards a Realistic Account. In: *Studies in Language* 26, 125–164.
- Tomasello, M./Stahl, D. (2004), Sampling Children’s Spontaneous Speech: How Much is Enough? In: *Journal of Child Language* 31, 101–121.
- Verhagen, A. (2005), *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition*. Oxford: Oxford University Press.
- Vihman, M. M. (1996), *Phonological Development: The Origins of Language in the Child*. Oxford: Blackwell.
- Yip, V./Matthews, S. (2000), Syntactic Transfer in a Bilingual Child. In: *Bilingualism: Language and Cognition* 3, 193–208.
- Zipf, G. (1935), *The Psycho-biology of Language*. Boston: Houghton Mifflin.

Holger Diessel, Jena (Germany)

58. Corpora and collocations

1. Introduction
2. What are collocations?
3. Cooccurrence and frequency counts
4. Simple association measures
5. Statistical association measures
6. Finding the right measure
7. Summary and conclusion
8. Literature

1. Introduction

1.1. The controversy around collocations

The concept of *collocations* is certainly one of the most controversial notions in linguistics, even though it is based on a compelling, widely shared intuition that certain words have a tendency to occur near each other in natural language. Examples of such collocations are *cow* and *milk*, *day* and *night*, *ring* and *bell*, or the infamous *kick* and *bucket*. Other words, like *know* and *glass* or *door* and *year*, do not seem to be particularly attracted to each other. J. R. Firth (1957) introduced the term “collocations” for charac-

- Slobin, D. I. (ed.) (1985), *The Cross-linguistic Study of Language Acquisition*, vol. 1–2. Hillsdale: Lawrence Erlbaum.
- Slobin, D. I. (ed.) (1992), *The Cross-linguistic Study of Language Acquisition*, vol. 3. Hillsdale: Lawrence Erlbaum.
- Slobin, D. I. (ed.) (1997), *The Cross-linguistic Study of Language Acquisition*, vol. 4–5. Hillsdale: Lawrence Erlbaum.
- Stern, C./Stern, W. (1907), *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung*. Leipzig: Barth.
- Tavakolian, S. L. 1977. Structural Principles in the Acquisition of Complex Sentences. Ph.D. dissertation, University of Massachusetts.
- Thompson, S. (2002), ‘Object Complements’ and Conversation: Towards a Realistic Account. In: *Studies in Language* 26, 125–164.
- Tomasello, M./Stahl, D. (2004), Sampling Children’s Spontaneous Speech: How Much is Enough? In: *Journal of Child Language* 31, 101–121.
- Verhagen, A. (2005), *Constructions of Intersubjectivity. Discourse, Syntax, and Cognition*. Oxford: Oxford University Press.
- Vihman, M. M. (1996), *Phonological Development: The Origins of Language in the Child*. Oxford: Blackwell.
- Yip, V./Matthews, S. (2000), Syntactic Transfer in a Bilingual Child. In: *Bilingualism: Language and Cognition* 3, 193–208.
- Zipf, G. (1935), *The Psycho-biology of Language*. Boston: Houghton Mifflin.

Holger Diessel, Jena (Germany)

58. Corpora and collocations

1. Introduction
2. What are collocations?
3. Cooccurrence and frequency counts
4. Simple association measures
5. Statistical association measures
6. Finding the right measure
7. Summary and conclusion
8. Literature

1. Introduction

1.1. The controversy around collocations

The concept of *collocations* is certainly one of the most controversial notions in linguistics, even though it is based on a compelling, widely shared intuition that certain words have a tendency to occur near each other in natural language. Examples of such collocations are *cow* and *milk*, *day* and *night*, *ring* and *bell*, or the infamous *kick* and *bucket*. Other words, like *know* and *glass* or *door* and *year*, do not seem to be particularly attracted to each other. J. R. Firth (1957) introduced the term “collocations” for charac-

teristic and frequently recurrent word combinations, arguing that the meaning and usage of a word (the *node*) can to some extent be characterised by its most typical *collocates*: “You shall know a word by the company it keeps” (Firth 1957, 179). Firth was clearly aware of the limitations of this approach. He understood collocations as a convenient first approximation to meaning at a purely lexical level that can easily be operationalised (cf. Firth 1957, 181). Collocations in this Firthian sense can also be interpreted as empirical statements about the predictability of word combinations: they quantify the “mutual expectancy” (Firth 1957, 181) between words and the statistical influence a word exerts on its neighbourhood. Firth’s definition of the term remained vague, though, and it was only formalised and implemented after his death, by a group of British linguists often referred to as the Neo-Firthian school. Collocations have found widespread application in computational lexicography (Sinclair 1966, 1991), resulting in corpus-based dictionaries such as COBUILD (Sinclair 1995; see also article 8).

In parallel to the development of the Neo-Firthian school, the term “collocations” came to be used in the field of phraseology for semi-compositional and lexically determined word combinations such as *stiff drink* (with a special meaning of *stiff* restricted to a particular set of nouns), *heavy smoker* (where *heavy* is the only acceptable intensifier for *smoker*), *give a talk* (rather than *make* or *hold*) and *a school of fish* (rather than *group*, *swarm* or *flock*). This view has been advanced forcefully by Hausmann (1989) and has found increasingly widespread acceptance in recent years (e. g. Grossmann/Tutin 2003). It is notoriously difficult to give a rigorous definition of collocations in the phraseological sense and differentiate them from restricted word senses (most dictionaries have separate subentries for the special meanings of *stiff*, *heavy* and *school* in the examples above). There is considerable overlap between the phraseological notion of collocations and the more general empirical notion put forward by Firth (cf. the examples given above), but they are also different in many respects (e. g., *good* and *time* are strongly collocated in the empirical sense, but *a good time* can hardly be understood as a non-compositional or lexically restricted expression). This poor alignment between two interpretations of the same term has resulted in frequent misunderstandings and has led to enormous confusion on both sides. The situation is further complicated by a third meaning of “collocations” in the field of computational linguistics, where it is often used as a generic term for any lexicalised word combination that has idiosyncratic semantic or syntactic properties and may therefore require special treatment in a machine-readable dictionary or natural language processing system. This usage seems to originate with Choueka (1988) and can be found in standard textbooks, where collocations are often defined in terms of non-compositionality, non-modifiability and non-substitutability (Manning/Schütze 1999, 184). It has recently been superseded by the less ambiguous term *multiword expression* (cf. Sag et al. 2002).

An excellent overview of the competing definitions of collocations and their historical development is given by Bartsch (2004). Interestingly, she takes a middle road with her working definition of collocations as “lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other” (Bartsch 2004, 76). For a compact summary, refer to Williams (2003).

1.2. Definitions and recommended terminology

In order to avoid further confusion, a consistent terminology should be adopted. Its most important goal is to draw a clear distinction between (i) the *empirical* concept of

recurrent and predictable word combinations, which are a directly observable property of natural language, and (ii) the *theoretical* concept of lexicalised, idiosyncratic multiword expressions, defined by linguistic tests and speaker intuitions. In this article, the term “*collocations*” is used exclusively in its empirical Firthian sense (i), and we may occasionally speak of “*empirical collocations*” to draw attention to this fact. Lexicalised word combinations as a theoretical, phraseological notion (ii) are denoted by the generic term “*multiword expressions*”, following its newly established usage in the field of computational linguistics. In phraseological theory, multiword expressions are divided into sub-categories ranging from completely opaque idioms to semantically compositional word combinations, which are merely subject to arbitrary lexical restrictions (*brush teeth* rather than *scrub teeth*) or carry strong pragmatic connotations (*red rose*). A particularly interesting category in the middle of this cline are semi-compositional expressions, in which one of the words is lexically determined and has a modified or bleached meaning (classic examples are *heavy smoker* and *give a talk*). They correspond to the narrow phraseological meaning of the term “*collocations*” (cf. Grossmann/Tutin 2003) and can be referred to as “*lexical collocations*”, following Krenn (2000). As has been pointed out above, it is difficult to give a precise definition of lexical collocations and to differentiate them e. g. from specialised word senses. Because of this fuzziness and the fact that many empirical collocations are neither completely opaque nor fully compositional, similar to lexical collocations, the two concepts are easily and frequently confused.

This article is concerned exclusively with empirical collocations, since they constitute one of the fundamental notions of corpus linguistics and, unlike lexicalisation phenomena, can directly be observed in corpora. It is beyond the scope of this text to delve into the voluminous theoretical literature on multiword expressions, but see e. g. Bartsch (2004) and Grossmann/Tutin (2003) for useful pointers. There is a close connection between empirical collocations and multiword expressions, though. A thorough analysis of the collocations found in a corpus study will invariably bring up non-compositionality and lexicalisation phenomena as an explanation for many of the observed collocations (cf. the case study in section 2.2.). Conversely, theoretical research in phraseology can build on authentic examples of multiword expressions obtained from corpora, avoiding the bias of relying on introspection or stock examples like *kick the bucket* (which is a rather uncommon phrase indeed: only three instances of the idiom can be found in the 100 million words of the British National Corpus). *Multiword extraction* techniques exploit the often confusing overlap between the empirical and theoretical notions of collocation. Empirical collocations are identified as candidate multiword expressions, and then the “false positives” are weeded out by manual inspection. A more detailed account of such multiword extraction procedures can be found in section 6.2.

Following the Firthian tradition (e. g. Sinclair 1991), we define a collocation as a combination of two words that exhibit a tendency to occur near each other in natural language, i. e. to *cooccur* (but see the remarks on combinations of three or more words in section 7.1.). The term “*word pair*” is used to refer to such a combination of two words (or, more precisely, word *types*; see article 36 for the distinction between types and tokens) in a neutral way without making a commitment regarding its collocational status. In order to emphasise this view of collocations as word pairs, we will use the notation *(kick, bucket)* instead of e. g. *kick (the) bucket*. In general, a word pair is denoted by (w_1, w_2) , with $w_1 = \text{kick}$ and $w_2 = \text{bucket}$ in the previous example; w_1 and w_2 are also referred to as the *components* of the word pair. The term “*word*” is meant in

the widest possible sense here and may refer to surface forms, case-folded surface forms, base forms, etc. (see article 25). While collocations are most commonly understood as combinations of orthographic words, delimited by whitespace and punctuation, the concept and methodological apparatus can equally well be applied to combinations of linguistic units at other levels, ranging from morphemes to phrases and syntactic constructions (cf. article 43).

In order to operationalise our definition of collocations, we need to specify the precise circumstances under which two words can be said to “cooccur”. We also need a formal definition of the “attraction” between words reflected by their repeated cooccurrence, and a quantitative measure for the strength of this attraction. The cooccurrence of words can be defined in many different ways. The most common approaches are (i) *surface cooccurrence*, where words are said to cooccur if they appear close to each other in running text, measured by the number of intervening word tokens; (ii) *textual cooccurrence* of words in the same sentence, clause, paragraph, document, etc.; and (iii) *syntactic cooccurrence* between words in a (direct or indirect) syntactic relation, such as a noun and its modifying adjective (which tend to be adjacent in most European languages) or a verb and its object noun (which may be far apart at the surface, cf. Goldman/Nerima/Wehrli (2001, 62) for French). These three definitions of cooccurrence are described in more detail in section 3, together with appropriate methods for the calculation of cooccurrence frequency data.

The hallmark of an attraction between words is their frequent cooccurrence, and collocations are sometimes defined simply as “recurrent cooccurrences” (Smadja 1993, 147; Bartsch 2004, 11). Strictly speaking, any pair of words that cooccur at least twice in a corpus is a potential collocation according to this view. It is common to apply higher *frequency thresholds*, however, such as a minimum of 3, 5 or even 10 cooccurrences. Evert (2004, chapter 4) gives a mathematical justification for this approach (see also section 7.1.), but a more practical reason is to reduce the enormous amounts of data that have to be processed. It is not uncommon to find more than a million recurrent word pairs ($f \geq 2$) in a corpus containing several hundred million running words, but only a small proportion of them will pass a frequency threshold of $f \geq 10$ or higher, as a consequence of Zipf’s law (cf. article 37). In the following, we use the term “*recurrent word pair*” for a potential collocation that has passed the chosen frequency threshold in a given corpus.

Mere recurrence is not a sufficient indicator for a strong attraction between words, though, as will be illustrated in section 4.1. An additional measure of attraction strength is therefore needed in order to identify “true collocations” among the recurrent word pairs, or to distinguish between “strong” and “weak” collocations. The desire to generalise from recurrent word pairs in a particular corpus (as a sample of language) to collocations in the full language or sublanguage, excluding word pairs whose recurrence may be an accident of the sampling process, has led researchers to the concept of *statistical association* (Sinclair 1966, 418). Note that this mathematical meaning of “association” describes a statistical attraction between certain events and must not be confused with psychological association (as e. g. in word association norms, which have no direct connection to the statistical association between words that is of interest here). By interpreting occurrences of words as events, statistical *association measures* can be used to quantify the attraction between cooccurring words. They thus complete the formal definition of empirical collocations.

The most important association measures will be introduced in sections 4 and 5, but many other measures have been suggested in the mathematical literature and in collocation studies. Such measures assign an *association score* to each word pair, with high scores indicating strong attraction and low scores indicating weak attraction (or even repulsion) between the component words. Association scores can then be used to select “true collocations” by setting a threshold value, or to rank the set of recurrent word pairs according to the strength of their attraction (so that “strong” collocations are found at the top of the list). These uses of association scores are further explained in section 2.1. It is important to keep in mind that different association measures may lead to entirely different rankings of the word pairs (or to different sets of “true collocations”). Section 6 gives some guidance on how to choose a suitable measure.

1.3. Overview of the article

Section 2 describes the different uses of association scores and illustrates the linguistic properties of empirical collocations with a case study of the English noun *bucket*. The three types of cooccurrence (surface, textual and syntactic) are defined and compared in section 3, and the calculation of cooccurrence frequency data is explained with the help of toy examples. Section 4 introduces the concepts of statistical association and independence underlying all association measures. It also presents a selection of simple measures, which are based on a comparison of observed and expected cooccurrence frequency. Section 5 introduces more complex statistical measures based on full-fledged contingency tables. The difficulty of choosing between the large number of available measures is the topic of section 6, which discusses various methods for the comparison of association measures. Finally, section 7 addresses some open questions and extensions that are beyond the scope of this article, and lists references for further reading.

Readers in a hurry may want to start with the “executive summaries” in section 4.3. and at the beginning of section 7, which give a compact overview of the collocation identification process with simple association measures. You should also skim the examples in section 3 to understand how appropriate cooccurrence frequency data are obtained from a corpus, find out in section 4.1. how to calculate the observed cooccurrence frequency O and expected frequency E , and refer to Figure 58.4 for the precise equations of various simple association measures.

2. What are collocations?

2.1. Using association scores

Association scores as a quantitative measure of the attraction between words play a crucial role in the operationalisation of empirical collocations, next to the formal definition of cooccurrence and the appropriate calculation of cooccurrence frequency data. While the interpretation of association scores seems straightforward (high scores indicate strong attraction), they can be used in different ways to identify collocations among the recurrent word pairs found in a corpus. The first contrast to be made is whether colloca-

tivity is treated as a categorical phenomenon or as a cline, leading either to *threshold* approaches (which attempt to identify “true collocations”) or to *ranking* approaches (which place word pairs on a scale of collocational strength without strict separation into collocations and non-collocations). A second contrast concerns the grouping of collocations: the *unit* view is interested in the most strongly collocated word pairs, which are seen as independent units; the *node-collocate* view focuses on the collocates of a given node word, i. e. “the company it keeps”. The two contrasts are independent of each other in principle, although the *node-collocate* view is typically combined with a ranking approach.

In a threshold approach, recurrent word pairs whose association score exceeds a (more or less arbitrary) threshold value specified by the researcher are accepted as “true collocations”. We will sometimes refer to them as an *acceptance set* for a given association measure and threshold value. In the alternative approach, all word pairs are ranked according to their association scores. Pairs at the top of the ranked list are then considered “more collocational”, while the ones at the bottom are seen as “less collocational”. However, no categorical distinction between collocations and non-collocations is made in this approach. A third strategy combines the ranking and threshold approaches by accepting the first n word pairs from the ranked list as collocations, with n either determined interactively by the researcher or dictated by the practical requirements of an application. Typical choices are $n = 100$, $n = 500$, $n = 1000$ and $n = 2000$. Such *n-best lists* can be interpreted as acceptance sets for a threshold value determined from the corpus data (such that exactly n word pairs are accepted) rather than chosen at will. Because of the arbitrariness of pre-specified threshold values and the lack of good theoretical motivations (cf. section 4.2.), *n-best* lists should always be preferred over threshold-based acceptance sets. It is worth pointing out that in either case the ranking, *n-best* list or acceptance set depends critically on the particular association measure that has been used. The *n-best* lists shown in Tables 58.2 and 58.3 in section 4.3. are striking examples of this fact.

The unit view interprets collocations as pairs of words that show a strong mutual attraction, or “mutual expectancy” (Firth 1957, 181). It is particularly suitable and popular for multiword extraction tasks, where *n-best* lists containing the most strongly associated word pairs in a corpus are taken as candidate multiword expressions. Such candidate lists serve e. g. as base material for dictionary updates, as terminological resources for translators and technical writers, and for the semi-automatic compilation of lexical resources for natural language processing systems (e. g. Heid et al. 2000). The node-collocate view, on the other hand, focuses on the predictability of word combinations, i. e. on how a word (the node) determines its “company” (the collocates). It is well suited for the linguistic description of word meaning and usage in the Firthian tradition, where a node word is characterised by ranked lists of its collocates (Firth 1957). Following Firth (1957, 195–196) and Sinclair (1966), this view has also found wide acceptance in modern corpus-based lexicography (e. g. Sinclair 1991; Kilgarriff et al. 2004), in particular for learner dictionaries such as *COBUILD* (Sinclair 1995) and the *Oxford Collocations Dictionary* (Lea 2002).

In addition to their “classic” applications in language description, corpus-based lexicography and multiword extraction, collocations and association scores have many practical uses in computational linguistics and related fields. Well-known examples include the construction of machine-readable dictionaries for machine translation and natural

language generation systems, the improvement of statistical language models, and the use of association scores as features in vector space models of distributional semantics. See Evert (2004, 23–27) for an overview and comprehensive references.

2.2. Collocations as a linguistic epiphenomenon

The goal of this section is to help readers reach an intuitive understanding of the empirical phenomenon of collocations and their linguistic properties. First and foremost, collocations are observable facts about language, i.e. primary data. From a strictly data-driven perspective, they can be interpreted as empirical predictions about the neighbourhood of a word. For instance, a verb accompanying the noun *kiss* is likely to be either *give, drop, plant, press, steal, return, deepen, blow* or *want*. From the explanatory perspective of theoretical linguistics, on the other hand, collocations are best characterised as an *epiphenomenon*: idioms, lexical collocations, clichés, cultural stereotypes, semantic compatibility and many other factors are hidden causes that result in the observed associations between words. In order to gain a better understanding of collocations both as an empirical phenomenon and as an epiphenomenon, we will now take a look at a concrete example, viz. how the noun *bucket* is characterised by its collocates in the British National Corpus (BNC, Aston/Burnard 1998). The data presented here are based on surface cooccurrence with a span size of 5 words, delimited by sentence boundaries (see section 3). Observed and expected frequencies were calculated as described in section 4.1. Collocates were lemmatised, and punctuation, symbols and numbers were excluded. Association scores were calculated for the measures MI and simple-ll (see section 4.2.).

A first observation is that different association measures will produce entirely different rankings of the collocates. For the MI measure, the top collocates are *fourteen-record, ten-record, full-track, single-record, randomize, galvanized, groundbait, slop, spade, Nessie*. Most of them are infrequent words with low cooccurrence frequency (e.g., *groundbait* occurs only 29 times in the BNC). Interestingly, the first five collocates belong to a technical sense of *bucket* as a data structure in computer science; others such as *groundbait* and *Nessie* (the name of a character in the novel *Worlds Apart*, BNC file ATE) are purely accidental combinations. By contrast, the top collocates according to the simple-ll measure are dominated by high-frequency cooccurrences with very common words, including several function words: *water, a, spade, plastic, size, slop, mop, throw, fill, with*.

A clearer picture emerges when different parts of speech among the collocates (e.g. nouns, verbs and adjectives) are listed separately, as shown in Table 58.1 for the simple-ll measure. Ideally, a further distinction should be made according to the syntactic relation between node and collocate (node as subject/object of verb, prenominal adjective modifying the node, head of postnominal *of-NP*, etc.), similar to the lexicographic *word sketches* of Kilgarriff et al. (2004). Parts of speech provide a convenient approximation that does not require sophisticated automatic language processing tools. A closer inspection of the lists in Table 58.1 underlines the status of collocations as an epiphenomenon, revealing many different causes that contribute to the observed associations:

Tab. 58.1: Collocates of *bucket* in the BNC (nouns, verbs and adjectives)

noun	f	simple-l1	verb	f	simple-l1	adjective	f	simple-l1
water	183	1063.90	throw	36	165.32	large	37	92.72
spade	31	338.21	fill	29	129.69	single-record	5	79.56
plastic	36	242.63	randomize	9	115.33	cold	13	52.63
slop	14	197.65	empty	14	106.51	galvanized	4	52.35
size	41	193.22	tip	10	62.65	ten-record	3	49.75
mop	16	183.97	kick	12	59.12	full	20	46.34
record	38	155.64	hold	31	58.52	empty	9	36.41
bucket	18	138.70	carry	26	55.68	steaming	4	36.37
ice	22	131.68	put	36	48.69	full-track	2	33.17
seat	20	78.35	chuck	7	48.40	multi-record	2	33.17
coal	16	76.44	weep	7	44.14	small	21	30.90
density	11	66.78	pour	9	39.35	leaky	3	30.14
brigade	10	66.78	douse	4	37.85	bottomless	3	29.04
algorithm	9	66.54	fetch	7	35.22	galvanised	3	28.34
shovel	7	64.53	store	7	30.77	iced	3	25.46
container	10	62.40	drop	9	21.76	clean	7	25.17
oats	7	62.32	pick	11	21.74	wooden	6	24.14
sand	12	61.91	use	31	20.93	old	19	18.83
Rhino	7	60.50	tire	3	20.58	ice-cold	2	17.66
champagne	10	59.28	rinse	3	20.19	anti-sweat	1	16.58

- the well-known idiom *kick the bucket*, although many of the cooccurrences represent a literal reading of the phrase (e. g. *It was as if God had kicked a bucket of water over.*, G0P: 2750);
- proper names such as *Rhino Bucket*, a hard rock band founded in 1987;
- both lexicalised and productively formed compound nouns: *slop bucket*, *bucket seat*, *coal bucket*, *champagne bucket* and *bucket shop* (the 23rd noun collocate);
- lexical collocations like *weep buckets*, where *buckets* has lost its regular meaning and acts as an intensifier for the verb;
- cultural stereotypes and institutionalised phrases such as *bucket and spade* (which people prototypically take along when they go to a beach, even though the phrase has fully compositional meaning);
- reflections of semantic compatibility: *throw*, *carry*, *kick*, *tip*, *take*, *fetch* are typical things one can do with a bucket, and *full*, *empty*, *leaky* are some of its typical properties (or states);
- semantically similar terms (*shovel*, *mop*) and hypernyms (*container*);
- facts of life, which do not have special linguistic properties but are frequent simply because they describe a situation that often arises in the real world; a prototypical example is *bucket of water*, the most frequent noun collocate in Table 58.1;
- linguistic relevance: it is more important to talk about *full*, *empty* and *leaky* buckets than e. g. about a rusty or yellow bucket; interestingly, *old bucket* ($f = 19$) is much more frequent than *new bucket* ($f = 3$, not shown); and
- “indirect” collocates (e. g. *a bucket of cold, warm, hot, iced, steaming water*), describing typical properties of the liquid contained in a bucket.

Obviously, there are entirely different sets of collocates for each sense of the node word, which are overlaid in Table 58.1. As Firth put it: “there are the specific contrastive collocations for *light/dark* and *light/heavy*” (Firth 1957, 181). In the case of *bucket*, a technical meaning, referring to a specific data structure in computer science, is conspicuous and accounts for a considerable proportion of the collocations (*bucket brigade algorithm*, *bucket size*, *randomize to a bucket*, *store records in bucket*, *single-record bucket*,

ten-record bucket). In order to separate collocations for different word senses automatically, a sense-tagged corpus would be necessary (cf. article 26).

Observant readers may have noticed that the list of collocations in Table 58.1 is quite similar to the entry for *bucket* in the *Oxford Collocations Dictionary* (OCD, Lea 2002). This is not as surprising as it may seem at first, since the OCD is also based on the British National Corpus as its main source of corpus data (Lea 2002, viii). Obviously, collocations were identified with a technique similar to the one used here.

3. Cooccurrence and frequency counts

As has already been stated in section 1.2., the operationalisation of collocations requires a precise definition of the cooccurrence, or “nearness”, of two words (or, more precisely, word *tokens*). Based on this definition, cooccurrence frequency data for each recurrent word pair (or, more precisely, pair of word *types*) can be obtained from a corpus. Association scores as a measure of attraction between words are then calculated from these frequency data. It will be shown in section 4.1. that *cooccurrence frequency* alone is not sufficient to quantify the strength of attraction. It is also necessary to consider the occurrence frequencies of the individual words, known as *marginal frequencies*, in order to assess whether the observed cooccurrences might have come about by chance. In addition, a measure of corpus size is needed to interpret absolute frequency counts. This measure is referred to as *sample size*, following statistical terminology.

The following notation is used in this article: O for the “observed” cooccurrence frequency in a given corpus (sometimes also denoted by f , especially when specifying frequency thresholds such as $f \geq 5$); f_1 and f_2 for the marginal frequencies of the first and second component of a word pair, respectively; and N for the sample size. These four numbers provide the information needed to quantify the statistical association between two words, and they are called the *frequency signature* of the pair (Evert 2004, 36). Note that a separate frequency signature is computed for every recurrent word pair (w_1, w_2) in the corpus. The set of all such recurrent word pairs together with their frequency signatures is referred to as a *data set*.

Three different approaches to measuring nearness are introduced below and explained with detailed examples: *surface*, *textual* and *syntactic* cooccurrence. For each type of cooccurrence, an appropriate procedure for calculating frequency signatures (O, f_1, f_2, N) is described. The mathematical reasons behind these procedures will become clear in section 5. The aim of the present section is to clarify the logic of computing cooccurrence frequency data. Practical implementations that can be applied to large corpora use more efficient algorithms, especially for surface cooccurrences (e. g. Gil/Dias 2003; Terra/Clarke 2004).

3.1. Surface cooccurrence

The most common approach in the Firthian tradition defines cooccurrence by surface proximity, i. e. two words are said to cooccur if they appear within a certain distance or *collocational span*, measured by the number of intervening word tokens. Surface cooccur-

rence is often, though not always combined with a node-collocate view, looking for collocates within the collocational spans around the instances of a given node word.

Span size is the most important choice that has to be made by the researcher. The most common values range from 3 to 5 words (e. g. Sinclair 1991), but many other span sizes can be found in the literature. Some studies in computational linguistics have focused on bigrams of immediately adjacent words, i. e. a span size of 1 (e. g. Choueka 1988; Schone/Jurafsky 2001), while others have used span sizes of dozens or hundreds of words, especially in the context of distributional semantics (Schütze 1998). Other decisions are whether to count only word tokens or all tokens (including punctuation and numbers), how to deal with multiword units (does *out of* count as a single token or as two tokens?), and whether cooccurrences are allowed to cross sentence boundaries.

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [...] There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it. The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over, as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

Fig. 58.1: Illustration of surface cooccurrence for the word pair (*hat, roll*)

Figure 58.1 shows surface cooccurrences between the words *hat* (in bold face, as node) and *roll* (in italics, as collocate). The span size is 4 words, excluding punctuation and limited by sentence boundaries. Collocational spans around instances of the node word *hat* are indicated by brackets below the text. There are two cooccurrences in this example, in the second and third span, hence $O = 2$. Note that multiple instances of a word in the same span count as multiple cooccurrences, so for *hat* and *over* we would also calculate $O = 2$ (with both cooccurrences in the third span). The marginal frequencies of the two words are given by their overall occurrence counts in the text, i. e. $f_1 = 3$ for *hat* and $f_2 = 3$ for *roll*. The sample size N is simply the total number of tokens in the corpus, counting only tokens that are relevant to the definition of spans. In our example, N is the number of word tokens excluding punctuation, i. e. $N = 111$ for the text shown in Figure 58.1. If we include punctuation tokens in our distance measurements, the sample size would accordingly be increased to $N = 126$ (9 commas, 4 full stops and 2 semicolons). The complete frequency signature for the pair (*hat, roll*) is thus (2,3,3,111). Of course, realistic data will have much larger sample sizes, and the marginal frequencies are usually considerably higher than the cooccurrence frequency.

Collocational spans can also be asymmetric, and are generally written in the form (L_k, R_n) for a span of k tokens to the left of the node word and n tokens to its right. The symmetric spans in the example above would be described as $(L4, R4)$. Asymmetric spans introduce an asymmetry between node word and collocate that is absent from most other approaches to collocations. For a one-sided span $(L4, R0)$ to the left of the node word, there would be 2 cooccurrences of the pair (*roll, hat*) in Figure 58.1, but none of the pair (*hat, roll*). A special case are spans of the form $(L0, R1)$, where cooccurrences are ordered pairs of immediately adjacent words, often referred to as bigrams in computational linguistics. Thus, *took place* would be a bigram cooccurrence of the lemma pair (*take, place*), but neither *place taken* nor *take his place* would count as cooccurrences.

3.2. Textual cooccurrence

A second approach considers words to cooccur if they appear in the same textual unit. Typically, such units are sentences or utterances, but with the recent popularity of Google searches and the Web as corpus (see article 18), cooccurrence within (Web) documents has found more widespread use.

One criticism against surface cooccurrence is the arbitrary choice of the span size. For a span size of 3, *throw a birthday party* would be accepted as a cooccurrence of (*throw*, *party*), but *throw a huge birthday party* would not. This is particularly counterintuitive for languages with relatively free word order, where closely related words can be far apart on the surface. In such languages, textual cooccurrence within the same sentence may provide a more appropriate collocational span. Textual cooccurrence also captures weaker dependencies, in particular those caused by paradigmatic semantic relations. For example, if an English sentence contains the noun *bucket*, it is quite likely to contain the noun *mop* as well (although the connection is far weaker than for *water* or *spade*), but the two nouns will not necessarily be near each other in the sentence.

A vast deal of coolness and a peculiar degree of judgement, are
requisite in catching a hat.

A man must not be precipitate, or he runs over it;

he must not rush into the opposite extreme, or he loses it
altogether.

There was a fine gentle wind, and Mr. Pickwick's hat rolled
sportively before it.

The wind puffed, and Mr. Pickwick puffed, and the hat rolled
over and over as merrily as a lively porpoise in a strong tide;

hat	—
—	over
—	—
hat	—
hat	over

Fig. 58.2: Illustration of textual cooccurrence for the word pair (*hat*, *over*)

The definition of textual cooccurrence and the appropriate procedure for computing frequency signatures are illustrated in Figure 58.2, for the word pair (*hat*, *over*) and sentences as textual segments. There is one cooccurrence of *hat* and *over* in the last sentence of this text sample, hence $O = 1$. In contrast to surface cooccurrence, the count is 1 even though there are two instances of *over* in the sentence. Similarly, the marginal frequencies are given by the number of sentences containing each word, ignoring multiple occurrences in the same sentence: hence $f_1 = 3$ and $f_2 = 2$ (although there are three instances each of *hat* and *over* in the text sample). The sample size $N = 5$ is the number of sentences in this case. The complete frequency signature of (*hat*, *over*) is thus (1,3,2,5), whereas for surface cooccurrence within the spans shown in Figure 58.1 it would have been (2,3,3,79).

3.3. Syntactic cooccurrence

In this more restrictive approach, words are only considered to be near each other if there is a direct syntactic relation between them. Examples are a verb and its object (or subject) noun, prenominal adjectives (in English and German) and nominal modifiers

(the pattern N *of* N in English, genitive noun phrases in German). Sometimes, indirect relations might also be of interest, e. g. a verb and the adjectival modifier of its object noun, or a noun and the adjective modifying a postnominal *of*-NP. The latter pattern accounts for several surface collocations of the noun *bucket* such as *a bucket of iced, cold, steaming water* (cf. Table 58.1). Collocations for different types of syntactic relations are usually treated separately. From a given corpus, one might extract a data set of verbs and their object nouns, another data set of verbs and subject nouns, a data set of adjectives modifying nouns, etc. Syntactic cooccurrence is particularly appropriate if there may be long-distance dependencies between collocates: unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time it does not introduce as much “noise” as textual cooccurrence. Syntactic cooccurrence is often used for multiword extraction, since many types of lexicalised multiword expressions tend to appear in specific syntactic patterns such as verb + object noun, adjective + noun, adverb + verb, verb + predicated adjective, delexical verb + noun, etc. (see Bartsch 2004, 11).

In an <i>open barouche</i> [...] stood a <i>stout old gentleman</i> , in a <i>blue coat</i>	open	barouche
and <i>bright buttons</i> , corduroy breeches and top-boots; two	stout	gentleman
<i>young ladies</i> in scarfs and feathers; a <i>young gentleman</i> apparently	old	gentleman
enamoured of one of the <i>young ladies</i> in scarfs and feathers; a lady	blue	coat
of <i>doubtful age</i> , probably the aunt of the aforesaid; and [...]	bright	button
	young	lady
	young	gentleman
	young	lady
	doubtful	age

Fig. 58.3: Illustration of syntactic cooccurrence (nouns modified by prenominal adjectives)

Frequency signatures for syntactic cooccurrence are obtained in a more indirect way, illustrated in Figure 58.3. First, all instances of the desired syntactic relation are identified, in this case modification of nouns by prenominal adjectives. Then the corresponding arguments are compiled into a list with one entry for each instance of the syntactic relation (shown on the right of Figure 58.3). Note that the list entries are lemmatised here, but e. g. case-folded word forms could have been used as well. Just as the original corpus is understood as a sample of language, the list items constitute a sample of the targeted syntactic relation, and Evert (2004) refers to them as “pair tokens”. Cooccurrence frequency data are computed from this sample, while all word tokens that do not occur in the relation of interest are disregarded. For the word pair (*young, gentleman*), we find one cooccurrence in the list of pair tokens, i. e. $O = 1$. The marginal frequencies are given by the total numbers of entries containing one of the component words, $f_1 = 3$ and $f_2 = 3$, and the sample size is the total number of list entries, $N = 9$. The frequency signature of (*young, gentleman*) as a syntactic adjective-noun cooccurrence is thus (1,3,3,9).

3.4. Comparison

Collocations according to surface cooccurrence have proven useful in corpus-linguistic and lexicographic research (cf. Sinclair 1991). They strike a balance between the restricted notion of syntactic cooccurrence (esp. when only a single type of syntactic rela-

tion is considered) and the very broad notion of textual cooccurrence. The number of recurrent word pairs extracted from a corpus is also more manageable than for textual cooccurrence. In this respect, syntactic cooccurrence are even more practical. A popular application of surface cooccurrence in computational linguistics are word space models of distributional semantics (Schütze 1998; Sahlgren 2006). As an alternative to the surface approach, Kilgarriff et al. (2004) collect syntactic collocates from different types of syntactic relations and display them as a *word sketch* of the node word.

Textual cooccurrence is easier to implement than surface cooccurrence, and more robust against certain types of non-randomness such as term clustering, especially when the textual units used are entire documents (cf. the discussion of non-randomness in article 36). However, it tends to create huge data sets of recurrent word pairs that can be challenging even for powerful modern computers.

Syntactic cooccurrence separates collocations of different syntactic types, which are overlaid in frequency data according to surface cooccurrence, and discards many indirect and accidental cooccurrences. It should thus be easier to find suitable association measures to quantify the collocativity of word pairs. Evert (2004, 19) speculates that different measures might be appropriate for different types of syntactic relations. Syntactic cooccurrence is arguably most useful for the identification of multiword expressions, which are typically categorised according to their syntactic structure. However, it requires an accurate syntactic analysis of the source corpus, which will have to be performed with automatic tools in most cases. For prenominal adjectives, the analysis is fairly easy in English and German (Evert/Kermes 2003), while for German verb-object relations, it is extremely difficult to achieve satisfactory results: recent syntactic parsers achieve dependency F-scores of 70%–75% (Schiehlen 2004). Outspoken advocates of syntactic cooccurrence include Daille (1994), Goldman/Nerima/Wehrli (2001), Bartsch (2004) and Evert (2004).

Leaving such practical and philosophical considerations aside, frequency signatures computed according to the different types of cooccurrence can disagree substantially for the same word pair. For example, the frequency signatures of (*short, time*) in the Brown corpus are: (16,135,457,59710) for syntactic cooccurrence (of prenominal adjectives), (27,213,1600,1170811) for (L5, R5) surface cooccurrence, and (32,210,1523,52108) for textual cooccurrence within sentences.

4. Simple association measures

4.1. Expected frequency

It might seem natural to use the cooccurrence frequency O as an association measure to quantify the strength of collocativity (e. g. Choueka 1988). This is not sufficient, however; the marginal frequencies of the individual words also have to be taken into account. To illustrate this point, consider the following example. In the Brown corpus, the bigram *is to* is highly recurrent. With $O = 260$ cooccurrences it is one of the most frequent bigrams in the corpus. However, both components are frequent words themselves: *is* occurs roughly 10,000 times and *to* roughly 26,000 times among 1 million word tokens. If the words in this corpus were rearranged in completely random order, thereby remov-

ing all associations between cooccurring words, we would still expect to see the sequence *is to* approx. 260 times. The high cooccurrence frequency of *is to* therefore does not constitute evidence for a collocation; on the contrary, it indicates that *is* and *to* are not attracted to each other at all. The expected number of cooccurrences for a completely “uncollocational” word pair has been derived by the following reasoning: *to* occurs 26 times every 1,000 words on average. If there is no association between *is* and *to*, then each of the 10,000 instances of *is* in the Brown corpus has a chance of 26/1,000 to be followed by *to*. Therefore, we expect around $10,000 \times (26/1,000) = 260$ occurrences of the bigram *is to*, provided that there is indeed no association between the words. Of course, even in a perfectly randomised corpus there need not be exactly 260 cooccurrences: statistical calculations compute averages across large numbers of samples (formally called *expectations*), while the precise value in a corpus is subject to unpredictable random variation (see article 36).

The complete absence of association, as between words in a randomly shuffled corpus, is called *independence* in mathematical statistics. What we have calculated above is the *expected value* for the number of cooccurrences in a corpus of 1 million words, under the *null hypothesis* that *is* and *to* are independent. In analogy to the *observed frequency* O of a word pair, the expected value under the null hypothesis of independence is denoted E and referred to as the *expected frequency* of the word pair. Expected frequency serves as a reference point for the interpretation of O : the pair is only considered collocational if the observed cooccurrence frequency is substantially greater than the expected frequency, $O \gg E$. Using the formal notation of section 3, the marginal frequencies of (*is*, *to*) are $f_1 = 10,000$ and $f_2 = 26,000$. The sample size is $N = 1,000,000$ tokens, and the observed frequency is $O = 260$. Expected frequency is thus given by the equation $E = f_1 \cdot (f_2 / N) = f_1 f_2 / N = 260$. While the precise calculation of expected frequency is different for each type of cooccurrence, it always follows the basic scheme $f_1 f_2 / N$. For textual and syntactic cooccurrence, the standard formula $E = f_1 f_2 / N$ can be used directly. For surface cooccurrence, an additional factor k represents the total span size, i. e. $E = k f_1 f_2 / N$. This factor is $k = 10$ for a symmetric span of 5 words (L5, R5), $k = 4$ for a span (L3, R1), and $k = 1$ for simple bigrams (L0, R1).

4.2. Essential association measures

A *simple association measure* interprets observed cooccurrence frequency O by comparison with the expected frequency E , and calculates an *association score* as a quantitative measure for the attraction between two words. The most important and widely used simple association measures are shown in Figure 58.4. In the following paragraphs, their mathematical background and some important properties will be explained.

$$\text{MI} = \log_2 \frac{O}{E} \quad \text{MI}^k = \log_2 \frac{O^k}{E} \quad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \quad \text{t-score} = \frac{O - E}{\sqrt{O}} \quad \text{simple-ll} = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

Fig. 58.4: A selection of simple association measures

The most straightforward and intuitive way to relate O and E is to use the ratio O/E as an association measure. For instance, $O/E = 10$ means that the word pair cooccurs 10 times more often than would be expected by chance, indicating a certain degree of collocativity. Since the value of O/E can become extremely high for large sample size (because $E \ll 1$ for many word pairs), it is convenient and sensible to measure association on a (base-2) logarithmic scale. This measure can also be derived from information theory, where it is interpreted as the number of bits of “shared information” between two words and known as (*pointwise*) *mutual information* or simply MI (Church/Hanks 1990, 23). A MI value of 0 bits corresponds to a word pair that cooccurs just as often as expected by chance ($O = E$); 1 bit means twice as often ($O = 2E$), 2 bits mean 4 times as often, 10 bits about 1000 times as often, etc. A negative MI value indicates that a word pair cooccurs less often than expected by chance: half as often for -1 bit, a quarter as often for -2 bits, etc. Thus, negative MI values constitute evidence for a “repulsion” between two words, the pair forming an *anti-collocation*.

The MI measure exemplifies two general *conventions for association scores* that all association measures should adhere to. (i) Higher scores indicate stronger attraction between words, i. e. a greater degree of collocativity. In particular, repulsion, i. e. $O < E$, should result in very low association scores. (ii) Ideally, an association measure should distinguish between *positive* association ($O > E$) and negative association ($O < E$), assigning positive and negative scores, respectively. A strong negative association would thus be indicated by a large negative value. As a consequence, the null hypothesis of independence corresponds to a score of 0 for such association measures. It is easy to see that MI satisfies both conventions: the more O exceeds E , the larger the association score will be; for $O = E$, the MI value is $\log_2 1 = 0$. Most, though not all association measures follow at least the first convention (we will shortly look at an important exception in the form of the simple-ll measure).

In practical applications, MI was found to have a tendency to assign inflated scores to low-frequency word pairs with $E \ll 1$, especially for data from large corpora. Thus, even a single cooccurrence of two word types might result in a fairly high association score. In order to counterbalance this low-frequency bias of MI, various heuristic modifications have been suggested. The most popular one multiplies the denominator with O in order to increase the influence of observed cooccurrence frequency compared to the expected frequency, resulting in the formula $\log_2(O^2/E)$. Multiplication with O can be repeated to strengthen the counterbalancing effect, leading to an entire family of measures MI^k with $k \geq 1$, as shown in Figure 58.4. Common choices for the exponent are $k = 2$ and $k = 3$. Daille (1994) has systematically tested values $k = 2, \dots, 10$ and found $k = 3$ to work best for her purposes. An alternative way to reduce the low-frequency bias of MI is to multiply the entire formula with O , resulting in the *local-MI* measure. Unlike the purely heuristic MI^k family, local-MI can be justified by an information-theoretic argument (Evert 2004, 89) and its value can be interpreted as bits of information. Although not immediately obvious from its equation, local-MI fails to satisfy the first convention for association scores in the case of strong negative association: for fixed expected frequency E , the score reaches a minimum at $O = E/\exp(1)$, and then increases for smaller O . Local-MI distinguishes between positive and negative association, though, and satisfies both conventions if only word pairs with positive association are considered. The measures MI^k satisfy the first convention, but violate the second convention for all $k > 1$. It has been pointed out above that MI assigns high association

scores whenever O exceeds E by a large amount, even if the absolute cooccurrence frequency is as low as $O = 1$ (and $E \ll 1$). In other words, MI only looks at what is known as *effect size* in statistics and does not take into account how much *evidence* the observed data provide. We will return to the distinction between effect-size measures and evidence-based measures in section 6. Here, we introduce three simple association measures from the latter group.

A *z-score* is a standardised measure for the amount of evidence provided by a sample against a simple null hypothesis such as $O = E$ (see article 36). In our case, the general rule for calculating z-scores leads to the equation shown in Figure 58.4. Z-scores were first used by Dennis (1965, 69) as an association measure, and later by Berry-Rogge (1973, 104). They distinguish between positive and negative association: $O > E$ leads to $z > 0$ and $O < E$ to $z < 0$. Z-scores can be interpreted by comparison with a standard normal distribution, providing theoretically motivated cut-off thresholds for the identification of “true collocations”. An absolute value $|z| > 1.96$ is generally considered sufficient to reject the null hypothesis, i. e. to provide significant evidence for a (positive or negative) association; a more conservative threshold is $|z| > 3.29$. When used as an association measure, z-score tends to yield much larger values, though, and most word pairs in a typical data set are highly significant. For instance, 80% of all distinct word bigrams in the Brown corpus have $|z| > 1.96$, and almost 70% have $|z| > 3.29$. Recent studies avoid standard thresholds and use z-scores only to rank word pairs or select n-best lists.

A fundamental problem of the z-score measure is the normal approximation used in its mathematical derivation, which is valid only for sufficiently high expected frequency E . While there is no clearly defined limit value, the approximation becomes very inaccurate if $E < 1$, which is often the case for large sample sizes (e. g., 89% of all bigrams in the Brown corpus have $E < 1$). Violation of the normality assumption leads to highly inflated z-scores and a low-frequency bias similar to the MI measure. In order to avoid this low-frequency bias, various other significance measures have been suggested, based on more “robust” statistical tests. One possibility is the *t-score* measure, which replaces E in the denominator of z-score by O . This measure has been widely used in computational lexicography following its introduction into the field by Church et al. (1991, section 2.2.). See Evert (2004, 82–83) for a criticism of its derivation from the statistical *t* test, which is entirely inappropriate for corpus frequency data.

Dunning (1993) advocated the use of likelihood-ratio tests, which are also more robust against low expected frequencies than z-score. For a simple measure comparing O and E , the likelihood-ratio procedure leads to the *simple-II* equation in Figure 58.4. It can be shown that simple-II scores are always non-negative and violate both conventions for association scores. Because the underlying likelihood-ratio test is a *two-sided* test, the measure does not distinguish between $O \gg E$ and $O \ll E$, assigning high positive scores in both cases. This detail is rarely mentioned in publications and textbooks and may easily be overlooked. A general procedure can be applied to convert a two-sided association measure like simple-II into a one-sided measure that satisfies both conventions: association scores are calculated in the normal way and then multiplied with -1 for all word pairs with $O < E$. This procedure is applicable if association scores of the two-sided measure are always non-negative and high scores are assigned to strong negative associations. For the resulting transformed measure, significance is indicated by the absolute value of an association score, while positive and negative association are distinguished by its sign.

Similar to the z-score measure, simple-II measures significance (i. e. the amount of evidence against the null hypothesis) on a standardised scale, known as a chi-squared distribution with one degree of freedom, or χ_1^2 for short. Theoretically motivated cut-off thresholds corresponding to those for z-scores are $|II| > 3.84$ and $|II| > 10.83$, but the same reservations apply: many word pairs achieve scores far above these thresholds, so that they are not a meaningful criterion for the identification of “true collocations”.

Article 36 gives detailed explanations of statistical concepts such as *significance*, *effect size*, *hypothesis test*, *one-sided* vs. *two-sided* test, *z-score* and *normal distribution* that have been used in this section.

4.3. Simple association measures in a nutshell

The preceding section has introduced a basic selection of simple association measures. These measures quantify the “attraction” between two words, i. e. their statistical association, by comparing observed cooccurrence frequency O against E , the expected frequency under the null hypothesis of independence (i. e. complete absence of association). E is important as a reference point for the interpretation of O , since two frequent words might cooccur quite often purely by chance. Most association measures follow the convention that higher association scores indicate stronger (positive) association. Many measures also differentiate between positive association ($O > E$), indicated by positive scores, and negative association ($O < E$), indicated by negative scores. Two-sided measures fail to make any distinction between positive and negative association, but can be converted into one-sided measures with an explicit test for $O > E$.

The association measures listed in Figure 58.4 offer a number of different angles on collocativity that are sufficient for many purposes. Except for the heuristic MI^k family, all measures have theoretical motivations, allowing a meaningful interpretation of the computed association scores. As has been exemplified with the standard z-score thresholds, one should not put too much weight on such interpretations, though. Cooccurrence data do not always satisfy the assumptions made by statistical hypothesis tests, and heuristic measures may be just as appropriate.

Association measures can be divided into two general groups: measures of *effect size* (MI and MI^k) and measures of *significance* (z-score, t-score and simple-II). The former ask the question “how strongly are the words attracted to each other?” (operationalised as “how much does observed cooccurrence frequency exceed expected frequency?”), while the latter ask “how much evidence is there for a positive association between the words, no matter how small effect size is?” (operationalised as “how unlikely is the null hypothesis that the words are independent?”). The two approaches to measuring association are not entirely unrelated: a word pair with large “true” effect size is also more likely to show significant evidence against the null hypothesis in a sample. However, there is an important difference between the two groups. Effect-size measures typically fail to account for sampling variation and are prone to a low-frequency bias (small E easily leads to spuriously high effect size estimates, even for $O = 1$ or $O = 2$), while significance measures are often prone to a high-frequency bias (if O is sufficiently large, even a small relative difference between O and E , i. e. a small effect size, can be highly significant).

Of the significance measures shown in Figure 58.4, simple-II is the most accurate and robust choice. Z-score has a strong low-frequency bias because the approximations used in its derivation are not valid for $E < 1$, while t-score has been derived from an inappropriate hypothesis test. Nonetheless, t-score has proven useful for certain applications, especially the identification of certain types of multiword expressions (see section 6.2.). It has to be kept in mind that simple-II is a two-sided measure and assigns high scores both to positive and negative associations. If only positive associations are of interest (as is the case for most studies), then word pairs with $O < E$ should be discarded. Alternatively, simple-II can be transformed into a one-sided measure that satisfies both conventions for association scores (by multiplying scores with -1 if a word pair has $O < E$).

Association measures with a background in information theory take a different approach, which at first sight seems appropriate for the interpretation of collocations as mutually predictable word combinations (e.g. Sinclair 1966, 414). They ask the question “to what extent do the occurrences of a word w_1 determine the occurrences of another word w_2 ”, and vice versa, based on the information-theoretic notion of mutual information (MI). Interestingly, different variants of MI lead to measures with entirely different properties: pointwise MI is a measure of effect size, while local-MI is very similar to simple-II and hence has to be considered a measure of significance.

It is probably impossible to choose a single most appropriate association measure (cf. the discussion in section 6). The recommended strategy is therefore to apply simple-II, t-score and MI as proven association measures with well-understood mathematical properties, in order to obtain three entirely different perspectives on the cooccurrence data. MI should always be combined with a frequency threshold to counteract its low-fre-

Tab. 58.2: Collocates of *bucket* in the BNC according to the association measures simple-II, t-score, MI, and MI with frequency threshold $f \geq 5$

collocate	<i>f</i>	<i>f</i> ₂	simple-II	collocate	<i>f</i>	<i>f</i> ₂	t-score
<i>water</i>	184	37012	1083.18	<i>a</i>	590	2164246	15.53
<i>a</i>	590	2164246	449.30	<i>water</i>	184	37012	13.30
<i>spade</i>	31	465	342.31	<i>and</i>	479	2616723	10.14
<i>plastic</i>	36	4375	247.65	<i>with</i>	196	658584	9.38
<i>size</i>	42	14448	203.36	<i>of</i>	497	3040670	8.89
<i>slop</i>	17	166	202.30	<i>the</i>	832	6041238	8.26
<i>mop</i>	20	536	197.68	<i>into</i>	87	157565	7.67
<i>throw</i>	38	11308	194.66	<i>size</i>	42	14448	6.26
<i>fill</i>	37	10722	191.44	<i>in</i>	298	1937966	6.23
<i>with</i>	196	658584	171.78	<i>record</i>	43	29404	6.12

collocate	<i>f</i>	<i>f</i> ₂	MI	collocate	<i>f</i> ≥ 5	<i>f</i> ₂	MI
<i>fourteen-record</i>	4	4	13.31	<i>single-record</i>	5	8	12.63
<i>ten-record</i>	3	3	13.31	<i>randomize</i>	10	57	10.80
<i>multi-record</i>	2	2	13.31	<i>slop</i>	17	166	10.03
<i>two-record</i>	2	2	13.31	<i>spade</i>	31	465	9.41
<i>a-row</i>	1	1	13.31	<i>mop</i>	20	536	8.57
<i>anti-sweat</i>	1	1	13.31	<i>oats</i>	7	286	7.96
<i>axe-blade</i>	1	1	13.31	<i>shovel</i>	8	358	7.83
<i>bastardling</i>	1	1	13.31	<i>rhino</i>	7	326	7.77
<i>dippermouth</i>	1	1	13.31	<i>synonym</i>	7	363	7.62
<i>Dok</i>	1	1	13.31	<i>bucket</i>	18	1356	7.08

Tab. 58.3: Most strongly collocated bigrams in the Brown corpus according to the association measures simple-ll, t-score, MI with frequency threshold $f \geq 10$, and MI with frequency threshold $f \geq 50$

bigram	$f \geq 10$	f_1	f_2	simple-ll
of the	9702	34036	58451	13879.8
in the	6018	19615	58451	9302.3
it is	1482	8409	9415	5612.9
on the	2459	5990	58451	4972.9
United States	395	480	600	4842.6
it was	1338	8409	9339	4831.2
to be	1715	25106	6275	4781.1
had been	760	5107	2460	4599.8
have been	650	3884	2460	4084.0
has been	567	2407	2460	3944.9

bigram	$f \geq 10$	f_1	f_2	t-score
of the	9702	34036	58451	76.30
in the	6018	19615	58451	61.33
on the	2459	5990	58451	41.83
to be	1715	25106	6275	37.23
it is	1482	8409	9415	36.24
it was	1338	8409	9339	34.22
at the	1654	5032	58451	32.72
to the	3478	25106	58451	31.62
from the	1410	4024	58451	30.66
he was	1110	9740	9339	30.32

bigram	$f \geq 10$	f_1	f_2	MI
Hong Kong	11	11	11	16.34
gon na	16	16	16	15.80
Viet Nam	14	16	14	15.80
Simms Purdew	12	16	12	15.80
Pathet Lao	10	10	17	15.71
El Paso	10	19	11	15.41
Lo Shu	21	21	21	15.40
Puerto Rico	21	24	21	15.21
unwed mothers	10	12	26	14.83
carbon tetrachloride	18	30	19	14.81

bigram	$f \geq 50$	f_1	f_2	MI
Los Angeles	50	51	50	14.12
Rhode Island	100	105	175	12.27
Peace Corps	55	171	109	11.39
per cent	146	371	155	11.17
United States	395	480	600	10.29
President Kennedy	54	374	156	9.72
years ago	138	793	246	9.33
fiscal year	58	118	701	9.32
New York	303	1598	309	9.12
United Nations	51	480	175	9.11

quency bias. As an example, and to illustrate the different properties of these association measures, Table 58.2 shows the collocates of *bucket* in the British National Corpus (following the case study in section 2.2.), according to simple-ll, t-score, MI without frequency threshold, and MI with an additional frequency threshold of $f \geq 5$. Table 58.3 gives a second example for word bigrams in the Brown corpus (excluding punctuation). Obviously, simple-ll and especially t-score focus on frequent grammatical patterns like *of the* or *to be*. More interesting bigrams can only be found if separate lists are generated for each part-of-speech combination. The top collocations according to MI, on the other hand, tend to be proper names and other very fixed combinations. Their cooccurrence frequency is often close to the applied frequency threshold.

5. Statistical association measures

The simple association measures introduced in section 4 are convenient and offer a range of different perspectives on collocativity. However, two serious shortcomings make this approach unsatisfactory from a theoretical point of view and may be problematic for certain types of applications. The first of these problems is most easily explained with a worked example. In a corpus of about a million words, you might find that the bigrams A = *the Iliad* and B = *must also* both occur $O = 10$ times, with the same expected frequency $E = 1$. Therefore, any simple measure will assign the same association score to both bigrams. However, bigram A is a combination of a very frequent word (*the* with, say, $f_1 = 100,000$) and an infrequent word (*Iliad* with $f_2 = 10$), while B combines two words of intermediate frequency (*must* and *also* with $f_1 = f_2 = 1,000$). Using the formula $E = f_1 f_2 / N$ from section 4.1., you can easily check that the expected frequency is indeed $E = 1$ for both bigrams. While O exceeds E by the same amount for *the Iliad* as for *must also*, it is intuitively obvious that bigram A is much more strongly connected than bigram

B. In particular, $O = 10$ is the highest cooccurrence frequency that can possibly be observed for these two words (since $O \leq f_1, f_2$): every instance of *Iliad* in the corpus is preceded by an instance of *the*. For bigram B, on the other hand, the words *must* and *also* could have cooccurred much more often than 10 times. One might argue that A should therefore obtain a higher association score than B, at least for certain applications.

The second limitation of simple association measures is of a more theoretical nature. We made use of statistical concepts and methods to define measures with a meaningful interpretation, but did not apply the procedures with full mathematical rigour. In statistical theory, measures of association and tests for the independence of events are always based on a cross-classification of a random sample of certain items. An appropriate representation of cooccurrence frequency data in the form of *contingency tables* is described in section 5.1., with different rules for each type of cooccurrence. Then several widely used statistical association measures are introduced in section 5.2.

We will see in section 6 that simple association measures often give close approximations to the more sophisticated association measures introduced below. Therefore, they are sufficient for many applications, so that the computational and mathematical complexities of the rigorous statistical approach can be avoided.

5.1. Contingency tables

A rigorous statistical approach to measuring association is based on contingency tables representing the cross-classification of a set of items. Such tables naturally take marginal frequencies into account, unlike a simple comparison of O against E . As a first step, we have to define the set of cooccurrence items in a meaningful way, which is different for each type of cooccurrence. Then a separate contingency table is calculated for every word pair (w_1, w_2) , using the presence of w_1 and w_2 in each cooccurrence item as factors for the cross-classification.

	w_2	$\neg w_2$			w_2	$\neg w_2$
w_1	O_{11}	O_{12}	$= R_1$	w_1	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$\neg w_1$	O_{21}	O_{22}	$= R_2$	$\neg w_1$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$
	$= C_1$	$= C_2$	$= N$			

Fig. 58.5: General form of the contingency table of observed frequencies with row and column marginals (left panel), and contingency table of expected frequencies under the null hypothesis of independence (right panel)

The resulting contingency table (left panel of Figure 58.5) has four *cells*, representing the items containing both w_1 and w_2 (O_{11} , equivalent to the observed cooccurrence frequency O), the items containing w_1 but not w_2 (O_{12}), the items containing w_2 but not w_1 (O_{21}), and the items containing neither of the two words (O_{22}). These *observed frequencies* add

up to the total number of items or *sample size*, since every item has to be classified into exactly one cell of the table. The row and column sums, also called *marginal frequencies* (as they are written in the margins of the table), play an important role in the statistical analysis of contingency tables. The first row sum R_1 corresponds to the number of cooccurrence items containing w_1 , and is therefore usually equal to f_1 (except for surface cooccurrence, see below), while the first column sum C_1 is equal to f_2 . This equivalence explains the name “marginal frequencies” for f_1 and f_2 .

As in the case of simple association measures, the statistical analysis of contingency tables is based on a comparison of the observed frequencies O_{ij} with expected frequencies under the null hypothesis that the factors defining rows and columns of the table are statistically independent (which is the mathematical equivalent of the intuitive notion of independence between w_1 and w_2 introduced in section 4.1.). In contrast to the simple approach, we are not only interested in the expected number of cooccurrences of w_1 and w_2 , but have to compute expected frequencies for all four cells of the contingency table, according to the equations shown in the right panel of Figure 58.5. Note that $O_{11} = O$ and $E_{11} = E$, so statistical contingency tables are a genuine extension of the previous approach. The statistical association measures introduced in section 5.2. below are formulated in terms of observed frequencies O_{ij} and expected frequencies E_{ij} , the marginals R_i and C_j , and the sample size N . This standard notation follows Evert (2004) and allows equations to be expressed in a clean and readable form.

The definition of appropriate contingency tables is most straightforward for syntactic cooccurrence. The pair tokens on the right of Figure 58.3 can naturally be interpreted as a set of cooccurrence items. If the first word is w_1 , an item is classified into the first row of the contingency table for the pair (w_1, w_2) , otherwise it is classified into the second row. Likewise, the item is classified into the first column if the second word is w_2 and into the second column if it is not. This procedure is illustrated in the left panel of Figure 58.6. The first row sum R_1 equals the total number of cooccurrence items containing w_1 as first element, and the first column sum equals the number of items containing w_2 as second element. This corresponds to the definition of f_1 and f_2 for syntactic cooccurrence in section 3.3. The example in the right panel of Figure 58.6 shows a contingency table for the word pair *(young, gentleman)* obtained from the sample of adjective-noun cooccurrences in Figure 58.3. Since there are nine instances of adjectival modification of nouns in this toy corpus, the sample size is $N = 9$. There is one cooccurrence of *young* and *gentleman* ($O_{11} = 1$), two items where *gentleman* is modified by another adjective

	$* w_2$	$* \neg w_2$			$* \text{gent.}$	$* \neg \text{gent.}$	
$w_1 *$	O_{11}	O_{12}	$= f_1$		young *	1	2
$\neg w_1 *$	O_{21}	O_{22}			$\neg \text{young} *$	2	4
			$= f_2$				
				$= N$			
					$= 3$		
							$= 9$

Fig. 58.6: Contingency table of observed frequencies for syntactic cooccurrence, with concrete example for the word pair *(young, gentleman)* and the data in Figure 58.3 (right panel)

	$w_2 \in S$	$w_2 \notin S$			
$w_1 \in S$	O_{11}	O_{12}	$= f_1$		
$w_1 \notin S$	O_{21}	O_{22}			
			$= f_2$	$= N$	
					$= 2$
					$= 5$

Fig. 58.7: Contingency table of observed frequencies for textual cooccurrence, with concrete example for the word pair (*hat, over*) and the data in Figure 58.2 (right panel)

($O_{21} = 2$), two items where *young* modifies another noun ($O_{12} = 2$), and four items that contain neither the adjective *young* nor the noun *gentleman* ($O_{22} = 4$).

For textual cooccurrence, Figure 58.2 motivates the definition of cooccurrence items as instances of textual units. In this example, each item corresponds to a sentence of the corpus. The sentence is classified into the first row of the contingency table if it contains one or more instances of w_1 and into the second row otherwise; it is classified into the first column if it contains one or more instances of w_2 and into the second column otherwise (see Figure 58.7). Note that no distinction is made between single and multiple occurrence of w_1 or w_2 in the same sentence. Again, the first row and column sums correspond to the marginal frequencies f_1 and f_2 as defined in section 3.2. The right panel of Figure 58.7 shows a contingency table for the word pair (*hat, over*), based on the example in Figure 58.2. With five sentences in the toy corpus, sample size is $N = 5$. One of the sentences contains both *hat* and *over* ($O_{11} = 1$), two sentences contain *hat* but not *over* ($O_{12} = 2$), one sentence contains *over* but not *hat* ($O_{21} = 1$), and one sentence contains neither of the two words ($O_{22} = 1$).

	w_2	$\neg w_2$			
$near(w_1)$	O_{11}	O_{12}	$\approx k \cdot f_1$		
$\neg near(w_1)$	O_{21}	O_{22}			
			$= f_2$	$= N - f_1$	
					$= 3$
					$= 108$

	roll	\neg roll		
$near(hat)$	2	18	$= 20$	
$\neg near(hat)$	1	87		
			$= 3$	
				$= 108$

Fig. 58.8: Contingency table of observed frequencies for surface cooccurrence, with concrete example for *roll* as a collocate of the node *hat* according to Figure 58.1 (right panel)

The statistical interpretation of surface cooccurrence is less straightforward than for the other two types. The most sensible definition identifies cooccurrence items with the relevant word tokens in the corpus, but excluding instances of the node word w_1 , for which no meaningful cross-classification is possible. Each item, i. e. word token, is then classified into the first row of the contingency table if it cooccurs with the node word w_1 , i. e. if it falls into one of the collocational spans around the instances of w_1 ; it is classified into the second row otherwise. The item is classified into the first column of the table if it is an instance of the targeted collocate w_2 , and into the second column otherwise. The

procedure is illustrated in Figure 58.8, with a concrete example for the data of Figure 58.1 shown in the right panel. This toy corpus consists of 111 word tokens (excluding punctuation). Subtracting the three instances of the node word *hat*, we obtain a sample size of $N = 108$. Of the 108 cooccurrence items, 20 fall into the collocational spans around instances of *hat*, so that the first row sum is $R_1 = 20$. Two of these items are cooccurrences of *hat* and *roll* ($O_{11} = 2$), and the remaining 18 items are classified into the second cell ($O_{12} = 18$). The 88 items outside the collocational spans are classified analogously: there is one instance of the collocate *roll* ($O_{21} = 1$), and all other items are assigned to the last cell of the table ($O_{22} = 87$).

For syntactic and textual cooccurrence, the contingency tables can be calculated directly from frequency signatures (O, f_1, f_2, N) that have been obtained as described in sections 3.2. and 3.3., using the following transformation equalities:

$$\begin{array}{ll} O_{11} = O & O_{12} = f_1 - O \\ O_{21} = f_2 - O & O_{22} = N - f_1 - f_2 + O \end{array}$$

The use of frequency signatures in combination with the equalities above is usually the most practical and convenient implementation of contingency tables. Tables for surface cooccurrence cannot be simplified in the same way, and it is recommended to calculate them by the explicit cross-classification procedure explained above.

5.2. Selected measures

Statistical association measures assume that the set of cooccurrence items is a random sample from a large population (representing an extensional definition of language as the set of all utterances that have been or can be produced, cf. article 36) and attempt to draw inferences about this population. Like simple measures, they can be divided into the general groups of effect-size and significance measures.

Effect-size measures aim to quantify how strongly the words in a pair are attracted to each other, i. e. they measure statistical association between the cross-classifying factors in the contingency table. Liebetrau (1983) gives a comprehensive survey of such association coefficients and Evert (2004, 54–58) discusses their mathematical properties. Coefficients describe properties of a population without taking sampling variation into account. They can be used as association measures in a straightforward way if this fact is ignored and the observed frequencies are taken as direct estimates for the corresponding population probabilities. As a result, effect-size measures tend to be unreliable especially for low-frequency data.

MI is the most intuitive association coefficient, comparing observed cooccurrence frequency against the value expected under the null hypothesis of independence. The equation shown in Figure 58.4 is also meaningful as a statistical association measure, where it should more precisely be written $\log_2(O_{11}/E_{11})$. Two other association coefficients are the (logarithmic) *odds ratio* (Blaheta/Johnson 2001, 56) and the *Dice coefficient* (Smadja/McKeown/Hatzivassiloglou 1996), shown in Figure 58.9. The odds ratio measure satisfies both conventions for association scores, with a value of 0 corresponding to independence and high positive values indicating strong positive association. Its interpretation is less intuitive than that of MI, though, and it has rarely been applied to

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{chi-squared}_{\text{corr}} = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - N/2)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \quad \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1} \quad \text{odds-ratio} = \log \frac{\left(O_{11} + \frac{1}{2}\right)\left(O_{22} + \frac{1}{2}\right)}{\left(O_{12} + \frac{1}{2}\right)\left(O_{21} + \frac{1}{2}\right)}$$

Fig. 58.9: Some widely used statistical association measures

collocations. The Dice coefficient does not adhere to the second convention, as it does not assume a well-defined value in the case of independence. It cannot be used to identify word pairs with strong negative association, but is well-suited for rigid combinations such as fixed multiword units (Smadja/McKeown/Hatzivassiloglou 1996; Dias/Guilloré/Lopes 1999).

Statistical significance measures are based on the same types of hypothesis tests as the simple measures in section 4.2., viz. chi-squared tests (as a generalisation of z-scores) and likelihood-ratio tests. Unsurprisingly, there is no counterpart for t-score, which was based on an inappropriate test and hence cannot be translated into a rigorous statistical measure. The *chi-squared* measure adds up squared z-scores for all cells of the contingency table (Σ_{ij} indicates summation over all four cells, i. e. over indices $ij = 11, 12, 21, 22$). The normal approximation implicit in the z-scores becomes inaccurate if any of the expected frequencies E_{ij} are small, and chi-squared exhibits a low-frequency bias similar to the z-score measure. A better approximation is obtained by applying *Yates' continuity correction* (cf. DeGroot/Schervish 2002, section 5.8.). The continuity-corrected version is often written in the compact form shown as $\text{chi-squared}_{\text{corr}}$ in Figure 58.9, without explicit reference to expected frequencies E_{ij} . Chi-squared is a two-sided measure because the squared values are always positive. It can be transformed into a one-sided measure using the general procedure introduced in section 4.2. Chi-squared is often abbreviated X^2 , the symbol used for the chi-squared test statistic in mathematical statistics.

The *log-likelihood* measure (Dunning 1993) is a straightforward extension of simple-ll, replacing the term $O-E$ by a summation over the remaining three cells of the contingency table. It is a two-sided measure and is sometimes abbreviated G^2 in analogy to X^2 . Interestingly, the association scores of log-likelihood, simple-ll and chi-squared are all interpreted against the same scale, a χ^2 distribution (cf. section 4.2.). Mathematicians generally agree that the most appropriate significance test for contingency tables is *Fisher's exact test* (Agresti 2002, 91–93), which was put forward by Pedersen (1996) as an alternative to the log-likelihood measure. Unlike chi-squared and likelihood-ratio tests, this exact test does not rely on approximations that may be invalid for low-frequency data. Fisher's test can be applied as a one-sided or two-sided measure and provides a useful reference point for the discussion of other significance measures. However, it is computationally expensive and a sophisticated implementation is necessary to avoid numerical instabilities (Evert 2004, 93). Section 6.1. shows that log-likelihood provides an excellent approximation to association scores computed by Fisher's test, so there is little reason to use the complicated and technically demanding exact test. The informa-

tion-theoretic measure *average-MI* is identical to log-likelihood (except for a constant factor) and need not be discussed further here.

Note that simple association measures can also be computed from the full contingency tables, replacing O by O_{11} and E by E_{11} in the equations given in Figure 58.4. This shows clearly that many simple measures can be understood as a simplified version (or approximation) of a corresponding statistical measure. A more comprehensive list of association measures with further explanations can be found in Evert (2004, section 3) and online at:

<http://www.collocations.de/AM/>

Both resources describe simple as well as statistical association measures, using the notation for contingency tables introduced in this section and summarised in Figure 58.5.

6. Finding the right measure

The twelve equations in Figures 58.4 and 58.9 represent just a small selection of the many association measures that have been suggested and used over the years. Evert (2004) discusses more than 30 different measures, Pecina (2005) lists 57 measures, and new measures and variants are constantly being invented. While some measures have been established as de facto standards, e. g. log-likelihood in computational linguistics, t-score and MI in computational lexicography, there is no ideal association measure for all purposes. Different measures highlight different aspects of collocativity and will hence be more or less appropriate for different tasks: the n-best lists in Tables 58.2 and 58.3 are a case in point. The goal of this section is to help researchers choose a suitable association measure (or set of measures) for their study. While the primary focus is on understanding the characteristic properties of the measures presented in this article and the differences between them, the methods introduced below can also be applied to other association measures, allowing researchers to make an informed choice from the full range of options.

6.1. Mathematical arguments

Theoretical discussions of association measures are usually concerned with their mathematical derivation: the assumptions of the underlying model, the theoretical quantity to be measured, the validity and accuracy of the procedures used (especially if approximations are involved), and general mathematical properties of the measures (such as a bias towards low- or high-frequency word pairs). A first step in such discussions is to collect association measures with the same theoretical basis into groups. Measures within each group can often be compared directly with respect to their mathematical properties (since ideally they should measure the same theoretical quantity and hence lead to the same results), while different groups can only be compared at a general and rather philosophical level (does it make more sense to measure effect size or significance of association?).

As has already been mentioned in sections 4 and 5, the association measures introduced in this article fall into two major groups: *effect-size measures* (MI, Dice, odds ratio) and *significance measures* (z-score, t-score, simple-II, chi-squared, log-likelihood). The choice between these two groups is largely a philosophical issue: one cannot be considered “better” than the other. Instead, they highlight different aspects of collocativity and are plagued by different types of mathematical problems.

Significance measures are particularly amenable to mathematical discussions, since in principle they attempt to measure the same theoretical quantity: the amount of evidence provided by a sample against the null hypothesis of independence. Moreover, chi-squared, log-likelihood and simple-II use the same scale (the χ_1^2 distribution), so that their scores are immediately comparable. While z-score and t-score use a scale based on the normal distribution, their scores can easily be transformed to the χ_1^2 scale. The long-standing debate in mathematical statistics over appropriate significance tests for contingency tables has not completely been resolved yet (see Yates 1984), but most researchers consider Fisher’s exact test to be the most sensible and accurate measure of significance (Yates 1984, 446). We will therefore use it as a reference point for the comparison of association measures in the significance group. Fisher’s test calculates so-called p-values (cf. article 36), which are also transformed to the χ_1^2 scale for the comparison. The scatterplots in Figure 58.10 compare association scores calculated by various significance measures with those of Fisher’s exact test, using a synthetic data set in which cooccurrence and marginal frequencies have been varied systematically. The log-likelihood measure (G^2) and to some extent also simple-II (G^2_{simple}) give an excellent approximation to Fisher’s test, as all data points are close to the diagonal. Chi-squared and z-score overestimate significance drastically (points far above diagonal), while t-score underestimates significance to a similar degree (points far below diagonal).

For effect-size measures, there is no well-defined theoretical quantity that would allow a direct comparison of their scores (e.g. with scatterplots as in Figure 58.10). Numerous coefficients have been suggested as measures of association strength in the population, but statisticians do not agree on a theoretically satisfactory choice (see e.g. Liebetrau 1983). A common mathematical property of effect-size measures is the use of direct estimates that do not take sampling variation into account. As a result, association

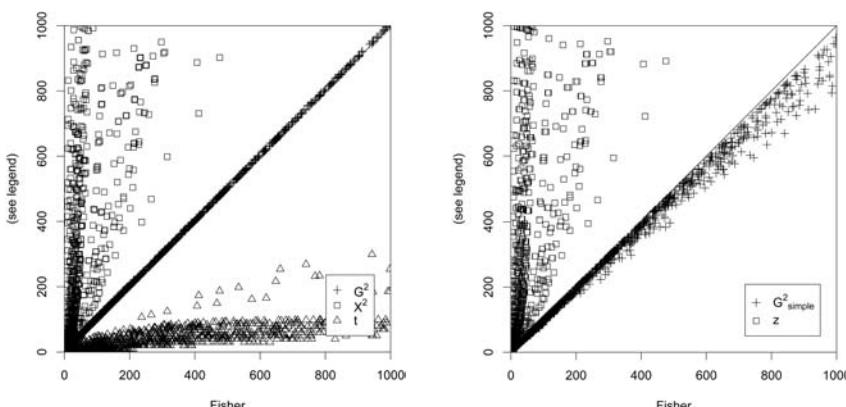


Fig. 58.10: Direct comparison of association scores on a synthetic data set, using Fisher’s exact test as a reference point (scores are transformed to χ_1^2 scale)

scores tend to become unreliable for low-frequency data. This effect is particularly severe for MI, odds ratio and similar measures that compare observed and expected frequency, since $E_{11} \ll 1$ for many low-frequency word pairs. Extending effect-size measures with a correction for sampling variation is a current topic of research and is expected to bridge the gap between the effect-size and significance groups (see section 7.1.).

It should be emphasised that despite their mathematical shortcomings, measures such as chi-squared and t-score may have linguistic merits that justify their use as heuristic measures for collocation identification. While clearly not satisfactory as measures of significance, they must not completely be excluded from the following discussion, which focuses on empirical and intuitive properties of association measures.

6.2. Collocations and multiword extraction

In those cases where mathematical theory does not help us choose between association measures, we can study their empirical properties independent of the underlying statistical reasoning. In this section, we specifically address empirical *linguistic* properties, i. e. we ask what kinds of word pairs are identified as collocations by the different association measures. A simple approach is to look at n-best lists as shown in Tables 58.2 and 58.3, which give a good impression of the different linguistic aspects of collocativity that the association measures capture. For instance, Table 58.2 indicates that simple-ll is a useful measure for identifying typical and intuitively plausible collocates of a node word. Without a frequency threshold, MI brings up highly specialised terms (**-record bucket*), but also many obviously accidental cooccurrences (such as *dippermouth* or *Dok*). A more thorough and systematic study along these lines has been carried out by Stubbs (1995).

More precise empirical statements than such impressionistic case studies can be made if there is a well-defined goal or application for the identified collocations. A particularly profitable setting is the use of association scores for multiword extraction, where the goal is usually to identify a particular subtype of multiword expressions, e. g. compounds (Schone/Jurafsky 2001), technical terminology (Daille 1994) or lexical collocations (Krenn 2000). Evert/Krenn (2001, 2005) suggest an evaluation methodology for such tasks that allows fine-grained quantitative comparisons between a large number of association measures. The evaluation follows the standard procedure for semi-automatic multiword extraction, where recurrent word pairs are obtained from a corpus, optionally filtered by frequency or other criteria, and ranked according to a selected association measure. Since there are no meaningful absolute thresholds for association scores (cf. section 2.1.), it is standard practice to select an n-best list of the 500, 1000 or 2000 highest-ranking collocations as candidate multiword expressions. The candidates are then validated by an expert, e. g. a professional lexicographer or terminologist.

In the evaluation setting, candidates in the n-best list are manually annotated as *true positives* (i. e. multiword expressions of the desired type) and *false positives*. These annotations are used to calculate the *precision* of the n-best list, i. e. the proportion of true positives among the n multiword candidates, and sometimes also *recall*, i. e. how many of all suitable multiword expressions that could have been extracted from the corpus are actually found in the n-best list. The precision values of different association measures can then be compared: the higher the precision of a measure, the better it is

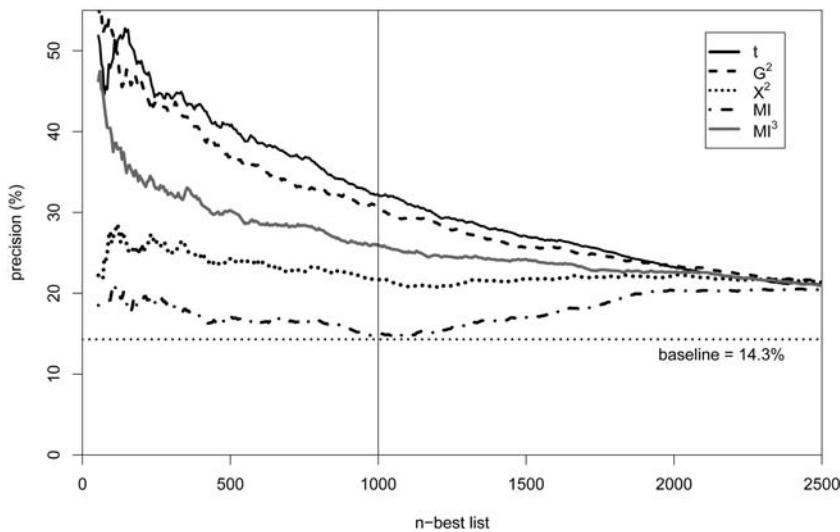


Fig. 58.11: Comparative evaluation of the association measures t-score (t), log-likelihood (G^2), chi-squared (X^2), MI and MI^3 on the data set of Krenn (2000)

suited for identifying the relevant type of multiword expressions. Such evaluation experiments could be used, e.g., to confirm our impression that MI reliably identifies multiword proper names among adjacent bigrams (Table 58.3).

Instead of large and confusing tables listing precision values for various association measures and n-best lists, evaluation results can be presented in a more intuitive graphical form as *precision plots*. Figure 58.11 illustrates this evaluation methodology for the data set of Krenn (2000), who uses PP-verb cooccurrences from an 8-million-word subset of the German *Frankfurter Rundschau* newspaper corpus to identify lexical collocations between prepositional phrases and verbs (including support verb constructions and figurative expressions). The lines in Figure 58.11 summarise the precision values of five different association measures for arbitrary n-best lists. The precision for a particular n-best list can easily be read off from the graph, as indicated by the thin vertical line for $n = 1000$: the solid line at the top shows that t-score achieves a precision of approx. 32 % on the 1000-best list, while log-likelihood (the dashed line below) achieves only 30.5 %. The precision of chi-squared (dotted line) is much lower at 21.5 %. Looking at the full lines, we see that log-likelihood performs much better than chi-squared for all n-best lists, as predicted by the mathematical discussion in section 6.1. Despite the frequency threshold, MI performs worse than all other measures and is close to the *baseline precision* (dotted horizontal line) corresponding to a random selection of candidates among all recurrent word pairs. Evaluation results always have to be interpreted in comparison to the baseline, and an association measure can only be considered useful if it achieves substantially better precision. The most striking result is that t-score outperforms all other measures, despite its mathematical shortcomings. This illustrates the limitations of a purely theoretical discussion: empirically, t-score is the best indicator for lexical PP-verb collocations among all association measures.

6.3. An intuitive geometrical model

In the previous section, we have looked at “linguistic” properties of association measures, viz. how accurately they can identify a particular type of multiword expressions or one of the other linguistic phenomena behind collocativity (see section 2.2.). If we take a pre-theoretic view of collocations as an observable property of language, though, the purpose of association scores is to measure this property in an appropriate way, not to match theoretical linguistic concepts. In this context, evaluation studies that depend on a theoretical or intuitive definition of true positives seem less appropriate. Instead, our goal should be to understand which quantitative aspects of collocativity each association measure singles out: we are interested in empirical mathematical properties of the measures.

Evert (2004, section 3.4.) proposes a geometric visualisation technique in order to reach the desired intuitive understanding of association measures. This technique works well for simple measures that require only two real numbers, O and E , to calculate an association score for a given word pair. Interpreting the numbers (O, E) as two-dimensional coordinates, we can thus represent each word pair in a data set by a point in the real Euclidean plane. The left panel of Figure 58.12 illustrates this “point cloud” view for adjacent bigrams in the Brown corpus. The data point representing the bigram *New York* (with $O = 303$ and $E \approx 0.54$) is marked with a circle. Its expected frequency $E \approx 0.5$ can be read off the x-axis, and its observed frequency $O = 303$ off the y-axis, as indicated by the thin horizontal and vertical lines. Note that both axes are on logarithmic scales in order to accommodate the wide range of observed and expected frequencies found in a typical data set. The frequency threshold $f \geq 10$ applied to the data set is clearly visible in the graph.

Association scores are usually compared against a cutoff threshold, whose value is either established in advance (in a threshold approach) or determined interactively (for n-best lists). In terms of the geometric model, the point cloud representing a data set is divided into *accepted* and *rejected* points by such a cutoff threshold. For any given association measure and cutoff threshold, this decision only depends on the coordinates of

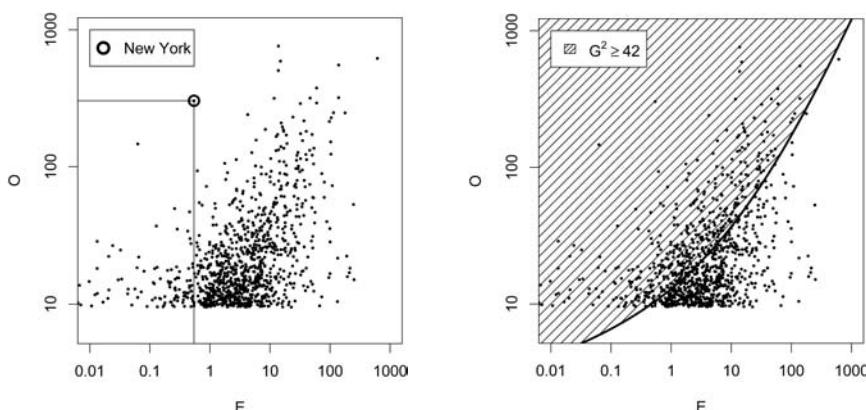


Fig. 58.12: Geometric visualisation of cooccurrence frequency data (left panel) and an acceptance region of the simple-ll association measure (right panel)

a point in the Euclidean plane, not on the word pair represented by the point. It is therefore possible to determine for any hypothetical point in the plane whether it would be accepted or rejected, i.e. whether the association score would be higher than the threshold or not. The right panel of Figure 58.12 shows an illustration for the simple-II measure and a cutoff threshold of 42. Any data point in the shaded region will be assigned a score $G^2 \geq 42$, and any point outside the region a score $G^2 < 42$.

It can be shown that for most association measures the set of accepted hypothetical points forms a simple connected *acceptance region*. The region is bounded below by a single increasing line referred to as a *contour* of the association measure. All points on a contour line have the same association score according to this measure; in our example, a simple-II score of 42. Every simple association measure is uniquely characterised by its contour lines for different threshold values. We can thus visualise and compare measures in the form of contour plots as shown in Figure 58.13. Each panel overlays the contour plots of two different association measures. Comparing the shapes of the contour lines, we can identify the characteristic mathematical properties of the measures and understand the differences between them. Reading contour plots takes some practice: keep in mind that contours connect points with the same association scores, just as the contour lines of a topographic map connect points of the same elevation.

MI only considers the ratio between O and E , even for very low observed frequency O . Hence its dashed contours in Figure 58.13 are straight lines. These straight lines of constant ratio O/E also provide a grid for the interpretation of other contour plots. A significance measure such as simple-II (left panel) is sensitive to the smaller amount of evidence provided by low-frequency data. Therefore, a higher ratio between O and E is required to achieve the same score, and the contour lines have a left curvature. There is a single straight contour line, which marks the null hypothesis of independence ($O = E$) and coincides with the corresponding contour line of MI. Contours for positive association are located above and to the left of the independence line. Contours for negative association show a right curvature and are located below and to the right of the independence line.

The centre panel of Figure 58.13 shows a contour plot for the t-score measure. Again, independence is marked by a straight line that coincides with the MI contour. For positive association, the t-score contours have a left curvature similar to simple-II, but much more pronounced. For very small expected frequencies, they flatten out to horizontal lines, creating an implicit frequency threshold effect. We may speculate that this implicit

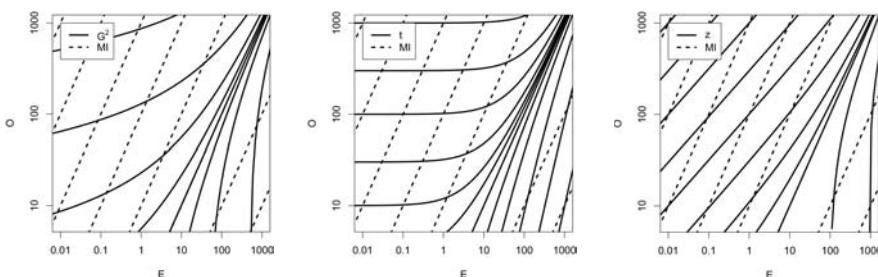


Fig. 58.13: Intuitive comparison of simple association measures represented by contour plots. The three panels compare simple-II (G^2 , left panel), t-score (centre panel) and z-score (right panel) against MI (dashed lines)

threshold is responsible for the good performance of t-score in some evaluation studies, especially if low-frequency word pairs are not discarded in advance. Interestingly, the contour lines for negative association are nearly parallel and do not seem to take random variation into account, in contrast to simple-II.

Finally, the right panel shows contour lines for z-score. Despite its mathematical background as a significance measure, z-score fails to discount low-frequency data. The contour lines for positive association are nearly parallel, although their slope is less steep than for MI. Thus, even data points with low observed frequency O can easily achieve high association scores, explaining the low-frequency bias of z-score that has been noted repeatedly. Interestingly, z-score seems to work well as a measure of significance for negative association, where its contour lines are very similar to those of simple-II.

The visualisation technique presented in this section can be extended to statistical association measures, but the geometric interpretation is more difficult and requires three-dimensional plots. See Evert (2004, section 3.4.) for details and sample plots.

7. Summary and conclusion

In this article, we have been concerned with the empirical Firthian notion of *collocations* as observations on the combinatorics of words in a language, which have to be distinguished clearly from lexicalised *multiword expressions* as pre-fabricated units, and in particular from *lexical collocations*, a subtype of multiword expressions. From the perspective of theoretical linguistics, collocations are often understood as an *epiphenomenon*, the surface reflections of compounds, idioms, lexical collocations and other types of multiword expressions, selectional preferences, semantic restrictions, cultural stereotypes, and to a considerable extent also conceptual knowledge (“facts of life”).

Introduced as an intuitively appealing, but fuzzy and pre-theoretical notion by Firth (1957), collocativity can be operationalised in terms of *cooccurrence* frequencies and quantified by mathematical *association measures*. High association scores indicate strong attraction between two words, but there is no standard scale of measurement to draw a clear distinction between collocations and non-collocations. Association measures and collocations have many uses, ranging from technical applications in computational linguistics to lexicographic and linguistic studies, where they provide descriptive generalisations about word usage. Collocations are closely related to lexicalised multiword expressions, and association measures are central to the task of automatic multiword extraction from corpora.

In order to identify and score collocations from a given corpus, the following steps have to be performed: (1) Choose an appropriate *type of cooccurrence* (surface, textual or syntactic). (2) Determine *frequency signatures* (i. e. cooccurrence frequency f and the marginal frequencies f_1 and f_2 in the corpus) for all relevant word pairs (w_1, w_2) as described in section 3 (Figures 58.1, 58.2 and 58.3 serve as a reminder), as well as sample size N . (3) Filter the cooccurrence data set by applying a *frequency threshold*. Theoretical considerations suggest a minimal threshold of $f \geq 3$ or $f \geq 5$, but higher thresholds often lead to even better results in practice. (4) Calculate the *expected frequencies* of the word pairs, using the general equation $E = f_1 f_2 / N$ for textual and syntactic cooccurrence, and the approximation $E = k f_1 f_2 / N$ for surface cooccurrence, where k is the total span size.

(5) Apply one of the *simple association measures* shown in Figure 58.4, or produce multiple tables according to different measures. Recall that the cooccurrence frequency f is denoted by O (for *observed frequency*) in these equations. (5) If collocations are treated as units, *rank* the word pairs by association score, or select a threshold to distinguish between collocations and non-collocations (or “strong” and “weak” collocations). In the node-collocate view, collocates w_2 are ranked separately for each node word w_1 .

If the data include word pairs with highly skewed marginal frequencies and you suspect that this may have distorted the results of the collocation analysis, you may want to apply *statistical association measures* instead of the simple measures. In order to do so, you have to compute a full 2×2 contingency table for each word pair, as well as a corresponding table of expected frequencies (see Figure 58.5). The precise calculation procedure depends on the type of cooccurrence and is detailed in section 5.1. (Figures 58.6, 58.7 and 58.8 serve as quick reminders). Then, one or more of the statistical measures in Figure 58.9 can be applied. Many further measures are found in Evert (2004) and online at <http://www.collocations.de/AM/> (both resources use the same notation as in this article).

The resulting set or ranking of collocations depends on many parameters, including the size and composition of the corpus, pre-processing (such as lemmatisation), application of frequency thresholds, the definition of cooccurrence used, and the choice of association measure. It is up to the researcher to find a suitable and meaningful combination of parameters, or to draw on results from multiple parameter settings in order to highlight different aspects of collocativity. While a particular type of cooccurrence is often dictated by the theoretical background of a study or practical restrictions (e.g., syntactic cooccurrence requires sufficiently accurate software for automatic syntactic analysis, or a pre-parsed corpus), other parameter values are more difficult to choose (e.g. span size for surface cooccurrence, or the frequency threshold).

A crucial step, of course, is to select one of well over 50 different association measures that are currently available (or to invent yet another measure). At this point, no definitive recommendation can be made. It is perhaps better to apply several measures with well-understood and distinct properties than attempt to find a single optimal choice. In any case, a thorough understanding of the characteristic properties of association measures and the differences between them is essential for a meaningful interpretation of the extracted collocations and their rankings. In section 6, various theoretical and empirical techniques have been introduced for this purpose, and the properties of several widely used measures have been discussed.

7.1. Open questions and extensions

The goal of this article was to present the current state of the art with regard to collocations and association measures. The focus has therefore been on established results rather than unsolved problems, open research questions, or extensions beyond simple word pairs. The following paragraphs give an overview of important topics of current research.

Like all statistical approaches in corpus linguistics, association measures suffer from the fact that the assumptions of their underlying statistical models are usually not met

by corpus data. In addition to the general question whether any finite corpus can be representative of a language (which is a precondition for the validity of statistical generalisations), *non-randomness* of corpus frequency data is a particularly serious problem for all statistical models based on random samples. A thorough discussion of this problem and possible solutions can be found in article 36 and in Evert (2006).

In addition to these common issues, cooccurrence data pose two specific difficulties. First, the null hypothesis of independence is extremely unrealistic. Words are never combined at random in natural language, being subject to a variety of syntactic, semantic and lexical restrictions. For a large corpus, even a small deviation from the null hypothesis may lead to highly significant rejection and inflated association scores calculated by significance measures. Effect-size measures are also subject to this problem and will produce inflated scores, e. g. for two rare words that always occur near each other (such as *déjà* and *vu*). A possible solution would be to specify a more realistic null hypothesis that takes some of the restrictions on word combinatorics into account, but research along these lines is still at a very early stage.

Second, word frequency distributions are highly skewed, with few very frequent types and a large number of extremely rare types. This property of natural language, often referred to as *Zipf's law* (see articles 37 and 41), is even more pronounced for cooccurrence data. In combination with the quantisation of observed frequencies (it is impossible to observe $O = 0.7$ cooccurrences), Zipf's law invalidates statistical corrections for sampling variation to the extent that accidental cooccurrences between low-frequency words may achieve very high association scores. An extensive study of this effect has resulted in the recommendation to apply a frequency threshold of $f \geq 5$ in order to weed out potentially spurious collocations (Evert 2004, chapter 4). Non-randomness effects may exacerbate the situation and necessitate even higher thresholds. Current research based on more sophisticated models of Zipfian frequency distributions aims to develop better correction techniques that are less drastic than a simple frequency threshold.

Intuitively, "mutual expectancies" often hold between more than two words. This is particularly obvious in the case of multiword expressions: *kick ... bucket* is always accompanied by the definite article *the*, *humble pie* usually occurs with *eat*, and the bigram *New York* is often followed by *City*. Applying association measures to word pairs will only bring up fragments of such larger collocations, and the missing pieces have to be filled in from the intuition of a linguist or lexicographer. It is therefore desirable to develop suitable measures for word triples and larger n -tuples. First attempts to formulate such measures are straightforward generalisations of the equations of MI (Lin 1998), log-likelihood (Zinsmeister/Heid 2003), or the Dice coefficient (da Silva/Lopes 1999). Obviously, a deep understanding of the mathematical properties of association measures for word pairs as well as their shortcomings is essential for a successful extension.

With the extension to n -word collocations, regular patterns become more noticeable: in addition to the well-known collocation *carry emotional baggage*, we also find *carry cultural, historical, ideological, intellectual, political, ... baggage* (some of them even more frequent than *emotional baggage*). This evidence suggests a productive *collocational pattern* of the form *carry Adj baggage*, with additional semantic restrictions on the adjective. Many instances of such patterns are too rare to be identified in corpora by statistical means, but would intuitively be considered as collocations by human speakers (think of *carry phraseological baggage*, for instance). There has been little systematic research on the productivity of collocations so far, notable exceptions being Lüdeling/Bosch (2003) and Stevenson/Fazly/North (2004).

Many collocations are intuitively felt to be *asymmetric*. For instance, in the bigram *the Iliad*, *the* is a more important collocate for *Iliad* than *Iliad* is for *the*. In the terminology of Kjellmer (1991), the bigram is left-predictive, but not right-predictive. Although such asymmetries are often reflected in skewed marginal frequencies (the collocation being more important for the less frequent word), hardly any of the known association measures make use of this information. Preliminary research suggests that measures of *directed association* could be based on the ratios O/f_1 and O/f_2 (as estimators for the conditional probability that w_1 is accompanied by w_2 and vice versa), or could be formulated by putting the association score of a word pair (w_1, w_2) in relation to the scores of all collocates of w_1 and w_2 , respectively (Michelbacher/Evert/Schütze 2007).

Although many association measures are available, there is still room for improvement and it would be desirable to develop measures with novel properties. Most existing measures fall into one of two major groups, viz. effect-size and significance measures. Both groups have their strengths and weaknesses: effect-size measures do not correct for sampling variation, while significance measures are biased towards high-frequency word pairs with small effect sizes (which tend to be uninteresting from a linguistic point of view). New association measures might be able to combine aspects of effect-size and significance measures, striking a balance between the low-frequency bias of the former and the high-frequency bias of the latter. First steps in this direction are summarised by Evert (2004, section 3.1.8.), but have not led to satisfactory results yet.

7.2. Further reading

Evert (2004) gives a more detailed account of statistical models for association in contingency tables and their limitations, together with a comprehensive inventory of association measures and methods for the comparison and evaluation of different measures. An online version of the inventory can be found at <http://www.collocations.de/AM/>. Contingency tables and the statistical tests that form the basis of many association measures are explained in standard textbooks on mathematical statistics (e.g. DeGroot/Schervish 2002). Advanced books (e.g. Agresti 2002) introduce more sophisticated models for the analysis of contingency tables. Although these models have not found widespread use as association measures yet, they may become important for the development of novel measures and their extension beyond simple word pairs.

Bartsch (2004) offers an insightful theoretical discussion of collocations and their properties, as well as an excellent overview of the various empirical and phraseological definitions of the term. Exemplary proponents of the two views are Sinclair (1991) and Sinclair et al. (2004) on the empirical side, and standard textbooks (e.g. Burger/Buhofer/Sialm 1982) for the phraseological view. Current research on collocations and multiword expressions is collected in the proceedings of ACL Workshops on Multiword Expressions (Daille/Williams 2001; Levin/Tokunaga/Lenci 2003; Tanaka et al. 2004; Villada Moirón et al. 2006; Grégoire/Evert/Kim 2007) and in Grossmann/Tutin (2003).

Relevant articles in this volume are article 24 (on word segmentation and part-of-speech tagging), article 25 (on lemmatisation), article 26 (on word sense disambiguation) and article 28 (on automatic syntactic annotation), as well as article 10 (on text corpora). Article 36 is a general introduction to the statistical analysis of corpus frequency data,

including most of the techniques on which association measures are based. Important applications of collocations can be found in the articles on computational lexicography (article 8) and word meaning (article 45).

We have followed a traditional view of collocations as simple word pairs here, but association measures and related techniques can equally well be applied to cooccurrences of other linguistic units (e. g. lexical items and syntactic constructions in article 43).

8. Literature

- Agresti, Alan (2002), *Categorical Data Analysis*, 2nd edition. Hoboken: John Wiley & Sons.
- Aston, Guy/Burnard, Lou (1998), *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press. See also the BNC homepage at <http://www.natcorp.ox.ac.uk/>.
- Bartsch, Sabine (2004), *Structural and Functional Properties of Collocations in English*. Tübingen: Narr.
- Berry-Rogghe, Godelieve L. M. (1973), The Computation of Collocations and their Relevance to Lexical Studies. In: Aitken, Adam J./Bailey, Richard W./Hamilton-Smith, Neil (eds.), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press, 103–112.
- Blaheta, Don/Johnson, Mark (2001), Unsupervised Learning of Multi-word Verbs. In: *Proceedings of the ACL Workshop on Collocations*. Toulouse, France, 54–60.
- Burger, Harald/Buhofer, Annelies/Salm, Ambros (1982), *Handbuch der Phraseologie*. Berlin etc.: Walter de Gruyter.
- Choueka, Yaacov (1988), Looking for Needles in a Haystack. In: *Proceedings of RIAO '88*. Cambridge, MA, 609–623.
- Church, Kenneth W./Hanks, Patrick (1990), Word Association Norms, Mutual Information, and Lexicography. In: *Computational Linguistics* 16(1), 22–29.
- Church, Kenneth/Gale, William A./Hanks, Patrick/Hindle, Donald (1991), Using Statistics in Lexical Analysis. In: Zernick, Uri (ed.) *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Hillsdale, NY: Lawrence Erlbaum, 115–164.
- da Silva, Joaquim Ferreira/Lopes, Gabriel Pereira (1999), A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multi-word Units from Corpora. In: *6th Meeting on the Mathematics of Language*. Orlando, FL, 369–381.
- Daille, Béatrice (1994), *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.
- Daille, Béatrice/Williams, Geoffrey (eds.) (2001), *Proceedings of the 2001 ACL Workshop on Collocation*. Toulouse, France.
- DeGroot, Morris H./Schervish, Mark J. (2002), *Probability and Statistics*, 3rd edition. Boston: Addison Wesley.
- Dennis, Sally F. (1965), The Construction of a Thesaurus Automatically from a Sample of Text. In: Stevens, Mary E./Giuliano, Vincent E./Heilprin, Lawrence B. (eds.), *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation*. (National Bureau of Standards Miscellaneous Publication 269.) Washington: National Bureau of Standards, 61–148.
- Dias, Gaël/Guilloré, Sylvie/Lopes, José G. P. (1999), Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. In: *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*. Cargèse, Corsica, France, 333–338.
- Dunning, Ted E. (1993), Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* 19(1), 61–74.
- Evert, Stefan (2004), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714.

- Evert, Stefan (2006), How Random is a Corpus? The Library Metaphor. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177–190.
- Evert, Stefan/Kermes, Hannah (2003), Experiments on Candidate Data for Collocation Extraction. In: *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, 83–86.
- Evert, Stefan/Krenn, Brigitte (2001), Methods for the Qualitative Evaluation of Lexical Association Measures. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 188–195.
- Evert, Stefan/Krenn, Brigitte (2005), Using Small Random Samples for the Manual Evaluation of Statistical Association Measures. In: *Computer Speech and Language* 19(4), 450–466.
- Firth, John Rupert (1957), A Synopsis of Linguistic Theory 1930–55. In: *Studies in Linguistic Analysis*. Oxford: The Philological Society, 1–32. Reprinted in Palmer 1968, 168–205.
- Gil, Alexandre/Dias, Gaël (2003), Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora. In: *Proceedings of the ACL Workshop on Multiword Expressions*. Sapporo, Japan, 25–32.
- Goldman, Jean-Philippe/Nerima, Luka/Wehrli, Eric (2001), Collocation Extraction Using a Syntactic Parser. In: *Proceedings of the ACL Workshop on Collocations*. Toulouse, France, 61–66.
- Grégoire, Nicole/Evert, Stefan/Kim, Su Nam (eds.) (2007), *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Prague, Czech Republic.
- Grossmann, Francis/Tutin, Agnès (eds.) (2003), *Les collocations: Analyse et traitement*. Amsterdam: De Werelt.
- Hausmann, Franz Josef (1989), Le dictionnaire de collocations. In: Hausmann, Franz Josef/Reichmann, Otto/Wiegand, Herbert Ernst (eds.) *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch*. Berlin: Mouton de Gruyter, 1010–1019.
- Heid, Ulrich/Evert, Stefan/Docherty, Vincent/Worsch, Wolfgang/Wermke, Matthias (2000), A Data Collection for Semi-automatic Corpus-based Updating of Dictionaries. In: Heid, Ulrich/Evert, Stefan/Lehmann, Egbert/Rohrer, Christian (eds.), *Proceedings of the 9th EURALEX International Congress*. Stuttgart, Germany, 183–195.
- Kilgarriff, Adam/Rychly, Pavel/Smrz, Pavel/Tugwell, David (2004), The Sketch Engine. In: *Proceedings of the 11th EURALEX International Congress*. Lorient, France, 105–116.
- Kjellmer, Göran (1991), A Mint of Phrases. In: Aijmer, Karin/Altenberg, Bengt (eds.), *English Corpus Linguistics*. London: Longman, 111–127.
- Krenn, Brigitte (2000), *The Usual Suspects: Data-oriented Models for the Identification and Representation of Lexical Collocations*. (Saarbrücken Dissertations in Computational Linguistics and Language Technology 7.) Saarbrücken: DFKI & Universität des Saarlandes.
- Lea, Diana (ed.) (2002), *Oxford Collocations Dictionary for Students of English*. Oxford etc.: Oxford University Press.
- Levin, Lori/Tokunaga, Takenobu/Lenci, Alessandro (eds.) (2003), *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan.
- Liebetrau, Albert M. (1983), *Measures of Association*. (Sage University Papers Series on Quantitative Applications in the Social Sciences 32.) Newbury Park: Sage.
- Lin, Dekang (1998), Extracting Collocations from Text Corpora. In: *Proceedings of the First Workshop on Computational Terminology*. Montreal, Canada, 57–63.
- Lüdeling, Anke/Bosch, Peter (2003), Identification of Productive Collocations. In: *Proceedings of the 8th International Symposium on Social Communication*. Santiago de Cuba, Cuba.
- Manning, Christopher D./Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Michelbacher, Lukas/Evert, Stefan/Schütze, Hinrich (2007), Asymmetric Association Measures. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*. Borovets, Bulgaria. Available at: <http://www.cogsci.uni-osnabrueck.de/~severt/PUB/MichelbacherEtc2007.pdf>.
- Palmer, Frank R. (ed.) (1968), *Selected Papers of J. R. Firth 1952–59*. London: Longmans.

- Pecina, Pavel (2005), An Extensive Empirical Study of Collocation Extraction Methods. In: *Proceedings of the ACL Student Research Workshop*. Ann Arbor, MI, 13–18.
- Pedersen, Ted (1996), Fishing for Exactness. In: *Proceedings of the South-Central SAS Users Group Conference*. Austin, TX, 188–200.
- Sag, Ivan A./Baldwin, Timothy/Bond, Francis/Copestake, Ann/Flickinger, Dan (2002), Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*. Mexico City, Mexico, 1–15.
- Sahlgren, Magnus (2006), *The Word Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces*. PhD thesis, Department of Linguistics, Stockholm University.
- Schiehlen, Michael (2004), Annotation Strategies for Probabilistic Parsing in German. In: *Proceedings of COLING 2004*. Geneva, Switzerland, 390–396.
- Schone, Patrick/Jurafsky, Daniel (2001), Is Knowledge-free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. Pittsburgh, PA, 100–108.
- Schütze, Hinrich (1998), Automatic Word Sense Discrimination. In: *Computational Linguistics* 24(1), 97–123.
- Sinclair, John (1966), Beginning the Study of Lexis. In: Bazell, Charles E./Catford, John C./Halliday, Michael A. K./Robins, Robert H. (eds.), *In Memory of J. R. Firth*. London: Longmans, 410–430.
- Sinclair, John (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, John (ed.) (1995), *Collins COBUILD English Dictionary*. New edition, completely revised. London: Harper Collins.
- Sinclair, John/Jones, Susan/Daley, Robert/Krishnamurthy, Ramesh (2004), *English Collocation Studies: The OSTI Report*. London etc.: Continuum Books. Originally written in 1970 (unpublished).
- Smadja, Frank (1993), Retrieving Collocations from Text: Xtract. In: *Computational Linguistics* 19(1), 143–177.
- Smadja, Frank/McKeown, Kathleen R./Hatzivassiloglou, Vasileios (1996), Translating Collocations for Bilingual Lexicons: A Statistical Approach. In: *Computational Linguistics* 22(1), 1–38.
- Stevenson, Suzanne/Fazly, Afsaneh/North, Ryan (2004), Statistical Measures of the Semi-productivity of Light Verb Constructions. In: *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*. Barcelona, Spain, 1–8.
- Stubbs, Michael (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.
- Tanaka, Takaaki/Villavicencio, Aline/Bond, Francis/Korhonen, Anna (eds.) (2004), *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing*. Barcelona, Spain.
- Terra, Egidio/Clarke, Charles L. A. (2004), Fast Computation of Lexical Affinity Models. In: *Proceedings of COLING 2004*. Geneva, Switzerland, 1022–1028.
- Villada Moirón, Begoña/Villavicencio, Aline/McCarthy, Diana/Evert, Stefan/Stevenson, Suzanne (eds.) (2006), *Proceedings of the ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Sydney, Australia.
- Williams, Geoffrey (2003), Les collocations et l'école contextualiste britannique. In: Grossmann/Tutin 2003, 33–44.
- Yates, Frank (1984), Tests of Significance for 2×2 Contingency Tables. In: *Journal of the Royal Statistical Society, Series A* 147(3), 426–463.
- Zinsmeister, Heike/Heid, Ulrich (2003), Significant Triples: Adjective + Noun + Verb Combinations. In: *Proceedings of the 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003)*. Budapest, Hungary, 92–101.

Stefan Evert, Osnabrück (Germany)

59. Corpora and text re-use

1. Notions of text re-use
2. Corpora and text re-use
3. Detecting and measuring text re-use
4. Applications of detecting text re-use
5. Conclusion
6. Literature

1. Notions of text re-use

Text re-use is the activity whereby pre-existing written texts are used again to create a new text or version (Aizawa (2003) calls this *text recycling*). By text we mean an entire document, or parts of a document such as a grammatical unit (sentence, paragraph) or fixed-length block of text. In some cases, entire documents are reproduced or repeated, perhaps as a result of duplication, e. g. the replication of entire websites across different servers. However, it is far more likely that part of a document is re-used rather than the whole, either copied word-for-word (verbatim) or rewritten (e. g. paraphrased or summarised) to fit within a different context. Examples include the re-telling of tales in literary and historical texts, the revision, summarisation or translation of existing sources, the re-use of news agency text by journalists, and probably the ‘classic’ example plagiarism: the unacknowledged or unethical re-use of text.

Given that the volume of readily available electronic texts is growing at a rapid rate, it is not surprising that the computational study and analysis of text re-use is becoming a popular research theme. In particular, the reliable, automatic detection of text re-use is both an interesting intellectual problem and one whose solution promises practical benefits to individuals and organisations. For example, teachers assessing the originality of a student’s work, companies wanting to find breaches of ownership or wishing to track or monitor the dissemination of their digital content, and Web search engines wishing to eliminate duplicate content prior to presenting search results to the user all stand to benefit from reliable automated techniques for detecting text re-use.

The aim of this article is to present an overview of the area of text re-use, including: examples of types of re-use of significant interest; corpora created specifically to assist in the study of text re-use; algorithmic methods for detecting and measuring text re-use; and practical applications of these detection algorithms. Much of this article addresses the third of these topics – computational methods for identifying text re-use. Although often easy to distinguish manually, discriminating re-use from incidental similarity automatically can be very difficult (Whale 1990). To address this issue, most automatic approaches use some measure of similarity to distinguish derived and non-derived texts. Therefore in discussing text re-use we also review approaches for measuring similarity between texts.

1.1. Examples of text re-use

Let us take, by way of initial illustration, two examples of text re-use: the unacknowledged re-use of published texts (plagiarism) and the re-use of newswire texts by the journalist.

1.1.1. Plagiarism

According to Hannabuss' (2001, 313) definition, plagiarism is the “unauthorised use or close imitation of the ideas and language/expression of someone else and involves representing their work as your own.” Plagiarism is closely linked with intellectual property and copyright, both of which are set in place to protect the ownership of ideas whose visible expression is in the form of a text. As Osen (1997, 15) comments: “if plagiarism is the bane of the academic world, copyright infringement is the scourge of the legal one.” There are cases when what appears to be plagiarism is not, e. g. appropriate self-re-use of one’s own work (Samuelson 1994; Collberg/Kobourov 2005), or poor citation. These cases can be resolved through manual inspection.

Plagiarism can take several distinct forms, including word-for-word plagiarism, paraphrasing, plagiarism of secondary sources, plagiarism of ideas, and plagiarism of authorship (Martin 1994). Given a source text, the most obvious form of plagiarism is word-for-word copying of the source. This can often be automatically detected using the simplest of methods, but occurrences by students are often due to the fact that they are uncertain as to how to cite or paraphrase source texts legitimately. Other forms, such as paraphrasing and the re-use of structure can also be identified relatively easily, but get progressively harder as the plagiarist uses more complex rewrites to hide the original text, or re-uses only ideas and not the content. These forms of plagiarism are not just harder to detect, but also harder to prove.

In recent years, plagiarism (and its detection) has received much attention from both the academic and commercial communities. This has been particularly true in academia, as students have used technology to fabricate texts (e. g. using pre-written texts from essay banks or paper mills, using word processors to manipulate texts and finding potential source texts using online search engines). Furthermore, the change of culture brought about by electronic ‘cyber-space’ has caused concern to authors surrounding the ownership of their written material (Wilks 2004). As Mallon (1989, 247) suggests: “the origin and ownership of all electronic documents is now peculiarly evanescent; one click of the ‘Save As’ button can give a whole new name and identity, instantly, to someone else’s creation.”

1.1.2. News production

While the re-use of others’ text without acknowledgement is, in academic life, a cardinal sin, there is one industry where this is not only accepted behaviour, but is in fact standard business practice. In the production of news, most newspapers rely very heavily upon press agencies as their primary source of written news material (Bell 1996, 20–22). Upon payment of a subscription fee, the newspaper is free to re-use this material verbatim, or to edit it in whatever way it sees fit, often without having to acknowledge the source.

The process of gathering, editing and publishing newspaper stories is a complex and highly specialised task often operating within specific publishing constraints such as short deadlines, prescriptive writing practice, limits of physical size during page layout, readability and audience comprehension, editorial bias, and a newspaper’s house style. Although news agency copy is re-used in the creation of a news story, due to the afore-

mentioned publishing constraints, it is unlikely that agency copy gets re-used word-for-word, and almost invariably differences will arise. The journalist may decide to rewrite, re-order, delete or paraphrase agency copy, rather than re-use the text verbatim depending on a wide variety of external influences and personal writing style. On the other hand, even independently-written texts will share similarities given that journalists write news stories in certain and predictable ways (Bell 1991; Keeble 1998; van Dijk 1988) and use conventional expressions and terminology in their style (Waterhouse 1993; Reah 1998).

Previous research has identified the major rewriting operations used by journalists and editors such as deletion, lexical substitution, changes in syntax and summarisation (Bell 1991; Fries 1987; van Dijk 1988). More specifically, these include deletion of redundant information and deletion resulting from syntactic changes, substitution of synonymous words and phrases, changes in word order, changes in tense, passive to active voice and verb/noun nominalisation. For examples of text re-use in journalism, see e.g. Bell (1991); Clough et al. (2002).

1.2. Defining text re-use

The preceding section gave two common examples of text re-use. Before proceeding to discuss resources that have been created to study re-use and techniques that have been developed to identify it automatically, it is useful to analyse the notion of text re-use a bit further.

Core to this notion is that an author uses one text in the process of generating another. Two aspects of this process deserve further attention: the nature of the use and the intentions of the author.

It is common to distinguish the content or information in a text from the words or form of expression by which the information is conveyed. Does the phenomenon of text re-use extend to cover any re-use of content or ideas obtained from one text in another? Or does it require the preservation in the new text of at least some of the forms of expression used in the original and if so, how many? We propose that the notion text re-use implies more than the re-use of some content or ideas drawn from the original. Thus an historian who reads an account of the Battle of Stalingrad and later, without re-consulting the text, uses facts drawn from this account in an article on the Soviet Union is not, on this view, engaged in text re-use; nor was James Joyce engaged in text re-use when he wrote *Ulysses*. However, there are cases of text re-use where very little, if any, of the original forms of expression are retained. So, what sort of relation must there be between the two texts in order for text re-use to have occurred? We propose that the key distinction between an author being influenced by, or drawing on the work of, another and re-using another's text is whether or not the author *consciously* applies some process of transformation to the form of expression in the source text to arrive at the final text. In other words, he intentionally re-writes one text to obtain the other.

Of course determining what this transformation process is and whether it has occurred may not be straightforward. Shared sequences of identical text may suggest, beyond any reasonable doubt, that one text was used in writing the other. But in general, information may be required from beyond the shared texts, for example, the acknowl-

edgement of the author that re-use occurred, knowledge of how likely it is that the author has read the candidate source text, or observations of stylistic discontinuity between sections of the new text suggesting some part of it has been imported.

2. Corpora and text re-use

Many text collections used in language studies contain repeated or duplicated material, sometimes resulting from text re-use (Bouayad-Agha/Kilgarriff 1999). For example, the Reuters-21578 news corpus (<http://trec.nist.gov/data/reuters/reuters.html>), the Text REtrieval Conference (TREC) collections (<http://trec.nist.gov/>) and Linguistic Data Consortium's English Gigaword corpus (<http://www.ldc.upenn.edu/Catalog/-CatalogEntry.jsp?catalogId=LDC2003T05>) all contain repeated newswire stories which occur because press agencies often repeat previously published text, adding new information as a story unfolds. The Web is also being used as a corpus to provide researchers with a source of linguistic data (Kilgarriff/Grefenstette 2003; cf. article 18). This also exhibits duplication resulting from multiple copies (or versions) of the same texts (Broder 1998; Shivakumar/Garcia-Molina 1995). While many text collections have not been designed to study text re-use, some researchers have used them for this purpose. For example, Sanderson (1997), Chowdhury et al. (2002), Aizawa (2003) and Metzler et al. (2005) have used the Reuters-21578 and TREC collections to study duplication and re-use and gather statistics that can be used to inform the document ranking processes of search engines.

Creating publicly accessible resources to investigate text re-use and evaluate computational approaches for detecting re-use is often difficult. For example, to collect and distribute confirmed instances of plagiarism is often impossible due to copyright and privacy restrictions. However, samples from the news industry are less restricted and have become available to the research community. We discuss two text collections which have been used to analyse text re-use in the production of news stories: the METER corpus and the Lancaster Newsbooks corpus.

2.1. The METER corpus

The METER corpus is a novel resource created to support the study of text re-use in journalism (Gaizauskas et al. 2001; Clough/Gaizauskas/Piao 2002). The corpus is a small hand-annotated collection of 1,716 news texts (over 500,000 words) using the UK Press Association (PA) – the major UK news agency – as the source, and stories about the same news events as published in nine British daily newspapers who subscribe to the PA as candidate re-users. Stories from five ‘popular’ papers (*The Sun*, *The Daily Mail*, *Daily Star*, *Daily Express* and *Daily Mirror*) and four ‘quality’ papers (*Daily Telegraph*, *The Guardian*, *The Independent* and *The Times*) and corresponding PA copy are included. The corpus has been encoded in a machine-readable format.

The material has been selected from a 12-month period (July 1999 to June 2000) from the areas of law and court reporting (769 stories) and show business (175 stories). All newspaper texts have been manually classified at the document-level as to whether they are wholly-derived (300), partially-derived (438) or non-derived (206) from the PA source

text. In addition, 355 of the derived texts have been classified according to a lexical-level scheme whereby every word sequence in the text is annotated as either verbatim, rewritten or new, depending on its relation to material in the PA text from which it is derived. Although it is impossible to be certain about the derivation relationship between texts in this domain, due to restricted access to the news workers who create the texts, trained journalists with experience and knowledge of the UK news industry were used to assess re-use and carry out the annotation.

The corpus is available from the authors and has been used for various research investigations including the study of newswire-newspaper text re-use (Clough et al. 2002), semantic tagging (Piao et al. 2004), and measuring the similarity between document pairs based on their thematic content and textual expression (Uzuner/Davis/Katz 2004).

2.2. The Lancaster Newsbooks corpus

The Lancaster Newsbooks corpus (<http://bowland-files.lancs.ac.uk/newsbooks/project.htm>) is a collection of news stories (newsbooks) used to study text re-use within texts written in the seventeenth century. As with journalists in modern news production, the writers of news in this era often duplicated or imitated existing sources or used the same text to report similar events in different newsbooks.

Originals were transcribed from microfilm or paper from the period between December 1653 and May 1654 comprising approximately 800,000 words. Using a sentence alignment algorithm developed for finding newspaper stories derived from newswire sources which had been trained/tested on the METER corpus (Piao et al. 2003), newsbooks were analysed for their similarity to determine the extent to which one had copied from the other or made use of a shared third source. The corpus is available from the project co-coordinators at Lancaster University and to date only limited analysis of text re-use within this small corpus has been undertaken.

3. Detecting and measuring text re-use

The solution to many problems in computational text analysis requires some measurement of *similarity* (or difference) between texts. For example, the retrieval of documents to fulfil a user's information need (Korfhage 1997), clustering documents according to some criterion (Willett 1988), summarising multiple documents (Mani/Maybury 1999; Hatzivassiloglou/Klavans/Eskin 1999; cf. article 60), aligning sentences from one language with those in another (Brown/Lai/Mercer 1991), finding exact and near duplicates of documents (Brin/Davis/Garcia-Molina 1995), plagiarism detection (Clough 2000), identifying the flow of information through texts (Metzler et al. 2005), classifying documents according to genre (Karlgren/Cutting 1994; Kessler/Nunberg/Schütze 1997; cf. article 38) and identifying authorship attribution (McEnery/Oakes 2000; Holmes 1994; Rudman 1998).

Providing a general definition of similarity is difficult, since the notion is often problem-specific (Hatzivassiloglou/Klavans/Eskin 1999). For example, the similarity between

a query and document in Information Retrieval (IR) is usually taken to mean *thematic* or *topical* similarity: a document is judged *relevant* if it is on the same *theme* or *topic* as the user's query (Korfhage 1997). The extent to which texts (entire documents or parts of a document) are considered topically similar to one another can be measured on a scale ranging from broad topical similarity at one end, to identical documents at the other (Metzler et al. 2005). Text re-use, however, is a different relation between texts than topical similarity (see section 1.2. above), though clearly re-used text will be topically similar to its source. As a consequence identifying text re-use cannot be limited to comparing the topical content of documents; it must also take into account the similarity of *expression* (Uzuner/Davis/Katz 2004) and the *style/authorship* of writing (Biber 1988; McEnery/Oakes 2000).

3.1. A framework for comparing texts

There exists a multitude of approaches to compare texts; however most consist of at least the following three stages:

1. pre-process input texts into an intermediate representation suitable for comparison
2. compare the intermediate representations derived in the pre-processing step
3. output a quantitative measure of similarity or visualise the output of the comparison step.

In the first stage, the input texts are pre-processed in order to create an intermediate representation for comparison. For example, in IR it is common to apply some form of suffix removal to increase the chances of matching between query-document terms; for detecting plagiarism in software code, it is common to compare parse trees rather than program code itself and replace variable names with unique identifiers. In natural language, one can imagine comparing alternative views of the text such as grammatical or semantic representations of the input texts (e. g. dependency trees).

The main goal of this pre-processing stage is to reduce the effect of differences due to systematic changes resulting from lexical variation (e. g. the use of synonymous terms) or syntactic variation (e. g. changes in tense or voice). Further decisions at this stage could include the size of unit to compare (e. g. words, sentences or fixed-length chunks) and to what extent language-dependent resources are involved.

In the second stage, the representations of the input texts derived in the first stage are compared, typically using some sort of similarity measure. For example, if the texts are represented by *unordered* sets of lexical items or pairs/triples of lexical items occurring within them then various lexical proximity or overlap measures can be used to compute the degree of similarity between them (see sections 3.2.1. and 3.2.2. below). If the texts are represented as *ordered* sequences of characters or words, then a sequence comparison method, such as edit distance, can be used to align the texts (see section 3.2.3.). A further consideration at this stage is whether texts are to be compared globally or whether a set of more local comparisons is to be made between segments of the texts. For example comparison of two texts could be made by pair-wise comparison of their sentences or paragraphs using one of the similarity measures mentioned above for each comparison. The resulting set of measures can then be used to compute a single global similarity measure or used to inform some other decision about the text.

The final stage selects an appropriate form of output after comparing texts: a quantitative measure or a qualitative representation of similarity (or difference). Quantitative results typically take the form of a number between 0 (maximum difference or total dissimilarity) and 1 (maximum similarity or identity). Qualitative representations can include visualisations showing the position of matches (section 3.2.6.), a list of the common substrings found between two texts, a summary of the matches found in the form of a histogram, a list of aligned sentences or an edit/alignment script.

Similarity measures are often used in deciding whether text re-use has occurred, but an important observation is that a single global similarity measure might in itself not be the most appropriate discriminator. For example if after comparing documents a single match of 10 consecutive words is found, then over the entire length of the document this might result in an apparently low global similarity score. However, a match of 10 consecutive words is very significant and highlights almost certain re-use (unless a formulaic phrase or direct quote). To address this, documents must be divided into smaller units and similarity computed between these or alternative discriminators of re-use must be found (e. g. matches between document pairs could be ranked by alternative measures such as the length of shared word sequences (Wise 1993), the similarity scores could be weighted by the length of matching sequence(s) or the similarity between documents visualised rather than quantified).

3.2. Techniques for determining text re-use

In this section we describe a number of approaches to measuring text similarity that have been applied to the problem of detecting text re-use. This review cannot be complete – there are simply too many techniques to cover exhaustively. Instead we introduce the approaches that have had the most influence on research in this area in sufficient detail that the reader may, we hope, grasp the central ideas underlying the approaches. References are provided to the many variants of these approaches that have been tried in various applications. Other comparative studies of similarity measures for various problems in text analysis (including plagiarism detection) have been carried out: e. g., Lopresti (2001); Shivakumar/Garcia-Molina (1995); Uitdenbogerd/Zobel (1999); Petrakis/Tzeras (2000); Metzler et al. (2005). The approaches most commonly described include those from information retrieval, sequence comparison and n-gram matching, and hence we begin with these.

3.2.1. Lexical similarity

The most common task addressed by IR is *ad-hoc* retrieval in which a user poses a *query* describing their information need against a relatively static collection of documents. The system then returns a list of documents which match the user's query, possibly ranking the returned documents in order of their presumed relevance to the query. One of the classic models for *ad-hoc* ranked retrieval is the *vector space model*. This approach treats documents and queries as vectors in a high-dimensional space (*n-space*) where each dimension corresponds to a term in the document collection (i. e. *n* is the number of terms

in the index). The vector space model ranks documents according to the similarity between the query and each document retrieved. The most relevant documents are assumed to be those represented by the vectors closest to the query vector. The similarity between vectors can be expressed, for example, as the angle between two vectors given by the *cosine measure*. A ranked list is produced by sorting the documents retrieved in decreasing order of similarity score. The vector-space model has proven to be very robust as well as simple and has been shown to perform well on many data sets.

Not all terms in a document have the same ‘significance’ or ‘importance’ and *term weighting* is used to reflect this. Index terms are given a weighting value indicating how well they discriminate between documents and this is used to weight vectors in the vector space model. *Term frequency* captures the importance of a word within a document: the more frequent the term, the more likely the term captures something of a document’s content. *Document frequency* captures the importance of a word within a collection: a term appearing in many documents of a collection is less able to discriminate between documents than if it appears in just a few. A standard term weighting is the *inverse document frequency* that weights an index term appearing in more documents across a collection as less significant than one which appears in fewer documents. Term frequency and document frequency are typically combined, together with document length normalisation, to form the *tf-idf* family of weighting schemes. These are among the most widely used and effective schemes for term weighting in any IR model. Many variations on the basic tf-idf weightings have been proposed.

In the simplest case, an IR system does not use structure or discourse when matching user requests to documents in a collection. Typically, documents are treated as *bags-of-words* because interdependencies between terms are not preserved. Similarity measures for IR systems tend to be based on the overlap of index terms (or lexical items) between documents and the query. Therefore similarity as computed by IR systems is fundamentally *lexical similarity*. While vector space models have been widely used for gathering topically similar documents in IR systems they have not been widely used for detecting text re-use. More information about IR and the vector space model can be found in Baeza-Yates/Ribeiro-Neto (1999) and Manning/Schütze (1999, chapter 15).

3.2.2. Overlap of n-grams

N-grams (or q-grams/k-grams) have been a popular area of research for many years and are most commonly used in empirical linguistics to build statistical models of language (Jurafsky/Martin 2000). These models can be found in many areas of language research such as speech and text processing. N-grams have also been used to measure similarity between document pairs for applications including information retrieval (Kimbrell 1988), language identification and categorisation (Damashek 1995), finding plagiarised documents (Lyon/Malcolm/Dickerson 2001), document compression (Broder 1998; Heintze 1996), the evaluation of machine translation results by comparison against multiple reference translations (Papineni et al. 2002; Doddington 2002; Sorice/Brill 2004) and spelling-error detection and correction (Willett 1988).

An n-gram is a string of n adjacent characters or words. N-grams can be used as an intermediate representation for texts above the level of sets of characters or words. As a representation, an n-gram contains more information than just the word or character

alone because it also provides their surrounding context. A set of n-grams can be generated from a document by representing it as a string of tokens (e.g. characters or words) and moving an n-token window from the beginning to the end of the string, one token at a time. This generates overlapping n-grams. For a string of length n tokens and a window of length m , $(n - m) + 1$ n-grams are generated. For example, the character bigrams resulting from the word SUBSTRING are: SU, UB, BS, ST, TR, RI, IN, NG. In character-based applications, it is common to find $n-1$ padding spaces at the beginning and end of the string to account more fully for all possible n-grams.

In the simplest case, n-grams are treated as independent, overlapping units. However, their location, order, length and frequency of occurrence in a collection can all be used to enhance a similarity score. Further variations include whether the n-grams are variable or fixed-length, overlapping windows or not, whether all n-grams are used or only a selection (e.g. as for copy detection) and whether n-gram types or tokens are counted (Brin/Davis/Garcia-Molina 1995; Shivakumar/Garcia-Molina 1996; Lyon/Malcolm/Dickerson 2001; Broder 1998). The final aspect to consider is how the similarity between n-gram sets is computed. Practically, n-grams might be captured using data structures such as sets, lists or vectors, and different similarity measures utilized to measure the overlap between them. For example, Damashek (1995) uses a cosine similarity measure between character n-grams, Shivakumar/Garcia-Molina (1996) use a modified version of a cosine coefficient to capture partial copies (called the subset measure), Lyon/Malcolm/Dickerson (2001) use resemblance and containment (Broder 1998; Chakrabarti 2002, 67–72) on word n-gram sets and Willett (1988) uses the Dice coefficient for document clustering using character n-grams.

Because the way in which people express themselves varies from person to person, even if writing about the same subject, it is unlikely that the topic will be expressed in exactly the same way (e.g. the same program structure or grammatical syntax), and using exactly the same words (i.e. the same vocabulary). As McEnery/Wilson (1996, 7) state about finding the same sentence more than once: “unless it is a very formulaic sentence (such as those appearing as part of a legal disclaimer at the beginning of a book), it is deeply unlikely that you will find it repeated in its exact form in any book, in any library, anywhere.”

When using n-gram based approaches to detect text re-use, the value of n must be derived empirically and is important: values too small will result in simply matching common words or idioms across independently-written texts; too large and matches between dependent texts will be missed. For example, Lyon/Malcolm/Dickerson (2001) show that within a collection of texts, word trigrams can be used to discriminate texts copied from those written independently; Shivakumar/Garcia-Molina (1995) find that for copy detection best results are obtained using word unigrams; Bharat/Broder (1999) conclude word bi-grams appear best for finding duplicate web host pairs.

3.2.3. String or sequence comparison

A different method for comparing texts is to compute a difference score based on string or sequence comparison, usually expressed as the number of elementary transformation operations required to turn one string into another. In this approach the *ordering* of strings is considered important and used during comparison. The granularity with which

texts are represented as strings can vary from characters, through words to entire lines or sentences. The indivisible unit of comparison between strings is called the token (or symbol). Independent of token size, many string comparison algorithms exist both to quantify similarity and highlight where the similarities or differences between two strings lie. This is necessary for applications such as file comparison where token differences between strings are used to track file revisions.

Sankoff/Kruskal (1983) suggest that one sequence can be transformed into another using the following *edit operations*: insertions and deletions (or indels), substitutions (can be analysed as insertion-deletion pairs), compressions and expansions, and transpositions (or swaps). The distance between two sequences can be expressed using these edit operations. Where an insertion or deletion edit operation occurs, this is known as a *gap* (denoted by ‘-’). Consider the following four elementary one-character edit operations:

- (a,a): a match (no change)
- (a,-): the deletion of a
- (-,b): the insertion of b
- (a,b): the replacement of a by b (for $a \neq b$)

The alignment of two sequences is achieved by applying a series of elementary edit operations to one sequence to transform it into the other. For example, given the sequences S=STARWARS and T=STAIRS, alignments including the following can be made:

S	STARWARS	STARWA-RS
T	STAI--RS	ST---AIRS

The first of these is obtained using the following single character edit operations: (S,S), (T,T), (A,A), (R,I), (W,-), (A,-), (R,R), (S,S).

The number of possible alignments in the previous example is much greater than two; therefore the question becomes how to test for equality between two strings allowing some errors and optimising over all possible correspondences that satisfy suitable conditions such as preserving the order of sequence elements. To turn the alignment into a quantitative distance score, the edit operations are assigned weights or costs and a distance computed as the sum of weights for each alignment. For example, suppose the following weights are assigned to the previous edit operations, and apply to arbitrary characters in the sequence alphabet:

$$\begin{aligned} w(a,a) &= 0 \\ w(a,-) &= w(-,b) = 1 \\ w(a,b) &= 1 \text{ for } a \neq b \end{aligned}$$

Given this weighting scheme, the alignments in the STARWARS/STAIRS example have costs 3 and 4 respectively.

It is often necessary to find the minimum number of differences between x and y or the cost of changing x into y. This can be translated into finding the edit operations required to transform x into y with the minimum cost (the *edit distance*). If the cost of indels is 1, the cost of a change 2 and the cost of a match 0, this gives rise to the most popular string dissimilarity measure: the *Levenshtein distance* (or unit cost model). This is the minimum cost for a sequence of edit operations to change one string into another.

If only insertion and deletion edit operations are allowed, it turns out that the problem of computing the minimum number of edit operations is equivalent to finding the *longest common subsequence* or *lcs*. A subsequence is any string that can be obtained by deleting zero or more tokens from a given string. The *lcs* between strings A and B is the longest consecutive sequence of elements (of maximum weight) taken in order from A and B. For example, *air* is a common subsequence between the character strings A=*reagir* and B=*repair*, while the *lcs* between A and B is the string *reair*. The *lcs* can be computed using an edit distance allowing only insertions and deletions where each has a weight of 1. The following relation holds between edit distance and lcs. Suppose $d(a, b)$ is the edit distance between strings A and B, $q(A, B)$ is the length of the *lcs* between A and B, and $\text{length}(A)$ and $\text{length}(B)$ are the lengths of A and B respectively. Then $d(A, B) = \text{length}(A) + \text{length}(B) - 2q(A, B)$ (Coggins 1999, 312).

Edit distance (and the *lcs*) can be computed using a technique known as *dynamic programming*. This was developed independently by Needleman/Wunsch (1970) for biosequence alignment, and Wagner/Fisher (1974) to solve the string-to-string correction problem. This algorithm computes an optimal edit distance, but due to the quadratic time and space complexity, several other approaches have been suggested which either approximate an optimal solution, or weaken constraints of the problem (Sankoff/Kruskal 1983).

The alignment described so far is known in the biological domain as a *global alignment*, that is, an alignment based upon the entirety of both sequences. Sometimes two sequences might appear dissimilar overall, but may contain portions which are similar, e. g. if one sequence contains a portion of another or if a group of contiguous tokens in one sequence can be rearranged to create another sequence (known as a *block move*). In these cases, a pair-wise global alignment will not exhibit clear regions of high similarity. For example, suppose we compare the word-tokenised string A=“Lee Harvey Oswald shot the President” to the two strings B=“The President was shot by Lee Harvey Oswald” and C=“Lee Harvey Oswald lives in Dallas”. In each case the *lcs* between the initial string and the comparison strings is 3 words in length (“Lee Harvey Oswald”), yet there is much more in common between A and B than there is between A and C. In the A-B case, the *lcs* has skipped over the differences and does not reflect all substring matches that include “shot” and “The President” (block moves).

The problem of block moves has been noted by researchers from various communities including those involved in biological sequence comparison (Smith/Waterman 1981), file comparison (Heckel 1978; Tichy 1984), and plagiarism detection (Gitchell/Tran 1999; Wise 1993, 1996). Sequence comparison methods have been used in plagiarism detection (Gitchell/Tran 1999; Wise 1993, 1996) and spelling error detection and correction (Kukich 1992), for sentence and word alignment (Church/Helfman 1993), for identification of cognate pairs (Tiedemann 1999), for file comparison (Heckel 1978; Tichy 1984), to capture differences between spoken and written language (Murata/Isahara 2002), string comparison (Lopresti 2001; Ristad/Yianilos 1998), adaptive name matching in information integration (Bilenko et al. 2003) and building paraphrase corpora using parallel news sources (Dolan/Quirk/Brockett 2004).

3.2.4. Sentence alignment

One important form of text re-use not discussed so far is the translation of documents from one language to another. Viewing translation as a form of text re-use suggests that

techniques developed to support machine translation may have application to detecting monolingual text re-use. For example, in automatic machine translation, text alignment is commonly used to create lexical resources that aid the translation process, e. g. bilingual dictionaries and parallel grammars (cf. Manning/Schütze 1999). An important stage is the alignment of bilingual texts (or bitexts) such that sentences and paragraphs in one language correspond to sentences and paragraphs in another with the same content. Over the past decade, several successful algorithms have been proposed for alignment of multilingual corpora (reviews can be found in Manning/Schütze 1999 and Wu 2000; cf. article 32).

When a document is translated from its original form (called the *source*) to a new form in the destination language (called the *target*), sentences might not appear in the same way. For example, sentences can be merged together, broken up or deleted. The target text may even contain sentences that are new and cannot be found in the source text. The task of sentence alignment is to seek a group of sentences (called a bead) in the target text corresponding to a group of semantically related sentences in the source, where either group can be empty to allow insertion and deletion of sentences. The question becomes how much content must overlap before sentences can be said to be in an alignment relation (an empirical problem).

Manning/Schütze (1999) make the distinction between alignment and correspondence algorithms. The distinction depends upon whether an algorithm allows crossing dependencies or not; that is, whether sentences in the target text must appear in the same order as in the source. Several approaches have been suggested for sentence alignment in bilingual texts, but most follow one of three method types. The first is length-based methods in which the underlying assumption is that short sentences get translated as short sentences and long sentences as long sentences (Brown/Lai/Mercer 1991; Gale/Church 1991). The second is lexical-based methods that use lexical content of sentences to guide the alignment. This process is iterative such that sentences are aligned from the alignment of words which can be found in a bilingual dictionary (Kay/Röscheisen 1993; Chen 1993; Catizone/Russell/Warwick 1989). Finally there are the so-called cognate-based methods, which assume that some words are similar across languages due to, for example, common or borrowed linguistic ancestors. For example, the French *supérieur* and English cognate *superior* are similar and could be found using approximate string matching. These methods assume that cognates can be used as anchor points to identify lexical correspondences within sentences and therefore align sentences themselves (Simard/Foster/Isabelle 1992; Melamed 1999).

3.2.5. Summarisation and paraphrasing

It is not uncommon when text is re-used for some words and phrases to be re-used verbatim while other words and phrases are not re-used literally but rather have substituted for them other words and phrases with similar meaning (also called *paraphrases*). Thus, when measuring similarity to detect text re-use it may be useful to consider that tokens, n-grams or sequences may match not just when their elements are identical but when there is a more complex relationship of semantic equivalence between them. Considerable work has been undertaken to assemble collections of paraphrases empirically from corpora, mostly in the context of research into multi-document summarization

where the detection and elimination of redundancy in multiple texts on the same topic is important.

In work on multi-document summarization, McKeown et al. (1999) describe a process for detecting similarities and differences between newswire texts reporting the same story in order to produce a coherent summary. To detect semantically similar paragraphs, they use a method to find new paraphrase examples that employs a supervised learning technique trained over feature vectors generated from 10,345 sentence pairs which have been manually judged as similar or not-similar (Hatzivassiloglou/Klavans/Eskin 1999). Each feature vector consists of text units which can be either primitive (consisting of one unit) or composite (consisting of pairs of primitive units). Primitive indicators include matching noun phrases, common semantic classes for verbs, Wordnet synonyms, word co-occurrence and shared proper nouns. Composite features are derived from the primitives by placing different types of restrictions on the participating primitive features including ordering, distance and which primitives can match.

Barzilay/McKeown (2001) describe a method of paraphrase identification using cognate-based sentence alignment, and an unsupervised learning method, to learn paraphrase rules from a parallel corpus of multiple English translations of the same source text. Two translations are aligned using the co-occurrence patterns of words, with the assumption that phrases in aligned sentences will appear in similar contexts. Given a pair of aligned sentences, an unsupervised learning approach uses a number of syntactic rules (e.g. verb-object and noun-modifier relations) to make a binary decision as to whether two sentences are paraphrases or not. Contexts surrounding all those words found to match (between two possible paraphrases) are used to ‘bootstrap’ the learning algorithm to learn ‘good’ contexts, i.e. find more paraphrase rules. Lexical and syntactic features are used to identify paraphrase pairs. Evaluation of the method based on manual examination of randomly sampled paraphrases discovered by the algorithm shows promising results (more than 85% of the paraphrases are judged correct).

The preceding two approaches to paraphrase acquisition are indicative of the type of work going on in this area. Closely related is work of Jing/Mckeown (1999) on determining the most likely source sentence in a document for each sentence in a summary and of Mani/Bloedorn (1999) and Shinyama et al. (2002) on using information extraction techniques to find regions of similarity and difference (Mani/Bloedorn) or paraphrases (Shinyama et al.) across multiple news stories about the same event. All of this work has been carried out in the context of research into automatic summarization. Despite its relevance to detecting text re-use, no one has, to the best of our knowledge, yet applied paraphrase acquisition and recognition to that problem.

3.2.6. Visual methods

Visual approaches have also been used to aid the manual inspection of similarity in a range of applications including plagiarism detection (Culwin/MacLeod/Lancaster 2001; Ducasse/Rieger/Demeyer 1999) and the visualisation of music structure (Foote/Cooper 2001). One of the most popular approaches is the Dotplot. A Dotplot is a two dimensional graphical representation with one of the input sequences to be compared along the X-axis, the other along the Y-axis, and a dot placed at each (X, Y) co-ordinate where the X and Y tokens match. In this representation possible alignments between the two

input sequences appear as visual patterns. Many patterns can be formed, the most notable being diagonal lines which indicate a series of consecutive matching tokens. While detecting such lines may be left to manual inspection, researchers from various fields have developed methods for extracting these diagonal lines automatically (Melamed 1999; Chen 1993; Fung 1995).

4. Applications of detecting text re-use

In this section we present an overview of three main examples of how the techniques described in section 3 have been applied in practice: plagiarism detection, copy detection and newswire monitoring.

4.1. Plagiarism detection

Since the 1970s, the level of interest in and the focus of research into automatic plagiarism detection has changed significantly. To begin with, empirical research came from the programming community, particularly in academia where computer science departments built tools to identify “unusual” similarity between programming assignments (or “constrained” texts). There is still great interest in identifying similarity between large software programs (e.g. duplication, redundant code and similarity between revisions), especially in industry (Hislop 1998). More detailed information regarding techniques used to detect plagiarism in software code can be found in (Clough 2000, 2003).

More recently interest has shifted towards identifying plagiarism between natural language (or “unconstrained”) texts. Particular areas of concern include identifying verbatim cut-and-paste (with minor changes) from Web-based sources and identifying paraphrased versions of identical content. This change is reflected in the increase in on-line services to detect plagiarism (e.g. Plagiarism.org and Turnitin.com) from on-line resources, particularly term paper mills which can supply pre-made essays to students for a given fee. Services to track and monitor commercial content have also received increased interest as the media report more cases of stolen digital content (e.g. contentguard.com).

Various methods have been investigated for detecting plagiarism in natural language texts. Identifying plagiarism manually often begins with spotting inconsistencies within a text such as changes in the author’s writing style, or recognising passages with a familiar feel to them. When comparing between several texts, plagiarism detection involves finding similarities which are more than just coincidence and likely to be the result of copying or collaboration between multiple authors (i.e. *collusion*). In some cases, a single text is first read and certain characteristics found which suggest plagiarism. The second stage is to then find possible source texts either using tools such as web search engines for on-line sources, or manually checking non-digital collections.

Automatic methods of detection in natural language have originated from file comparison, IR, authorship attribution, and compression and copy detection. Techniques to investigate or detect similarity within a closed-set of texts (e.g. a group of student essays) include: the longest common subsequence (e.g. UNIX *diff*), approximate string match-

ing (e.g. MOSS (Schleimer/Wilkerson/Aiken 2003) and Turnitin.com), the overlap of longest common substrings (e.g. YAP (Wise 1996) and JPLAG (Prechelt/Malpohl/Philippse 2000)), the proportion of shared content words (CopyCatch (Woolls/Coulthard 1998)), the overlap of consecutive word sequences or word n-grams (e.g. Ferret (Lyon/Malcolm/Dickerson 2001), SCAM (Shivakumar/Garcia-Molina 1996), COPS (Brin/Davis/Garcia-Molina 1995), and Koala (Heintze 1996)), and compressed versions of the texts (Medori et al. 2002). Methods have also been developed to visualise the similarity between texts: VAST (Culwin/Lancaster 2000), Dotplot (Church/Helfman 1993), and Duploc (Ducasse/Rieger/Demeyer 1999).

While much research has been directed at finding possible copies and collusion between members of a closed-set, work has also been carried out on investigating plagiarism from open sets of documents such as the Web. Faced with a vast number of candidate source documents from which plagiarism could have taken place and given a test document to check for plagiarism, the problem breaks down into two parts: (1) narrow the set of candidate source documents as far as possible (2) compare each candidate source document with the test document for signs of plagiarism. Most plagiarism research assumes that (1) has been solved and focuses on methods for (2). However, (1) is not a trivial problem, as brute force comparison with all possible source documents is not possible. Typically some form of segmentation of the test document (by paragraph, sentence or fixed length window) is carried out and each segment used as a query to a search engine to retrieve a candidate set of documents for the stage (2) comparison. Another approach is to detect stylistic inconsistency within documents (see, e.g. Glover/Hirst 1996), as a possible signal of plagiarised material. Stylistically anomalous segments within texts can be used to focus the search for candidate source documents.

4.2. Copy and duplicate detection

Closely related to the problem of plagiarism detection, and arguably subsuming it, is that of copy and duplicate detection. Copy detection systems deal with storage, indexing, searching and the securing of documents in an electronic repository (e.g. a digital library). Work in this area has a somewhat different, broader focus than just plagiarism detection, and emerges as much from concerns about data management and integrity as from the narrow concern of demonstrating that one text contains material taken from another without acknowledgement. In particular, copy detection systems address issues of scale and efficiency, e.g. in detecting duplicates across very large document collections.

Leaving aside deliberate re-use of text, the rapid growth of texts available in electronic form (sometimes multiple forms per text) and hasty construction of digital text collections can easily lead to duplication within collections. Identifying such duplicates is an area of significant practical concern. Replication of documents in a collection can lead to skewed collection statistics. For search engines, identifying repeated information can help to eliminate redundant hits from search results (Chowdhury et al. 2002). As noted in the previous section, in plagiarism detection, narrowing the potential candidate sources of plagiarism is an important part of the process and copy detection systems are geared to support this. As Monostori/Zaslavsky/Schmidt (2001) state: "different styles

within the paper often reveal plagiarism, but without presenting the actual source it is hard to prove plagiarism."

Two methods exist for protecting documents from illicit or accidental replication: (1) prevention and (2) detection. Prevention involves the use of encryption or information hiding to inhibit copying, and detection involves discovering potential copies. It is common for copy detection systems to detect both complete or exact copies (duplicates), and partial copies from document collections that typically contain texts in various formats (e.g. HTML, Latex, Postscript, PDF and Microsoft Word), of various sizes and relating to different topics. Schleimer/Wilkerson/Aiken (2003) suggest that a copy detection system should have the following three desirable properties: (1) matches should be unaffected by formatting, whitespace and typological differences, (2) matches must be long enough to imply the material has been copied rather than matched coincidental language usage, and (3) changing the order of sentences in a document, inserting new sentences and removing sentences should not affect the set of discovered matches.

Two main approaches of copy detection exist (Hoad/Zobel 2003; Schleimer/Wilkerson/Aiken 2003): fingerprinting and ranking. Ranking is based on information retrieval techniques (see section 3.2.1.). Fingerprinting computes hash values (unique numbers) to represent some key content of the document (Manber 1994), typically in the form of n-grams. Similarity is then computed based on the number of fingerprints shared between documents. Four key areas in designing a fingerprinting process are: the generation of fingerprints (i.e. the hash function used), the granularity of fingerprints (i.e. the size of n-gram), the number of fingerprints used to represent a document (e.g. a fixed number or a function of document length), and the selection of suitable fingerprints to represent a document (e.g. removing non-discriminating fingerprints). While fingerprinting is not guaranteed to yield a unique representation of a document, it does offer a very efficient means to test large numbers of documents for duplication with a reasonable degree of accuracy.

Systems for copy detection include Siff (Manber 1994), COPy detection System or COPS (Brin/Davis/Garcia-Molina 1995), Stanford Copy Analysis Mechanism or SCAM (Shivakumar/Garcia-Molina 1995), Koala (Heintze 1996), CHECK (Si/Leong/Lau 1997), the n-gram ‘shingling’ approach of Broder (1998), MatchDetectReveal or MDR (Monostori/Zaslavsky/Schmidt 2001), and the Ferret system by Lyon/Malcolm/Dicker-son (2001). Hoad/Zobel (2003) compare fingerprinting and ranking approaches to detecting co-derivatives – pairs of documents both derived from a common third source – and propose a new measure for ranking documents, showing it to be superior to existing fingerprinting and ranking techniques.

4.3. Newswire monitoring

Being able to measure text re-use reliably and accurately is of great commercial interest to news agencies, as it may enable them to quantify the re-use of text by customers automatically (i.e. monitor re-use). Given the importance of the news agency’s role within the newspaper industry, accurately measuring text re-use could have the following practical benefits: enable the news agency to charge customers on a pay-per-usage basis rather than a flat-fee (more competitive and fairer); find “illegal” copies of newswire

copy on the Internet or within some other set of electronic documents; identify material no customers are using and hence eliminate redundant or unnecessary services; help establish trends and patterns about re-use by paying customers; and, enable them to determine the amount of self-re-use (i. e. re-duplication of the same text). Because news agencies generate large volumes of text on a daily basis, manually analysing re-use in the news is a complex and time-consuming process. Using computational techniques offers a practical way of generating re-use statistics for news agencies.

Using the METER corpus described in section 2.1., according to which a newspaper text is classified as to whether it is wholly derived, partially derived or non-derived from a newswire source, Clough/Gaizauskas/Piao (2002) have investigated three computational techniques for identifying text re-use automatically: n-gram matching (section 3.2.2.), sequence comparison (section 3.2.3.) and sentence alignment (section 3.2.4.). In the first approach, n-gram matches of varying lengths were used together with a containment score (Broder 1998); in the second a substring matching algorithm called Greedy String Tiling (Wise 1993) was used to compute the longest possible substrings between newswire-newspaper pairs; and in the final approach sentences between the source and candidate text pairs were automatically aligned.

Closely related to the problem of identifying newspaper re-use of newswire copy is that of finding different versions of the same news story. Church/Helfman (1993) use the Dotplot method (section 3.2.6.) to compare multiple news stories reporting the same event based on co-occurring character n-grams (consecutive sequences). Steinberger et al. (2003) use a number of measures including the overlap of fixed-length sequences (word n-grams) and words, and the length of newspaper to determine the most likely source and revisions of that source from a group of news stories. McKeown et al. (1999) describe a process for detecting similarities and differences between newswire texts reporting the same story to produce a coherent summary. Their methods of multi-document summarisation are used in a practical application at Columbia University called NewsBlaster (<http://www1.cs.columbia.edu/nlp/-newsblaster/>). Finally, Kirriemuir/Willett (1995) show how hierachic agglomerative clustering methods can be applied to the output of text searches on a database of newspapers, magazines, publications and reports from a major UK telecommunications provider. Based on term similarities, documents can be successfully grouped together which are identical or closely-related (i. e. revisions or rewrites of an existing article).

5. Conclusion

In this article we have presented *text re-use*, a problem in computational text analysis which is both intellectually interesting and of significant practical concern. Many different kinds re-use relationship can exist between texts – plagiarism, copy rewrite, duplication or reformatting – and solutions to identify these relationships automatically almost always involve computing some measure of similarity or difference. The selection of an appropriate method for automatically identifying text re-use involves determining which measure of similarity (or difference) best captures a derivation relation. Approaches to solving problems in related areas such as the retrieval of documents to fulfill a user's information needs, clustering documents according to some criterion, summaris-

ing multiple documents, aligning sentences from one language with those in another, finding exact and near duplicates of documents, detecting breaches of ownership, classifying documents according to genre, and identifying authorship attribution are all relevant to text re-use. The explosion in volume of digital text together with global networked access to text sources mean the study of text re-use and of methods to detect it automatically is likely to receive increasing attention in years to come.

6. Literature

- Aizawa, A. (2003), Analysis of Source Identified Text Corpora: Exploring the Statistics of the Re-used Text and Authorship. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL'03)*. Sapporo, Japan, 383–390.
- Baeza-Yates, R./Ribeiro-Neto, B. (1999), *Modern Information Retrieval*. Harlow, UK: Addison-Wesley.
- Barzilay, R./McKeown, K. (2001), Extracting Paraphrases from a Parallel Corpus. In: *Proceedings of ACL'01*. Toulouse, France, 50–57.
- Bell, A. (1991), *The Language of News Media*. Oxford: Blackwell.
- Bell, A. (1996), Text, Time and Technology in News English. In: Goodman, S./Graddol, D. (eds.), *Redesigning English: New Texts, New Identities*. London/New York: Routledge, 3–26.
- Bharat, K./Broder, A. (1999), Mirror Mirror on the Web: A Study of Host Pairs with Replicated Content. In: *Proceedings of the 8th International World Wide Web Conference*. Toronto, Canada, 1579–1590.
- Biber, D. (1988), *Variation across Speech and Writing*. Cambridge, UK: Cambridge University Press.
- Bilenko, M./Mooney, R./Cohen, W./Ravikumar, P./Fienberg, S. (2003), Adaptive Name Matching in Information Integration. In: *IEEE Intelligent Systems*, 18(5), 16–23.
- Bouayad-Agha, N./Kilgarriff, A. (1999), *Duplication in Corpora*. Technical Report ITRI-99-07, Information Technology Research Institute, University of Brighton, UK.
- Brin, S./Davis, J./Garcia-Molina, H. (1995), Copy Detection Mechanisms for Digital Documents. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. San Jose, CA, 398–409.
- Broder, A. Z. (1998), On the Resemblance and Containment of Documents. In: *Proceedings of the Compression and Complexity of Sequences*. Washington, DC: IEEE Computer Society, 21–29.
- Brown, P. F./Lai, J. C./Mercer, R. L. (1991), Aligning Sentences in Parallel Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA, 169–176.
- Catizone, R./Russell, G./Warwick, S. (1989), Deriving Translation Data from Bilingual Texts. In: *Proceedings of the First International Lexical Acquisition Workshop, (AAAI-89)*. Detroit, MI, 1–7.
- Chakrabarti, S. (2002), *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, CA: Morgan-Kaufmann Publishers.
- Chen, S. F. (1993), Aligning Sentences in Bilingual Corpora Using Lexical Information. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*. Columbus, OH, 9–16.
- Chowdhury, A./Frieder, O./Grossman, D./McCabe, M. C. (2002), Collection Statistics for Fast Duplicate Document Detection. In: *ACM Transactions on Information Systems* 20(2), 171–191.
- Church, K. W./Helfman, J. I. (1993), Dotplot: A Program for Exploring Self-similarity in Millions of Lines of Text and Code. In: *Journal of Computational and Graphical Statistics* 2(2), 153–174.
- Clough, P. D. (2000), *Plagiarism in Natural and Programming Languages: An Overview of Current Tools and Technologies*. Technical Report CS-00-05, Department of Computer Science, University of Sheffield, UK.

- Clough, P. D. (2003), Measuring Text Reuse. PhD thesis, University of Sheffield.
- Clough, P./Gaizauskas, R./Piao, S. L. (2002), Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Spain, 1678–1685.
- Clough, P./Gaizauskas, R./Piao, S. L./Wilks, Y. (2002), Measuring Text Reuse. In: *Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics*. Philadelphia, PA, 152–159.
- Coggins, J. M. (1999), Dissimilarity Measures for Clustering Strings. In: Sankoff, D./Kruskal, J. (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, 2nd edition. Stanford, CA: CSLI Publications, 311–321.
- Collberg, C./Kobourov, S. (2005), Self-plagiarism in Computer Science. In: *Communications of the ACM* 48(4), 88–94.
- Culwin, F./Lancaster, T. (2000), A Review of Electronic Services for Plagiarism Detection in Student Submissions. In: *Proceedings of 8th Annual Conference on the Teaching of Computing*. Edinburgh, UK. Available at: http://www.ics.heacademy.ac.uk/events/presentations/317_Culwin.pdf.
- Culwin, F./MacLeod, A./Lancaster, T. (2001), *Source Code Plagiarism in UK HE Computing Schools: Issues, Attitudes and Tools*. Technical Report SBU-CISM-01-01, School of Computing, Information Systems and Mathematics, South Bank University, London.
- Damashek, M. (1995), Gauging Similarity with N-grams: Language-independent Categorisation of Text. In: *Science* 267, 842–848.
- Doddington, G. (2002), Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In: *Proceedings of the Second International Conference on Human Language Technology*. San Diego, CA, 138–145.
- Dolan, W./Quirk, C./Brockett, C. (2004), Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: *Proceedings of COLING 2004*. Geneva, Switzerland, 350–356.
- Donaldson, J. L./Lancaster, A./Sposato, P. H. (1981), A Plagiarism Detection System. In: *ACM SIGSCI Bulletin* 13(1), 21–25.
- Ducasse, S./Rieger, M./Demeyer, S. (1999), A Language Independent Approach for Detecting Duplicated Code. In: *Proceedings of the International Conference on Software Maintenance (ICSM'99)*. Oxford, UK, 109–118.
- Foote, J./Cooper, M. (2001), Visualizing Musical Structure and Rhythm via Self-similarity. In: *Proceedings of the International Conference on Computer Music*. Havana, Cuba. Available at: <http://www.fxpal.com/publications/FXPAL-PR-01-152.pdf>.
- Fries, U. (1987), Summaries in Newspapers: A Textlinguistic Investigation. In: Fries, U. (ed.), *The Structure of Texts*. (Swiss Papers in English Language and Literature 3.) Tübingen: Gunter Narr Verlag, 47–63.
- Fung, P. (1995), A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In: *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA, 236–243.
- Gaizauskas, R./Foster, J./Wilks, Y./Arundel, J./Clough, P./Piao, S. (2001), The Meter Corpus: A Corpus for Analysing Journalistic Text Re-use. In: *Proceedings of the Corpus Linguistics Conference 2001*. Lancaster, UK, 214–223.
- Gale, W. A./Church, K. W. (1991), A Program for Aligning Sentences in Bilingual Corpus. In: *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 177–184.
- Gitchell, D./Tran, N. (1999), Sim: A Utility for Detecting Similarity in Computer Programs. In: *Proceedings of 13th SIGSCI Technical Symposium on Computer Science Education*. New Orleans, LA, 266–270.
- Glover, A./Hirst, G. (1996), Detecting Stylistic Inconsistencies in Collaborative Writing. In: Sharples, M./van der Geest, T. (eds.), *The New Writing Environment: Writers at Work in a World of Technology*. London: Springer, 147–168.

- Hannabuss, S. (2001), Contested Texts: Issues of Plagiarism. In: *Library Management* 22(6–7), 311–318.
- Hatzivassiloglou, V./Klavans, J./Eskin, E. (1999), Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-1999)*. College Park, MD, 203–212.
- Heckel, P. (1978), A Technique for Isolating Differences Between Files. In: *Communications of the ACM* 21(4), 264–268.
- Heintze, N. (1996), Scalable Document Fingerprinting. In: *Proceedings of the Second USENIX Workshop on Electronic Commerce*. Oakland, CA, 191–200.
- Helfman, J. (1993), Dotplot: A Program for Exploring Self-similarity in Millions of Lines of Text and Code. In: *Journal of Computational and Graphical Statistics* 2(2), 153–174.
- Helfman, J. I. (1996), Dotplot Patterns: A Literal Look at Pattern Languages. In: *Theory and Practice of Object Systems (TAPOS)* 2(1), special issue on patterns, 31–41.
- Hislop, G. W. (1998), Analyzing Existing Software for Software Re-use. In: *Journal of Systems and Software* 41, 33–40.
- Hoad, T. C./Zobel, J. (2003), Methods for Identifying Versioned and Plagiarized Documents. In: *Journal of American Society for Information Science and Technology* 54(3), 203–215.
- Holmes, D. I. (1994), Authorship Attribution. In: *Computers and the Humanities* 28(2), 87–106.
- Jing, H./McKeown, K. (1999), The Decomposition of Human-written Summary Sentences. In: *Proceedings of SIGIR'99*. Berkeley, CA, 129–136.
- Jurafsky, D./Martin, J. H. (2000), *Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.
- Karlsgren, J./Cutting, D. (1994), Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: *Proceedings of Coling 94*. Kyoto, Japan, 1071–1075.
- Kay, M./Röscheisen, M. (1993), Text-translation Alignment. In: *Computational Linguistics* 19(1), 121–142.
- Keeble, R. (1998), *The Newspapers Handbook*. London: Routledge.
- Kessler, B./Nunberg, G./Schütze, H. (1997), Automatic Detection of Text Genre. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 32–38.
- Kilgarriff, A./Grefenstette, G. (2003), Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics* 29(3), 333–347.
- Kimbrell, R. E. (1988), Searching for Text? Send an N-gram! In: *Byte* 13(5), 297–312.
- Kirriemuir, J. W./Willett, P. (1995), Identification of Duplicate and Near-duplicate Full-text Records in Database Search-outputs Using Hierarchic Cluster Analysis. In: *Program* 29(3), 241–256.
- Korfhage, R. R. (1997), *Information Storage and Retrieval*. New York: John Wiley.
- Kukich, K. (1992), Techniques for Automatically Correcting Words in Text. In: *ACM Computing Surveys* 24(4), 377–439.
- Lopresti, D. P. (2001), A Comparison of Text-based Methods for Detecting Duplication in Scanned Document Databases. In: *Information Retrieval* 4(2), 153–173.
- Lyon, C./Malcolm, J./Dickerson, B. (2001), Detecting Short Passages of Similar Text in Large Document Collections. In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. Philadelphia, PA, 118–125.
- Mallon, T. (1989), *Stolen Words: Forays into the Origins and Ravages of Plagiarism*. New York: Ticknor and Fields.
- Manber, U. (1994), Finding Similar Files in a Large File System. In: *Proceedings of 1994 Winter Usenix Technical Conference*. San Francisco, CA, 1–10.
- Mani, I./Bloedorn, E. (1999), Summarizing Similarities and Differences among Related Documents. In: *Information Retrieval* 1(1–2), 35–67.

- Mani, I./Maybury, M. (1999), *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press.
- Manning, C. D./Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Martin, B. (1994), Plagiarism: A Misplaced Emphasis. In: *Journal of Information Ethics* 3(2), 36–47.
- McEnery, A. M./Oakes, M. P. (2000), Authorship Identification and Computational Stylometry. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 545–562.
- McEnery, A./Wilson, A. (1996), *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McKeown, K./Klavans, J./Hatzivassiloglou, V./Barzilay, R./Easkin, E. (1999), Towards Multidocument Summarization by Reformulation: Progress and Prospects. In: *Proceedings of AAAI/IAAI*. Orlando, FL, 453–460.
- Medori, J./Atwell, E./Gent, P./Souter, C. (2002), Customising a Copying-identifier for Biomedical Science Student Reports: Comparing Simple and Smart Analyses. In: O'Neill, M./Sutcliffe, R. F. E./Ryan, C./Eaton, M./Griffith, N. (eds.), *Artificial Intelligence and Cognitive Science, 13th Irish International Conference, AICS 2002. Lecture Notes in Computer Science* 2464, 228–233.
- Melamed, D. I. (1999), Bitext Maps and Alignment Via Pattern Recognition. In: *Computational Linguistics* 25(1), 107–130.
- Metzler, D./Bernstein, Y./Croft, W. B./Moffat, A./Zobel, J. (2005), Similarity Measures for Tracking Information Flow. In: *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2005)*. Bremen, Germany, 517–524.
- Monostori, K./Zaslavsky, A./Schmidt, H. (2001), Efficiency of Data Structures for Detecting Overlaps in Digital Documents. In: *Proceedings of Australasian Computer Science Conference (ACSC '01)*. Gold Coast, Queensland, Australia, 140–147.
- Murata, M./Isahara, H. (2002), Automatic Extraction of Differences between Spoken and Written Languages, and Automatic Translation from the Written to the Spoken Language. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Spain. Available at: <http://citeseer.ist.psu.edu/551475.html>.
- Needleman, S. B./Wunsch, C. D. (1970), A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. In: *Journal of Molecular Biology* 48, 443–453.
- Osen, J. (1997), The Cream of Other Men's Wit: Plagiarism and Misappropriation in Cyberspace. In: *Computer Fraud and Security* 11, 13–19.
- Papineni, K./Roukos, S./Ward, T./Zhu, W.-J. (2002), BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of ACL*. Philadelphia, PA, 311–318.
- Petrakis, E. G. M./Tzeras, K. (2000), Similarity Searching in the CORDIS Text Database. In: *Software Practice and Experience* 30(13), 1447–1464.
- Piao, S. L./Rayson, P./Archer, D./McEnery, T. (2003), Extracting Multiword Expressions with a Semantic Tagger. In: *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, 49–56.
- Piao, S. L./Rayson, P./Archer, D./McEnery, T. (2004), Evaluating Lexical Resources for a Semantic Tagger. In: *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, 499–502.
- Prechelt, L./Malpohl, G./Philippssen, M. (2000), *JPlag: Finding Plagiarisms among a Set of Programs*. Technical Report 2000-1, Fakultät für Informatik, Universität Karlsruhe.
- Reah, D. (1998), *The Language of Newspapers*. London: Routledge.
- Ristad, S. E./Yianilos, P. N. (1998), Learning String Edit Distance. In: *IEEE Transactions on Pattern Recognition and Machine Intelligence* 20(5), 522–532.
- Rudman, J. (1998), The State of Authorship Attribution Studies: Some Problems and Solutions. In: *Computers and the Humanities* 31, 351–365.
- Samuelson, P. (1994), Self-plagiarism or Fair Use? In: *Communications of the ACM* 37(8), 21–25.

- Sanderson, M. (1997), *Duplicate Detection in the Reuters Collection*. Technical Report TR-1997-5, Department of Computing Science at the University of Glasgow.
- Sankoff, D./Kruskal, J. (eds.) (1983), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, 2nd edition. Stanford, CA: CSLI Publications.
- Schleimer, S./Wilkerson, D. S./Aiken, A. (2003), Winnowing: Local Algorithms for Document Fingerprinting. In: *Proceedings of SIGMOD Conference 2003*. San Diego, CA, 76–85.
- Shinyaama, Y./Sekine, S./Sudo, K./Grishman, R. (2002), Automatic Paraphrase Acquisition from News Articles. In: *Proceedings of the Second International Conference on Human Language Technology Research*. San Diego, CA, 313–318.
- Shivakumar, N./Garcia-Molina, H. (1995), Scam: A Copy Detection Mechanism for Digital Documents. In: *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*. Austin, TX. Available at: <http://citeseer.ist.psu.edu/shivakumar95scam.html>.
- Shivakumar, N./Garcia-Molina, H. (1996), Building a Scalable and Accurate Copy Detection Mechanism. In: *Proceedings of 1st ACM Conference on Digital Libraries DL'96*. Bethesda, MD, 160–168.
- Si, A./Leong, H. V./Lau, R. W. H. (1997), Check: A Document Plagiarism Detection System. In: *Proceedings of ACM Symposium for Applied Computing*. San Jose, CA, 70–77.
- Simard, M./Foster, G./Isabelle, P. (1992), Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation (TMI92)*. Montreal, Canada, 67–81.
- Smith, T. F./Waterman, M. S. (1981), Identification of Common Molecular Subsequences. In: *Journal of Molecular Biology* 147, 195–197.
- Soricut, R./Brill, E. (2004), A Unified Framework for Automatic Evaluation using 4-gram Co-occurrence Statistics. In: *Proceedings of ACL'04*. Barcelona, Spain, 613–620.
- Steinberger R./Pouliquen, B./Scheer, S./Ribeiro, A. (2003), Continuous Multi-source Information Gathering and Classification. In: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation*. Vienna, Austria. Available at: http://langtech.jrc.it/Documents/CIMCA-03_Ribeiro-Steinberger-et-al.pdf.
- Tichy, W. F. (1984), The String-to-string Correction Problem with Block Moves. In: *ACM Transactions on Computer Systems* 2(4), 309–321.
- Tiedemann, J. (1999), Automatic Construction of Weighted String Similarity Measures. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, MD, 213–219. Available at: <http://acl.ldc.upenn.edu/W/W99/W99-0626.pdf>.
- Uitdenbogerd, A./Zobel, J. (1999), Melodic Matching Techniques for Large Music Databases. In: *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1) (Orlando, Florida, USA) MULTIMEDIA '99*. New York: ACM Press, 57–66.
- Uzuner, O./Davis, R./Katz, B. (2004), Using Empirical Methods for Evaluating Expression and Content Similarity. In: *Proceedings of the 37th Hawaiian International Conference on System Sciences (HICSS-37)*. Big Island, HI, 40104a.
- van Dijk, T. (1988), *News Analysis: Case Studies of International and National News in the Press*. Hillsdale, NJ: Lawrence Erlbaum.
- Wagner, R. A./Fischer, M. J. (1974), The String-to-string Correction Problem. In: *Journal of the ACM* 21, 168–178.
- Waterhouse, K. (1993), *Waterhouse on Newspaper Style*. London: Penguin.
- Whale, G. (1990), Identification of Program Similarity in Large Populations. In: *The Computer Journal* 33(2), 140–146.
- Wilks, Y. (2004), On the Ownership of Text. In: *Computers and the Humanities* 38(2), 115–127.
- Willett, P. (1988), Recent Trends in Hierarchic Document Clustering: A Critical Review. In: *Information Processing and Management* 24(5), 557–597.
- Wise, M. (1993), *Running Karp-Rabin Matching and Greedy String Tiling*. Technical Report 463, Basser Department of Computer Science Technical Report, Sydney University.

- Wise, M. (1995), Neweyes: A System for Comparing Biological Sequences Using the Running Karp-Rabin Greedy String Tiling Algorithm. In: *Proceedings of Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, UK, 339–401.
- Wise, M. (1996), Yap3: Improved Detection of Similarities in Computer Programs and Other Texts. In: *Proceedings of SIGCSE'96*. Philadelphia, PA, 130–134.
- Wools, D./Coulthard, M. (1998), Tools for the Trade. In: *Forensic Linguistics* 5(1), 33–57.
- Wu, D. (2000), Alignment. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 415–458.

Paul Clough and Rob Gaizauskas, Sheffield (UK)

60. Corpora for text summarisation

1. Introduction
2. Background information
3. Corpora for single document summarisation
4. Corpora for multi-document summarisation
5. Corpora used to investigate characteristics of human summaries
6. Conclusions
7. Literature

1. Introduction

As in many other fields in computational linguistics, corpora play an important role in automatic summarisation and are used to both train and evaluate summarisation methods. Corpora have also been employed in a limited number of cases to investigate features of summaries in order to design summarisation methods and to learn more about human summarisation. This article discusses the issues which need to be addressed when corpora are built for automatic summarisation and how these corpora are exploited. The structure is as follows: in section 2 we provide a brief background to automatic summarisation and the use of corpora in the field. Section 3 describes the issues and methods involved in building corpora for single document summarisation. The use and creation of corpora for multi-document summarisation, including multilingual summarisation, is discussed in section 4, followed by section 5, which presents corpus based investigations of human produced abstracts. The article finishes with conclusions and notes on further readings/information.

2. Background information

Automatic summarisation is the process which produces summaries from one or more source texts using fully automatic means (Hovy 2003). First introduced in the late 1950s, the field of automatic summarisation experienced a rapid development in the late 1990s

- Wise, M. (1995), Neweyes: A System for Comparing Biological Sequences Using the Running Karp-Rabin Greedy String Tiling Algorithm. In: *Proceedings of Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, UK, 339–401.
- Wise, M. (1996), Yap3: Improved Detection of Similarities in Computer Programs and Other Texts. In: *Proceedings of SIGCSE'96*. Philadelphia, PA, 130–134.
- Wools, D./Coulthard, M. (1998), Tools for the Trade. In: *Forensic Linguistics* 5(1), 33–57.
- Wu, D. (2000), Alignment. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 415–458.

Paul Clough and Rob Gaizauskas, Sheffield (UK)

60. Corpora for text summarisation

1. Introduction
2. Background information
3. Corpora for single document summarisation
4. Corpora for multi-document summarisation
5. Corpora used to investigate characteristics of human summaries
6. Conclusions
7. Literature

1. Introduction

As in many other fields in computational linguistics, corpora play an important role in automatic summarisation and are used to both train and evaluate summarisation methods. Corpora have also been employed in a limited number of cases to investigate features of summaries in order to design summarisation methods and to learn more about human summarisation. This article discusses the issues which need to be addressed when corpora are built for automatic summarisation and how these corpora are exploited. The structure is as follows: in section 2 we provide a brief background to automatic summarisation and the use of corpora in the field. Section 3 describes the issues and methods involved in building corpora for single document summarisation. The use and creation of corpora for multi-document summarisation, including multilingual summarisation, is discussed in section 4, followed by section 5, which presents corpus based investigations of human produced abstracts. The article finishes with conclusions and notes on further readings/information.

2. Background information

Automatic summarisation is the process which produces summaries from one or more source texts using fully automatic means (Hovy 2003). First introduced in the late 1950s, the field of automatic summarisation experienced a rapid development in the late 1990s

as a result of the vast increase in the amount of information needing to be processed by humans.

Depending on the source which needs to be summarised, a distinction has to be made between *single document summarisation*, where summaries are produced from only one document, and *multi-document summarisation (MDS)*, which produces summaries from several, usually related, documents. A special case of multi-document summarisation is *multilingual summarisation* where the input texts are written in different languages. Some authors (see Mani 2001) differentiate between multilingual summarisation, where the texts to be summarised are written in several languages and the summary produced is in one of these languages, and cross-lingual summarisation, where the input is in one or several languages and the output is in a different language. This distinction does not influence the way corpora are produced for and used in summarisation, and for this reason the discussion in section 4.6. covers both multilingual and cross-lingual summarisation. When corpora are built for text summarisation, different problems need to be addressed for each of the main types introduced above.

Different approaches to the production of summaries mean that a distinction also needs to be made between *extracts* and *abstracts*. Extracts are summaries produced by concatenating parts from the source (i.e. paragraphs, sentences, clauses, etc.) without any modification. For abstracts, the important concepts are extracted from the source and then presented in a coherent way without keeping the original wording of the text. The majority of corpora produced for automatic summarisation are annotated in such a way that they can be used mainly by methods which produce extracts. However, recent advances in automatic summarisation evaluation in the Document Understanding Conferences (DUC) have proposed alternative approaches which lessen the need for a distinction between abstracts and extracts. As a result of this, unannotated corpora can be used by both extraction and abstraction methods (see section 4.2.).

A summary is supposed to retain only the most important information from the source. However, how this important information is defined depends very much on how a summary is used (Spärck Jones 2001). It is possible to have *general summaries* which try to present all the important topics in the source, and *user-focused summaries* which focus only on topics indicated by the user. General summaries are usually produced by single document summarisation methods, whereas user-focused summaries are most often created using multi-document summarisation methods.

The length of automatic summaries is generally expressed as a maximum number of words or as a percentage of either the number of sentences or words in the source (the compression rate). The fact that automatic summarisation methods can produce summaries of different lengths can pose problems when corpora are annotated because an annotation is often applied with one summary length in mind, and therefore the methods which exploit this annotated corpus will usually be restricted to this length only (see section 3.1. for more details). Radev/Jing/Budzikowska (2000) proposed methods to adapt annotated corpora for summaries of any length, depending on how the annotation was applied but these methods first require several people to annotate the same text.

The way most summarisation methods use annotated corpora assumes that a corpus is annotated with information which indicates the relevance of sentences to a document or to a specific topic. In this way, a summarisation method can extract a set of features considered important for each sentence, and then use the annotation to determine which combinations of features indicate sentences to be included in the summary. When a

corpus is used to evaluate a summarisation method, the sentences extracted by the method are compared with those marked as relevant in the corpus. This approach is very similar to approaches used in other fields of computational linguistics (see articles 23–32 in the preprocessing section). However, in contrast to other areas where the annotation can be defined more precisely (e.g. part-of-speech tagging, see article 23), it is very difficult to decide how to annotate a sentence for automatic summarisation. As a result, a strict set of guidelines which indicate how the important topic(s) of the document(s) can be identified is needed. These guidelines are dependent on the individual context and can differ greatly from corpus to corpus. Therefore, we do not present guidelines in this article, but focus instead on issues of corpus building and use. Questions which need to be addressed in the evaluation of automatic summarisation methods using corpora are also discussed.

The recent evaluation methods proposed in DUC do not necessarily require annotated corpora. Some of them rely on a gold standard of human produced summaries to assess automatic summaries. As in the case of annotated corpora, these summaries need to be produced according to some guidelines. In contrast to evaluation methods which use annotated corpora, these methods do not expect a perfect match between the sentences in the summaries and the sentences in the gold standard, instead measuring the extent to which the concepts from the gold standard are represented in the summaries.

3. Corpora for single document summarisation

Building annotated corpora for automatic summarisation has proved to be a daunting task because of the difficulty of defining what exactly an important unit (e.g. clause, sentence or paragraph) is. This difficulty stems from the fact that the decision as to whether a sentence is important enough to be marked as such is highly subjective, and as a result, the agreement between different annotators is often low. In all fields which use annotated corpora, inter-annotator agreement is important because it shows how reliably humans can identify the phenomenon to be annotated, giving an indication about the quality of the corpus and how useful it can be for a given task.

The annotation scheme normally used to label corpora for automatic summarisation is minimal in that it encodes information only about the “importance” of each sentence (section 3.1. gives more details about how “importance” is judged). In cases where a sentence is not important, no annotation is attached to the sentence. On top of this, additional information not normally used directly in the summarisation process, but which can have a beneficial impact on the quality of automatic summaries, can also be marked (Hasler/Orasan/Mitkov 2003).

There are several ways to produce corpora for single document summarisation. *Manual annotation* is the best-established one, and requires a human annotator to read the whole text and mark important units. Examples of this method are presented in section 3.1. The agreement computed between different annotators is discussed in section 3.2. Given that manual annotation is difficult and time-consuming, *automatic annotation methods* have been proposed as an alternative. An overview of these methods is presented in section 3.3.

Nowadays, most summarisation methods rely on corpora at least for evaluation. In an increasing number of cases, corpora are developed specifically for use in evaluation

conferences, where all participants perform the same task on a corpus prepared by the conference organisers. This means that the results of the evaluation are directly comparable. All of the major evaluation conferences focus on both single document and multi-document summarisation, in most cases with emphasis on multi-document summarisation, and are discussed in section 4.

3.1. Manually annotated corpora for automatic summarisation

The most common way to use corpora for single document summarisation requires human annotators to read the whole source text and manually mark units for importance according to a set of guidelines. The advantage of this is that an advanced program does not need to be employed to perform the annotation, as it is, for example, in the case of other tasks related to discourse (see article 29), coreference (see article 26) and multimodal corpora (see article 31), but the drawback is that it is time-consuming and difficult.

Corpora were first used to train and test summarisation methods by Edmundson (1969). Here, the most important sentences from a heterogeneous corpus consisting of 200 documents in the fields of physics, life science, information science and humanities were annotated by human judges. In order to ensure the consistency of the annotation, judges were asked to follow a set of guidelines and to select those sentences which indicate the subject area, why the research is necessary, how the problem is solved and the findings of the reported research. These rules broadly correspond to the moves in a scientific paper (Swales 1990). In addition, the annotators were advised to choose those sentences that minimise redundancy and maximise coherence.

In an experiment which attempted to establish how corpora can be used to evaluate automatic summarisation methods, Jing et al. (1998) asked 5 human judges to mark the important sentences in 40 documents from the TREC collection. Each person created two summaries for each document, one at 10% length and one at 20% length. Because there were several annotations for each document, the “ideal” summary was produced by considering the majority opinion of the human constructed summaries.

Hasler/Orasan/Mitkov (2003) present an enhanced annotated corpus which differs from the majority of available resources in that it contains more information. In addition to encoding information about the importance of sentences, it also indicates parts which can be removed from sentences marked as *essential/important*. Another innovation is that it provides a different label for those sentences which are not significant enough to be marked as important in their own right, but which have to be considered as they contain information essential for the understanding of the content of other sentences marked as *essential/important* (i. e. sentences that contain the antecedents of anaphoric expressions). These two types of additional information were added to give users an insight into the conciseness and coherence of summaries, respectively. The corpus consists of 163 annotated newswire and scientific texts totalling almost 150,000 words, with some texts being annotated by two or three annotators. In order to ensure consistency, a detailed set of guidelines was given to the annotators who were asked to identify sentences containing the most important 15% of the text (*essential* sentences), and an additional 15% which is the next most important (*important* sentences).

As we show in section 3.2., it is very difficult to obtain high agreement between annotators, especially when the texts are long (e. g. scientific papers). In order to make the annotation task easier, Kupiec/Pedersen/Chen (1995) and Teufel/Moens (1997) asked human judges to align sentences from author produced summaries with sentences from the full text. As expected, a one-to-one mapping is not possible, and therefore partial matches were also considered. In order to facilitate the alignment procedure, a set of guidelines which indicate when a match is possible was provided for the annotators.

In a corpus of 188 scientific and technical documents, Kupiec/Pedersen/Chen (1995) found that 79% of the sentences from the summary could be directly matched with sentences from the full text. Teufel/Moens (1997) observed that only 31.7% of the sentences in the summary could be matched with sentences from the full text in their corpus of 202 articles from the field of computational linguistics. Marcu (1999) found, in a smaller scale experiment, that only 15% of the clauses in 10 abstracts taken from the Ziff-Davis corpus, a corpus of computer product announcements, could be aligned with clauses from the full text. These noticeable differences in the results obtained by researchers suggest that the success of the alignment depends very much on the type of text used and how strictly matching is defined.

3.2. Experiments for selection of important units by humans

As mentioned in section 3.1., the most common approach for building a gold standard is to ask human judges to read a text and decide which units are important. These units tend to be sentences, but there are attempts to identify sub-sentential units such as elementary discourse units (Marcu 1997), which usually correspond to a clause, or to leave it to the judges to decide which are the best units for selection (Tsou et al. 1997). However, without clear guidelines about how to determine a unit, the annotation may not be very consistent.

Rath/Resnick/Savage (1961) asked six judges to extract the 20 most important sentences from each of 10 articles and only 1.6 sentences per article were agreed on. However, in a more recent study Marcu (1997) argues that it is possible to obtain annotation with a high degree of agreement between annotators. He describes an experiment in which 13 judges were asked to rate each unit from 5 texts as *very important*, *important* or *unimportant*. In order to facilitate their decision, the texts to be annotated were already split into units. Comparison between the annotations showed that the judges were consistent when they were asked to mark *very important* and *unimportant* units, but less consistent with what they considered *important*. Simple majority voting (i. e. more than 7 judges chose the same category for a unit) could be applied in 87% of the cases to decide the importance of a unit. Statistical significance tests showed that the agreement between annotators is significant.

In a similar experiment Tsou et al. (1997) asked 6 groups of evaluators, 3 from North China and 3 from Taipei, to mark the most important 10% of units in a text, and then 15% of the next most important units, without giving them exact instructions about how to identify a unit. The importance of each unit was computed using a weighted average measure called *perceived importance*. Comparison between the perceived importance of propositions showed average overall inter-group consistency between North China and

Taiwan, and high intra-group consistency. Therefore, the authors conclude that the background of the annotators plays an important role in selecting the important units.

Salton et al. (1997) found that two judges asked to extract paragraphs which represent the most important 20% of the text agree only in 46% of cases. The authors noticed that because the annotated articles had, in many cases, a general first paragraph that was often selected by both subjects, the agreement for the rest of the text was even lower. Jing et al. (1998) show that agreement between annotators tends to decrease as the length of the summaries increases.

Interesting results were obtained when inter-annotator agreement was computed on the corpus described by Hasler/Orasan/Mitkov (2003). When the kappa statistic (Siegel/Castellan 1988) was used to compute the agreement between annotators, values which indicate little agreement were obtained. Manual investigation of the selected sentences revealed that this low agreement value is caused by the fact that in many cases, the annotators marked different sentences which convey similar information. In light of this, cosine similarity (Salton/McGill 1983) was used to compute the overlap in the information covered by the annotated sentences. Using only the occurrence of words and not senses, the cosine similarity indicated a substantial overlap in the information present in the selected sentences.

An observation which needs to be made about these very different results obtained by different researchers for inter-annotator agreement is that, in addition to the factors already discussed in this section, agreement is also influenced by the type of texts which are processed. For example, Jing et al. (1998) explain that one reason for their high agreement is the fact that the annotated texts are newswire texts which have a very clear structure to facilitate the identification of important information. The length of the text is another factor which determines inter-annotator agreement. In longer texts the annotators have more decisions to make about whether a unit is important or not because of the amount of information presented. Moreover, in long texts, annotators' mood and tiredness, along with lapses in concentration, are more likely to influence the quality of the annotation.

3.3. Automatic annotation of corpora for summarisation

Despite the fact that it is difficult for humans to align units from a summary with units from the whole document, particularly in cases where the documents are long and contain specialised language, several methods which automatically produce this alignment have been proposed. The underlying idea is that humans often create a summary through copy-paste of at least parts of sentences from the whole document (Endres-Niggemeyer 1998), so it should be possible to produce this alignment automatically. The main advantage of such automatic methods is that they can be used to produce large-scale corpora for summarisation with minimum effort.

Marcu (1999) proposed a greedy method which eliminates units from the full document that do not reduce the similarity between a human produced summary and the full document. If, on the basis of the similarity measure, it is not possible to shorten the text further, the rhetorical structure of the reduced document is used to eliminate more clauses. The method was evaluated on 10 randomly selected articles with an average

length of 1,066 words from the Ziff-Davis corpus and results showed that it was close to human performance. The method was subsequently used to create a corpus of 6,942 texts with important clauses annotated. A drawback of this method is the fact that it does not allow any control over the number of sentences identified as important. This means that it cannot be employed to create corpora which can be used directly to train and evaluate methods producing summaries of a predefined length. Orasan (2005) adapted Marcu's method to address this problem, but concluded that for short summaries the method is not suitable and proposed a genetic algorithm instead.

Another method using (*full document, abstract*) pairs to align sentences from a summary with sentences from the whole document was proposed by Jing/McKeown (1999). The abstract is seen as a sequence of words, some of which appear in the document, and therefore the problem of alignment is reformulated as a problem of finding the most likely position of the words from the abstract in the full document using a Hidden Markov Model. The method was also evaluated on the Ziff-Davies corpus, and similar results were obtained to those reported in Marcu (1999).

Mani/Bloedorn (1998) annotate a corpus by computing for each sentence in a text a score based on its similarity with the human produced summary accompanying the full text. Once these scores were calculated, the highest scoring sentences were considered to be part of the summary. The corpus used consisted of 198 full-text articles and their author-supplied summaries published in the cmp-lg archive (Teufel 1999).

Given the difficulty of annotating texts manually and the errors found in fully automatic annotation methods, Orasan (2002) discusses a semi-automatic method to annotate corpora for summarisation. In order to facilitate the annotation process, a user-friendly tool which integrates several automatic annotation methods was developed. The purpose of these automatic methods is to draw the users' attention to sentences they consider important and let users decide whether the sentences should be annotated or not.

4. Corpora for multi-document summarisation

As mentioned in section 2, multi-document summarisation (MDS) is concerned with creating summaries from more than one source document, which means that in addition to the challenges in single-document summarisation, issues such as redundancy, compression ratio and passage selection become even more important (Goldstein et al. 2000). The temporal dimension of texts also comes into play as different documents may contain different information depending on the time they were constructed. Cross-document coreference (see article 27) has to be addressed because of the need to identify discourse entities referring to the same events/participants in order to produce accurate summaries. These aspects make corpora for MDS far more difficult to build than those for single-document summarisation, as much more information in many more texts needs to be annotated in order for these texts to be properly useful for training and evaluation.

As a result of these issues, the type of annotation produced differs in several ways from that used in single document summarisation. The first and foremost difference is the fact that it is no longer necessary to annotate sentences which cover general important information in the documents. Multi-document summarisation is usually employed

to produce user-focused summaries, and therefore sentences are annotated with information about how important they are for a given topic. The most common way to annotate a corpus for MDS is to decide on the topics of interest, use a reliable search engine to retrieve documents relevant to the query, and then ask human judges to mark how relevant each sentence from the top retrieved documents is to the selected topic. Radev/Jing/Budzikowska (2000) used *utility judgement* which requires a score from 0 to 10 to be assigned to each sentence, regarding how relevant that sentence would be to the topic.

Where corpora are intended to be used only to evaluate multi-document summarisation methods, humans are required to produce “ideal summaries” from clusters of documents relevant to a topic (see section 4.2.). In these cases, automatic summaries are evaluated manually or automatically by comparing them with the “ideal summaries”. In the DUC, the SEE tool (Lin 2001) is used by judges to manually assess the summaries, and the ROUGE test (Lin 2004) and pyramid evaluation (Nenkova/Passonneau 2004) to automatically measure the quantity of information in the summaries.

Given the challenges of creating corpora for multi-document summarisation, many researchers use readily available corpora which tend to come from well-established, large-scale evaluation conferences such as the Document Understanding Conferences, the Topic Detection and Tracking (TDT) initiative and the Text REtrieval Conferences (TREC). Some of these corpora are specifically designed for summarisation evaluation conferences, whilst others are developed for the evaluation of systems in related fields such as topic detection and tracking or information retrieval. These corpora are useful because they are of a suitable size, and already contain the annotations or gold standard texts needed to exploit them to their full potential. Given their importance for the summarisation community, most of the discussion in this section will focus on these corpora. Some of the corpora have also been used for single document summarisation tasks. Where relevant, a brief discussion about how they were used will be provided.

In addition to the corpora developed by evaluation conferences, there are some smaller-scale corpora which have been built specifically for exploitation in multi-document summarisation. These corpora have been developed by researchers wishing to further the development of MDS using specific annotations or for a specific type of summarisation.

4.1. Corpora from SUMMAC

SUMMAC was the first evaluation conference organised in the field of automatic summarisation (Mani et al. 1998). It was part of Phase III of the TIPSTER Text Program which finished in 1998. Given that the main purpose of this conference was to explore the evaluation methods available for text summarisation, a corpus was not created specially for the conference. The corpus used in this conference was derived from the TREC data and was slightly different from one task to another. In the ad-hoc task, which concentrated on indicative user-focused summaries, 20 topics, each with 50 documents, were extracted from the TREC data. The annotation available for these texts was the relevance of each document to a query available from the TREC corpus. For the classification task, which required judges to classify documents into predefined classes on the basis of their summaries, only 10 topics, each with 100 documents, were selected. The

topics of each document were again taken from the TREC corpus. It should be pointed out that SUMMAC evaluated both single and multi-document summaries, and therefore the resources can be used for both tasks.

4.2. Corpora from the Document Understanding Conferences (DUC)

The increasing importance of summarisation was acknowledged by the research community through the organisation of the DUC evaluation conferences every year. These conferences have been part of the DARPA TIDES (Translingual Information Detection, Extraction, and Summarization) program since 2001. The purpose of these conferences is to evaluate summarisation systems on different tasks using corpora distributed specifically for this purpose. Each year since it began, the tasks and corpora for DUC have changed or been extended, making extensive resources available for the Computational Linguistics/Natural Language Processing community. The corpora from the earlier conferences are not available to non-participants, but the Linguistic Data Consortium (LDC) distributes the corpora from the later conferences.

Some of the DUC tasks over the years have included automatic summarisation of single documents and of multiple documents on the same topic, creation of a short (100 word) summary by viewpoint and also in response to a question, and creation of a very short (10 word) summary for cross-lingual single documents. In order to evaluate the performance of the participating systems on these tasks, assessors select topics of interest from the datasets and produce summaries according to each task. Hence, the DUC corpora consist not only of collections of documents about the same topic (for multi-document summarisation) but also of human-produced responses to the tasks for evaluation purposes. DUC uses other existing text collections, namely TDT and TREC, to build their corpus for each conference.

As an example, DUC2003 used 30 document clusters of around 10 texts each from the TDT and TREC Novelty Track collections. These texts comprised the Financial Times of London, the Federal Register, the Los Angeles Times, AP Newswire, New York Times Newswire, Xinhua News Agency (English version) and the FBIS (Foreign Broadcast Information Service), a service which translates daily broadcasts into English. In DUC2004, this collection was used for training the participating systems which then used a corpus of 50 TDT English clusters (AP Newswire and New York Times Newswire), 25 TDT Arabic clusters (Agence France Presse Arabic Newswire translated into English by fully automatic Machine Translation systems) and 50 TREC English clusters (AP Newswire, New York Times Newswire, Xinhua News Agency English version). The sentences in the corpus were not annotated for their importance to a topic, instead summaries of different lengths were produced by humans and used as a gold standard by the ROUGE evaluation method (Lin 2004). In addition, topics were attached to each news cluster.

DUC2005 experimented with the pyramid evaluation method (Nenkova/Passonneau 2004) which assumes that there are several human produced summaries for each text and requires annotators to identify summary content units (SCUs), text spans which contain potentially useful information for a summary. After this identification, the quality of a summary is measured on the basis of how many SCUs it contains. Harnly et al. (2005) propose a method to identify SCUs automatically.

The DUC corpora have been enriched with additional information and used by researchers for other summarisation purposes. Zhou/Ticrea/Hovy (2004) augmented the DUC2003 corpus by annotating 130 biographies of 12 people with biographical information which referred to 9 elements (birth and death information, fame factor, personality, personal, social, education, nationality, scandal, work). Machine learning was applied to this corpus to train a system which can produce bibliographies. Newman et al. (2004) used the DUC2003 corpus provided to the participants for the very short summaries task to identify pairs of sentences which contain the same information. On the basis of approximately 1,300 short summaries produced by the DUC judges, 202 pairs of similar sentences were extracted to determine redundant sentences in multi-document summaries.

4.3. Corpora from the Text Summarization Challenge (TSC)

The developments in the field of automatic summarisation as a result of the above evaluation conferences prompted the research community to organise such conferences for other languages. The TSC is a summarisation evaluation programme in Japan initially based on the SUMMAC evaluation. This series of evaluation conferences aims to develop resources for Japanese and to investigate evaluation methods. The participants have to complete various tasks, such as single document summarisation and user-focused summarisation from multiple documents at different compression rates, and their systems are evaluated at sentence-level against a corpus of human gold standard summaries (Okumura et al. 2004). For the single document tasks in the evaluation, human annotators were required to annotate important sentences in each article at 10%, 30% and 50% compression rates, and to produce free abstracts at 20% and 40% compression rates. For the multi-document tasks, human judges were asked to produce free summaries from clusters of documents. All three TSC evaluations to date have used articles from the Mainichi Newspaper Database as their text collections, along with other news articles from the Web.

4.4. Corpora from the Topic Detection and Tracking (TDT) initiative

Some of the most used corpora which were not specifically designed for summarisation evaluation but have been widely used in the field are the TDT corpora. The TDT initiative is part of DARPA's TIDES programme and is designed to evaluate systems which perform the task of topic detection and tracking, i. e. identifying the first mention of a topic in a text collection and then following its development. This can help to determine which documents are topically related, identify new documents which can be classified as topically related to an existing set, build clusters of texts that discuss the same topic and detect changes between topically cohesive sections. These tasks are clearly related to issues in multi-document summarisation, which is why annotated TDT corpora prove such a popular choice for the development and evaluation of systems that produce summaries of multiple documents.

The TDT corpora consist of newswire, broadcast radio, broadcast TV and website documents from various sources in English, Mandarin and Arabic, although there are

most data for English. Each year the participants are given different tasks, which have included story segmentation, topic labelling, first story detection, story link detection, and most recently topic selection, topic definition and topic research. The corpora are split into training, development and testing sets and are currently annotated by human judges for topics which fall into several categories (Allan et al. 1998). The corpora from TDT1 through TDT5 are available from the LDC and are often used as corpora for evaluations in other conferences such as DUC and ACE (Automatic Content Extraction).

Barzilay/McKeown/Elhadad (1999) used sentences from the TDT corpora to train a system to paraphrase information in summaries. They developed a corpora of “themes” and through a manual analysis found 7 main types of paraphrase which could be used for paraphrasing rules to provide the basis for their comparison algorithm. The system was tested by evaluating the paraphrased sentences against a human gold standard from the corpus. Fukumoto/Suzuki (2000) noticed that the distinction between *topic* and *event* which is made in the TDT initiative can be used to produce multi-document summaries. In order to develop a method which extracts the key paragraphs from documents related to the same topic, the TDT corpora were used.

4.5. CSTBank

Radev/Otterbacher/Zhang (2003) developed CSTBank, a corpus annotated for *Cross Structure Theory* (Radev 2000), which could be useful for multi-document summarisation as it provides a theoretical model for issues that arise when trying to summarise multiple texts. Radev (2000) details *Cross Structure Theory* (CST), a theory describing relationships between two or more sentences from different source documents related to the same topic. CST is related to *Rhetorical Structure Theory* (RST) (Mann/Thompson 1988) but takes into account the features of multi-document structure and does not have an underlying tree representation or assume writers’ intentions. There are 18 domain-independent relations such as *identity*, *equivalence*, *subsumption*, *contradiction*, *overlap*, *fulfilment* and *elaboration* between texts spans. Radev argues that being aware of these relations during multi-document summarisation could help to minimise redundancy or include contradictions from different sources, and therefore improve the quality of the summary. CSTBank contains different clusters of documents arranged in families based on source texts and clustering methods, and is created from a number of text sources, including some other existing corpora. The data sources for CSTBank are TDT, DUC, TREC/ACQUAINT, TREC Novelty Track, the Penn Treebank, the Prague Dependency Treebank, Hong Kong News, NewsInEssence, various online news agencies and Usenet groups.

Radev/Otterbacher/Zhang (2004) describe the annotation process for Phase I of CSTBank, in which they used 8 human judges to manually annotate the first 5 of 6 clusters of related texts using CST relations. Before annotation, the judges attended a training session, and received annotation guidelines which contained 15 practice pairs of sentences in each section for the annotators to ensure sufficient understanding of the task and the relations. However, due to the fact that more than one CST relation can be allocated, as the relations are not mutually exclusive, it was difficult to measure the inter-annotator agreement and this had to be based on the existence of relations rather than the relation type.

4.6. Multilingual summarisation

A recent trend in the field of automatic summarisation is to develop multilingual automatic summarisation methods. As explained in section 2, these methods represent a special type of multi-document summarisation which produces summaries from documents written in different languages. Therefore, in addition to the above issues, building corpora for multilingual summarisation has the added challenge of dealing with documents in more than one language. As the judges involved in the annotation process need to know all the languages of the input documents, corpora developed for multilingual summarisation usually contain texts only in two languages and are built as part of large evaluation conferences such as DUC.

SUMMBank is an English-Chinese parallel corpus annotated for single and multi-document summarisation which was developed as part of the Summer 2001 Johns Hopkins Workshop (Radev et al. 2001). The corpus is based on the Hong Kong Newspaper Corpus, a parallel corpus distributed by the LDC which contains translations and near translations of English and Chinese news articles, local administration announcements and descriptions of municipal events. SUMMBank consists of a mixture of automatic summaries, human summaries, and documents and summaries relevant to 20 queries. The automatic summaries were produced using four automatic summarisation methods and two baseline methods. These six methods were run at different compression rates to produce over 100 million summaries. For the manual summaries, three human annotators from the LDC judged the importance of each sentence to a query using utility judgment (see beginning of section 3). Using the scores assigned by humans to the sentences, over 10,000 abstracts and extracts have been produced. Summaries of 5%, 10%, 20%, 30% and 40% compression rates were produced for single documents.

The Multilingual Summarization Evaluations (MSE) organised in 2005 and 2006 addressed the problem of multilingual and multi-document summarisation. The training data used in these evaluations contains 25 clusters from DUC2004. These clusters comprise news stories in English and Arabic, as well as automatic translations of Arabic texts into English. Human annotators at the LDC produced 100-word summaries which were used as a gold standard by the ROUGE evaluation method.

5. Corpora used to investigate characteristics of human summaries

Corpora are not used exclusively in automatic summarisation to train and test summarisation methods. In a number of cases corpora have been built to investigate the characteristics of human produced summaries in order to develop automatic summarisation methods or to teach people how to write high-quality summaries. The structure of human produced summaries constituted the focus of research for several researchers (Liddy 1991; Salager-Meyer 1990; Orasan 2001) who assembled corpora of author produced summaries. The analysis was manual and no annotated corpora were produced. In a similar way, Saggion/Lapalme (1998) manually analysed a corpus of (*abstract, patent*) pairs to identify specific parts in patents which can be used to produce summaries. Hasler (2007) investigated the characteristics of human and automatically produced news summaries and proposed a set of guidelines to improve their readability and coherence.

Teufel/Moens (1998) manually annotated sentences from scientific articles with information which indicates their role (e. g. PROBLEM, STATEMENT, CONCLUSIONS, etc.) and used this annotation to train a machine learning method. The purpose of this method developed by them was to produce summaries which are similar to those written by humans. Paice/Jones (1993) used the structure of papers published in the domain of crop husbandry to produce summaries. A corpus of relevant sentences was collected and recurring patterns such as *This paper studies the effect of AGEN on the HLP of SPE* were identified, where AGEN, HLP and SPE represent classes of terms specific to the domain. These patterns were then used to produce summaries.

Narita/Kurokawa/Utsuro (2002) built a parallel corpus of English-Japanese abstracts in order to help Japanese researchers to write abstracts in English. This corpus is designed to allow researchers to search for words and patterns in their native language and then to see how these words and patterns are expressed in English.

6. Conclusions

As in many other fields of computational linguistics, research in automatic summarisation has become increasingly dependent on corpus-based and/or corpus-driven approaches. This article has discussed the issues of creating, annotating and using corpora in automatic summarisation, distinguishing different types of summarisation attempted in the field.

Single document summarisation usually requires annotators to identify the important units of information in a text. Once these units are identified, they can be used by summarisation methods which learn how to recognise them, or by evaluation methods which assess the performance of summarisation methods. However, agreement between different annotators for such tasks is relatively low, indicating the difficulty of the annotation and subjectivity of the notion of “importance”.

Recent evaluation conferences (e. g. DUC) have developed new evaluation methods which do not necessarily require annotated corpora. Instead they require one or several human produced summaries which constitute a gold standard, and which can be regarded as a type of annotation applied to the document to be summarised. Such gold standards are also employed in multi-document summarisation, where marking the important units in corpora is not a feasible approach due to the large number of units which have to be considered. Researchers preparing corpora for multilingual summarisation need to tackle the issues related to corpora for multi-document summarisation and consider the multilingual dimension as well.

In many cases the corpora used in automatic summarisation are not specially built for this field. Instead, existing corpora are enriched with information which can be used in the summarisation process. This is particularly true in evaluation conferences where corpora developed for TREC, TDT and TIDES are used. These corpora are distributed by the Linguistic Data Consortium (LDC) and for this reason their catalogue, which can be accessed at <http://ldc.upenn.edu>, can prove an invaluable resource for researchers who want to find more information.

The proceedings from the Document Understanding Conferences are an excellent source of further information. They can all be accessed online at: <http://duc.nist.gov/>. At

the same address, further details about the distributed data and the guidelines can be found. Publicly available parts of the CSTBank corpus can be downloaded from <http://tangra.si.umich.edu/clair/CSTBank/>. The CAST corpus described in Hasler/Orasan/Mitkov (2003) is one of the few freely available corpora for automatic summarisation and can be accessed at: <http://clg.wlv.ac.uk/projects/CAST>. For an overview of the domain of automatic summarisation, the <http://www.summarization.com> website is a very good starting point containing lists of links and a comprehensive bibliography.

7. Literature

- Allan, J./Carbonell, J./Doddington, G./Yamron, J./Yang, Y. (1998), Topic Detection and Tracking Pilot Study: Final Report. In: *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. Lansdowne, VA, 194–218.
- Barzilay R./McKeown K. R./Elhadad M. (1999), Information Fusion in the Context of Multi-document Summarization. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, MD, 550–557.
- Edmundson H. P. (1969), New Methods in Automatic Extracting. In: *Journal of the Association for Computing Machinery* 16(2), 264–285.
- Endres-Niggemeyer, B. (1998), *Summarizing Information*. Berlin: Springer.
- Fukumoto F./Suzuki Y. (2000), Extracting Key Paragraphs Based on Topic and Event Detection – Towards Multi-document Summarization. In: *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, 31–39.
- Goldstein, J./Mittal, V. O./Carbonell, J./Kantrowitz, M. (2000), Multi-document Summarization by Sentence Extraction. In: *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, 40–48.
- Harnly, A./Nenkova, A./Passonneau, R./Rambow, O. (2005), Automation of Summary Evaluation by the Pyramid Method. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 2005)*. Borovets, Bulgaria.
- Hasler, L./Orasan, C./Mitkov, R. (2003), Building Better Corpora for Summarisation. In: *Proceedings of Corpus Linguistics 2003*. Lancaster, UK, 309–319.
- Hasler, L. (2007), From Extracts to Abstracts: Human Summary Production Operations for Computer-aided Summarisation. PhD Thesis, University of Wolverhampton, UK.
- Hovy, E. (2003), Text Summarization. In: Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 583–598.
- Jing, H./Barzilay, R./McKeown, K./Elhadad, M. (1998), Summarization Evaluation Methods: Experiments and Analysis. In: *Intelligent Text Summarization. Papers from the AAAI Spring Symposium*. Stanford, CA, 60–68.
- Jing, H./McKeown, K. (1999), The Decomposition of Human-written Summary Sentences. In: *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*. Berkeley, CA, 129–136.
- Kupiec, J./Pedersen, J./Chen, F. (1995), A Trainable Document Summarizer. In: *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval*. Seattle, WA, 68–73.
- Liddy, E. D. (1991), Discourse-level Structure of Empirical Abstracts: An Exploratory Study. In: *Information Processing and Management* 27(1), 550–581.

- Lin, C.-Y. (2001), Summary Evaluation Environment (SEE). <http://www1.cs.columbia.edu/nlp/tides/SEEManual.pdf>.
- Lin, C.-Y. (2004), ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedings of the Workshop on Text Summarization Branches out*. Barcelona, Spain, 74–81.
- Mani, I. (2001), *Automatic Summarization*. Amsterdam: John Benjamins.
- Mani, I./Bloedorn, E. (1998), Machine Learning of Generic and User-focused Summarization. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI'98)*. Madison, WI, 821–826.
- Mani, I./House, D./Klein, G./Hirshman, L./Orbst, L./Firmin, T./Chrzanowski, M./Sundheim, B. (1998), *The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report MTR 98W0000138*. McLean, VA: The Mitre Corporation.
- Mann, W./Thompson, S. (1988), Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. In: *Text* 8(3), 243–281.
- Marcu, D. (1997), The Rhetorical Parsing of Natural Language Texts. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97/EACL'97)*. Madrid, Spain, 96–103.
- Marcu, D. (1999), The Automatic Construction of Large-scale Corpora for Summarization Research. In: *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*. Berkeley, CA, 137–144.
- Narita, M./Kurokawa, K./Utsuro, T. (2002), A Web-based English Abstract Writing Tool using a Tagged E-J Parallel Corpus. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 2115–2119.
- Nenkova, A./Passonneau, R. (2004), Evaluating Content Selection in Summarization: The Pyramid Method. In: *Proceedings of NAACL-HLT2004*. Boston, MA, 145–152.
- Newman, E./Doran, W./Stokes, N./Carthy, J./Dunnion, J. (2004), Comparing Redundancy Removal Techniques for Multi-document Summarisation. In: *The Proceedings of STAIRS*. Valencia, Spain, 223–228.
- Okumura M./Fukushima, T./Nanba, H./Hirao, T. (2004), Text Summarization Challenge 2: Text Summarization Evaluation at NTCIR Workshop 3. In: *SIGIR Forum* 38(1), 29–38.
- Orasan, C. (2001), Patterns in Scientific Abstracts. In: *Proceedings of Corpus Linguistics 2001 Conference*. Lancaster, UK, 433–443.
- Orasan, C. (2002), Building Annotated Resources for Automatic Text Summarisation. In: *Proceedings of LREC-2002*. Las Palmas, Spain.
- Orasan, C. (2005), Automatic Annotation of Corpora for Text Summarisation: A Comparative Study. In: *Lecture Notes in Computer Science* 3406, 670–681.
- Paice, C./Jones, P.A. (1993), The Identification of Important Concepts in Highly Structured Technical Papers. In: Korfhage, R./Rasmussen, E./Willett, P. (eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, PA, 69–78.
- Radev, D. (2000), A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-document Structure. In: *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong, People's Republic of China, 74–83.
- Radev, D./Jing, H./Budzikowska, M. (2000), Centroid-based Summarisation of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In: *ANLP/NAACL Workshop on Summarization*. Seattle, WA, 21–30.
- Radev, D./Otterbacher, J./Zhang, Z. (2003), *CSTBank: Cross-document Structure Theory Bank*. Available at: <http://tangra.si.umich.edu/clair/CSTBank>.
- Radev, D./Otterbacher, J./Zhang, Z. (2004), CSTBank: A Corpus for the Study of Cross-document Structural Relationship. In: *Proceedings of Language Resources and Evaluation Conference (LREC 2004)*. Lisbon, Portugal.
- Radev, D./Teufel, S./Saggion, H./Lam, W./Blitzer, J./Çelebi, A./Qi, H./Drabek, E./Liu, D. (2001), *Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework*. Johns

- Hopkins Summer Workshop Final Report. Baltimore, MD: Center for Language and Speech Processing, Johns Hopkins University.
- Rath, G. J./Resnick, A./Savage, R. (1961), The Formation of Abstracts by the Selection of Sentences. Part I: Sentence Selection by Man and Machines. In: *American Documentation* 12(2), 139–141.
- Saggion H./Lapalme, G. (1998), The Generation of Abstracts by Selective Analysis. In: *Intelligent Text Summarization. Papers from the AAAI Spring Symposium*. Stanford, CA, 130–132.
- Salager-Meyer, F. (1990), Discoursal Flaws in Medical English Abstracts: A Genre Analysis per Research- and Text-type. In: *Text* 10(4), 365–384.
- Salton, G./McGill, M. J. (1983), *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Salton, G./Singhal, A./Mitra, M./Buckley, C. (1997), Automatic Text Structuring and Summarization. In: *Information Processing and Management* 33(3), 193–207.
- Siegel, S./Castellan, N. J. (1988), *Nonparametric Statistics for the Behavioral Sciences*. 2nd edition. New York: McGraw-Hill.
- Spärck Jones, K. (2001), Factorial Summary Evaluation. In: *Proceedings of the 1st Document Understanding Conference*. New Orleans, LA.
- Swales, J. (1990), *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Teufel, S. (1999), Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, University of Edinburgh.
- Teufel, S./Moens, M. (1997), Sentence Extraction as a Classification Task. In: *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain, 58–59.
- Teufel, S./Moens, M. (1998), Sentence Extraction and Rhetorical Classification for Flexible Abstracts. In: *Intelligent Text Summarization. Papers from the AAAI Spring Symposium*. Stanford, CA, 16–25.
- Tsou, B. K./Lin, Hing-Lung/Lai, B. Y. T./Chan, S. W. K. (1997), A Comparative Study of Human Efforts in Textual Summarization. In: *Proceedings of PACLING '97*. Tokyo, Japan, 324–332.
- Zhou, L./Ticrea, M./Hovy, E. H. (2004), Multi-document Biography Summarization. In: *Proceedings of EMNLP-2004*. Barcelona, Spain, 434–441.

*Constantin Orasan, Laura Hasler
and Ruslan Mitkov, Wolverhampton (UK)*

61. Quantitative methods in corpus linguistics

1. Introduction
2. Corpus design issues
3. The unit of analysis in corpus-based studies
4. Type A designs: Corpus-based studies of a linguistic feature
5. Type B designs: Corpus-based studies of texts and text categories
6. Comparing Type A and Type B designs for register analyses
7. The role of inferential statistics in corpus linguistics
8. Conclusion
9. Literature

- Hopkins Summer Workshop Final Report. Baltimore, MD: Center for Language and Speech Processing, Johns Hopkins University.
- Rath, G. J./Resnick, A./Savage, R. (1961), The Formation of Abstracts by the Selection of Sentences. Part I: Sentence Selection by Man and Machines. In: *American Documentation* 12(2), 139–141.
- Saggion H./Lapalme, G. (1998), The Generation of Abstracts by Selective Analysis. In: *Intelligent Text Summarization. Papers from the AAAI Spring Symposium*. Stanford, CA, 130–132.
- Salager-Meyer, F. (1990), Discoursal Flaws in Medical English Abstracts: A Genre Analysis per Research- and Text-type. In: *Text* 10(4), 365–384.
- Salton, G./McGill, M. J. (1983), *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Salton, G./Singhal, A./Mitra, M./Buckley, C. (1997), Automatic Text Structuring and Summarization. In: *Information Processing and Management* 33(3), 193–207.
- Siegel, S./Castellan, N. J. (1988), *Nonparametric Statistics for the Behavioral Sciences*. 2nd edition. New York: McGraw-Hill.
- Spärck Jones, K. (2001), Factorial Summary Evaluation. In: *Proceedings of the 1st Document Understanding Conference*. New Orleans, LA.
- Swales, J. (1990), *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Teufel, S. (1999), Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, University of Edinburgh.
- Teufel, S./Moens, M. (1997), Sentence Extraction as a Classification Task. In: *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain, 58–59.
- Teufel, S./Moens, M. (1998), Sentence Extraction and Rhetorical Classification for Flexible Abstracts. In: *Intelligent Text Summarization. Papers from the AAAI Spring Symposium*. Stanford, CA, 16–25.
- Tsou, B. K./Lin, Hing-Lung/Lai, B. Y. T./Chan, S. W. K. (1997), A Comparative Study of Human Efforts in Textual Summarization. In: *Proceedings of PACLING '97*. Tokyo, Japan, 324–332.
- Zhou, L./Ticrea, M./Hovy, E. H. (2004), Multi-document Biography Summarization. In: *Proceedings of EMNLP-2004*. Barcelona, Spain, 434–441.

*Constantin Orasan, Laura Hasler
and Ruslan Mitkov, Wolverhampton (UK)*

61. Quantitative methods in corpus linguistics

1. Introduction
2. Corpus design issues
3. The unit of analysis in corpus-based studies
4. Type A designs: Corpus-based studies of a linguistic feature
5. Type B designs: Corpus-based studies of texts and text categories
6. Comparing Type A and Type B designs for register analyses
7. The role of inferential statistics in corpus linguistics
8. Conclusion
9. Literature

1. Introduction

As documented in the articles of this handbook, the central goal of corpus-based analysis is to describe and interpret generalizable patterns of language use. The subdisciplines of discourse analysis and functional linguistics are similar to corpus linguistics in having a primary interest in research questions relating to language use, but these disciplines form a cline in the extent to which they rely on quantitative methods and the value attached to the generalizability of findings. Discourse analytic studies often focus on detailed discussion of a few texts (see, e.g., the survey of studies contained in Schiffрин/Tannen/Hamilton 2001). Functional linguistic studies often focus on linguistic generalizations, with relatively little attention paid to the representativeness of the texts analyzed. In contrast, corpus linguistic investigations of language use are usually designed as quantitative studies with the goal of generalizable findings representing some domain of use.

The design of a corpus is a fundamentally important consideration for achieving this goal: the corpus must be representative of the target domain of use in order for subsequent analyses to be generalized to that domain (see also article 9). Quantitative analyses are also important for generalizable results, because they provide a measure of the extent to which a pattern holds in different domains of use. Computational techniques figure prominently in corpus analyses simply because they enable studies of a scope not feasible otherwise. Computers make it possible to identify and analyze complex patterns of language use, tracking the use of multiple linguistic features across large text collections. Further, computers provide consistent, reliable analyses; they do not change their minds or become tired during an analysis.

The present article provides a survey of the major research designs used in quantitative corpus-based analysis. The article begins with a brief overview of issues of corpus design, recognizing the central place of corpus representativeness for this research approach. We then organize the remainder of the article around the three major kinds of research questions investigated with corpus-based analysis: studies with the word as the primary unit of analysis; studies with a linguistic feature as the primary unit of analysis; and studies with texts as the primary unit of analysis. Finally, we briefly discuss the role of inferential statistics in corpus-based studies of language use.

2. Corpus design issues

To serve as the basis for linguistic investigations, corpora must be designed to represent particular registers, dialects, or other domains of use (see article 9). Part II of this handbook documents several of the most important domains of use that corpora have been designed to represent: written registers (article 10), spoken registers (article 11), historical genres (article 14), learner language (especially for learners of English; article 15), electronic texts (including computer-mediated communication and web registers; see articles 17–18), and translations of ‘parallel texts’ in two or more languages (article 16).

The representativeness of the corpus limits the kinds of research questions that can be addressed and the generalizability of the results of the research. For example, a corpus composed of newspaper texts would not provide the basis for a general investigation of

variation in English; it would not represent the patterns of use found in spoken registers or even in most written registers.

Representativeness is determined by two considerations: composition and size. The composition of a corpus refers to the text categories included in the design of the corpus. The size of a corpus refers to the number of words or number of texts in the corpus (see below).

Sampling methods are used to select texts in a way that represents the target text categories and size. There are two general approaches to sampling: proportional and stratified. A **proportional sample** represents groups to the extent that they occur in the larger population. For example, proportional samples are useful for political surveys, where predicting the final vote depends on the proportion of each demographic subgroup. However, proportional samples are usually not useful for studies of linguistic variation and use. For example, if we collected all the language produced and received for a week by residents of a city in the U.S., we could identify the actual proportions of language varieties that these people experienced – probably something like 80 % conversation, 10 % television shows, 1 % newspapers, 1 % novels, 2 % meetings, 2 % radio broadcasts, 2 % texts that they wrote (memos, email messages, letters), and 2 % other texts (signs, instructions, specialist written texts, etc.). A proportional corpus of this type would tell us how often a person is likely to encounter a certain word in the course of a typical week. However, a proportional corpus would be of limited use for studies of variation, because most of the corpus would be conversation. We could not use such a corpus to study language use patterns in other registers, because they would not be adequately represented. Even a popular register such as newspaper language would be minimally represented in such a corpus; specialist registers such as legal documents or medical/scientific research articles would be virtually nonexistent in the corpus. However, a general study of linguistic variation in English would normally want to consider the full range of registers considered to be important, regardless of their proportional use by typical speakers over the course of a normal week.

A **stratified sample** is constructed for these latter research purposes. A stratified corpus is designed to represent the full range of linguistic variation that exists in a language, not the proportions of variation. In a stratified corpus, the researcher begins by identifying the text categories that are the focus of research; texts are then selected from each of those categories so that each category is adequately represented. The text categories could be register differences, dialect differences, or other discourse domains, depending on the primary research questions.

Register variation is central to descriptions of language use: all speakers of a language control many registers, and every time people speak or write, they must choose a register. Therefore, a well-designed corpus will usually focus on a single specialized register, or it will be designed to represent a range of registers. Regional and/or social dialects are also important to consider if the research questions relate to dialect differences. In addition, more specific parameters, such as subject matter, can be important for some research questions.

The second major consideration that determines corpus representativeness is size. Although corpus size is often measured by the total number of words in the corpus, other considerations are at least as important: the number of texts from different categories, and the number of words in each text sample. If too few texts are included, a single text can have an undue influence on the results of an analysis. Additionally, the number

of texts is the most important consideration for studies with the text as the unit of analysis (see section 5). Enough texts must be included in each category to capture variation across speakers or authors (see Biber 1990, 1993).

The number of samples from a text also deserves attention, because the characteristics of a text can vary dramatically internally. A clear example of this is experimental research articles, where the introduction, methods, results, and discussion sections all have different patterns of language use. Thus, sampling that did not include all of these sections would misrepresent the language patterns found in research articles.

Finally, the number of words in each sample is important for providing a reliable count of features in a text. Most earlier corpora used relatively small text samples (1,000–2,000 words), while recent corpora often include complete texts (sometimes over 50,000 words long).

Lexicographic studies (see section 4.2.) require particularly large corpora. Many words and collocations occur with low frequencies, and a corpus must contain many millions of words, taken from many different texts, to enable investigations of word use. However, for all kinds of research, both size and composition are important considerations.

Biber (1993) describes how the representativeness of a corpus can be investigated empirically, depending on the patterns of variation for target linguistic features. For example, rare linguistic features (e.g., clefts) and features that show greater variability (e.g., relative clauses) will require larger samples – both longer texts and a greater number of texts. In contrast, common features with stable distributions, such as pronouns, can be represented accurately in a relatively small sample. However, all linguistic features vary across registers (see the survey of grammatical features in the *Longman Grammar of Spoken and Written English*; Biber et al. 1999). Thus, although the distribution of common features like pronouns and nouns can be represented accurately with relatively short text samples, a single-register corpus (like newspapers) would tell us nothing about the use of nouns and pronouns in other registers (like conversation or fiction).

3. The unit of analysis in corpus-based studies

One of the first decisions required when carrying out a quantitative corpus-based analysis is to determine what the unit of analysis is. This is a crucial decision because it determines the object of the research and the way data should be collected and organized, which in turn limits the research questions that can be asked and the statistical techniques that can be applied (see Biber/Conrad/Reppen 1998, 269–274).

Corpus-based studies generally have one of two primary research goals: 1) to describe the variants and use of a linguistic structure, or 2) to describe differences among texts and text varieties, such as registers or dialects.

Three major types of research design have been employed in corpus research. The primary difference among these research design types is the unit of analysis, which in turn makes each design type appropriate for one of the above two research goals. In Type A studies, the unit of analysis is each occurrence of a linguistic feature. Type A studies are thus designed for Research Goal 1 (describing the variants of a linguistic structure). In Type B studies, the unit of analysis is each individual text. Type B studies

are thus designed for Research Goal 2 (describing the differences among texts and text varieties). Finally, in Type C studies, the unit of analysis is the entire corpus (or different subcorpora). Type C studies can be used for either Research Goal 1 or 2, but they do not permit the use of inferential statistics (see below).

These units of analysis are the ‘observations’ of the study. For the purposes of quantitative analysis, the main difference between the three research design types is the nature of these observations. That is, the observations in Type A studies do not have quantitative characteristics, while the observations in Type B studies are analyzed in terms of quantitative characteristics. Type C studies differ from both of the others in that there are actually very few observations – usually only 2 or 3 observations – because each subcorpus is treated as an observation.

For example, a Type A study of relative clauses might have the goal of predicting the choice of relative pronoun (*who*, *which*, *that*). For this purpose, we could analyze each relative clause to record characteristics such as whether the clause is restrictive or non-restrictive and whether the head noun is human, animate but not human, or inanimate. In this research design, there are three **variables**: relative pronoun type, head noun type, and clause type. All three variables are **nominal**, meaning that the values are simply categories. Most importantly for our purposes here, these variables are *not* numeric: they do not describe greater-than or lesser-than characteristics. For example, there is no sense in which the relative pronoun *who* is quantitatively greater or lesser than *which*.

In contrast, the observations in Type B studies (each text) have quantitative characteristics, so they can be analyzed with respect to true numeric variables. For example, in a Type B comparison of newspapers and academic prose, we would treat each text as an observation. These texts could be analyzed to determine the rate of occurrence of linguistic features, such as nouns, verbs, and relative clauses. In this case, these variables have an **interval** scale, meaning that they represent greater-than and lesser-than relationships, with a unit of 1 being fixed. For example, a text with 17.5 relative clauses per 1,000 words has a greater rate of occurrence than a text with 14.5 relative clauses per 1,000 words. Because there are many observations in a Type B study (i.e., the total number of texts in the corpus), it is possible to compute mean scores and standard deviations, and to use inferential statistics to compare the rates of occurrence across text categories (see 6 below).

Type C studies are similar to Type B studies in that the observations have quantitative characteristics. That is, in a Type C study, each subcorpus is treated as an observation, and it is possible to compute the rate of occurrence for linguistic features in each subcorpus. However, in this case there are only 2–3 observations included in the entire study; and each text variety is represented by only a single observation (i.e. the subcorpus for that variety). As a result, it is not possible to compute a mean score or standard deviation in a Type C study, and no inferential statistics are possible.

Notice that although the three research design types address fundamentally different kinds of research questions, they could all be based on the same corpus. In Type A studies, however, the focus is on accounting for the variants of a single linguistic feature. In contrast, Type B studies describe the differences among texts.

In general, most previous corpus-based studies have used either a Type A design or a Type C design. This handbook provides several examples of both types of research design. For example, the articles on historical corpora (14), learner corpora (15), discourse analysis (49), dialectology (53), contrastive studies (54), collocations (58), and

grammatical colligations (43) all illustrate research with Type A designs. Article 36, on “Statistical methods for corpus exploitation”, provides detailed information on the statistical analysis of Type A designs. Several other articles in the handbook illustrate Type C designs, including the articles on language teaching (7), distributions in text (37) and recent language change (52). In contrast, corpus-based studies with Type B designs are less common; the article on multidimensional approaches (38) discusses several of those studies.

In the following sections, we provide more detailed descriptions and case studies of each research design type. In section 4, we describe Type A studies, which focus on the variants of a linguistic feature. In this section we also introduce studies that have individual words as the unit of analysis – a special case of Type A (see also article 37). Such lexicographic studies have been one of the most important applications of corpora, and a number of special analytical techniques have been developed specifically for this type of study; thus we treat these research designs in a separate subsection (4.2.). In section 5, we describe Type B studies, which compare the typical linguistic characteristics of different texts. Then, in section 5.3. we discuss Type C studies, which compare the typical linguistic characteristics of different subcorpora.

4. Type A designs: Corpus-based studies of a linguistic feature

Corpus-based studies that use Type A designs are focused on particular linguistic features. There are two main kinds of Type A design: those that focus on the contextual factors that lead language users to choose one variant of a linguistic feature over another, and those that focus on the co-occurrence of words (or *collocation*). In section 4.1. we describe studies of linguistic variants, and in section 4.2. we describe corpus-based studies of collocation.

4.1. Corpus-based studies of linguistic variants

4.1.1. The goal

Many functional studies of linguistic variation use a corpus-based approach. The subfield of functional linguistics is based on the premise that linguistic variability is not arbitrary; rather, there are systematic contextual factors that influence the choice of one linguistic variant over another. Corpus linguistics provides an ideal approach to the investigation of linguistic variation, because it allows the researcher to observe numerous tokens of the linguistic feature in natural contexts.

The goal of such studies is to analyze the distribution of linguistic variants across a range of contexts, in order to predict the choice of one variant over another (see also article 43). Several corpus studies have investigated the contextual and functional differences among seemingly equivalent linguistic variants, such as:

- dative movement (*give John a ball* versus *give the ball to John*);
- particle movement with phrasal verbs (*look up the answer* versus *look the answer up*);
- raising and extraposed constructions (*John is difficult to please* versus *It is difficult to please John* versus *To please John is difficult*);
- that*-deletion (*I think that/0 I should be able to go*);

Similar studies have investigated the use of related constructions, even when these are not strictly equivalent. These include studies of:

- active vs. passive constructions (*researchers consider many factors versus many factors were considered*);
- that*-clauses vs. *to*-clauses (*I hope that I can go* vs. *I hope to go*);
- WH*-clefts vs. *it*-clefts (*What happened three years ago was that I decided to go back to school* vs. *It was three years ago that I decided to go back to school*)

Corpus investigations are used to study natural occurrences of these constructions, to isolate the influence of factors such as phrase length, pronominal versus full noun objects, topic continuity, informational prominence, and purely lexical factors. Studies of this type include Thompson/Mulac (1991), Prince (1978), Collins (1991), and Oh (2000). The *Longman Grammar of Spoken and Written English* (LGSWE) also includes many corpus-based analyses of this type (Biber et al. 1999).

4.1.2. A case study: Complementizer *that* versus 0

An example of a Type A study is an analysis of the contextual factors that influence the choice between the complementizer *that* and 0 in *that*-clauses. These variants seem equivalent in many contexts, as in:

I do not think that the situation is slipping out of control.

versus

I don't think [] any of us would be willing to do that.

In a study of this linguistic choice, each occurrence of a *that*-complement clause is a separate observation. For each *that*-clause, we would record whether *that* or 0 was used; this is the variable that we are trying to predict. Other variables would record the characteristics of the context, so that we can determine which contextual factors are most strongly associated with each variant.

Several contextual factors might be influential in making the choice between these two variants (see Thompson/Mulac 1991; Biber et al. 1999, 681–683). For example, the complementizer *that* might be omitted more often with common matrix verbs, such as *think*, than with less common verbs, such as *show*. It also might be the case that the complementizer *that* is omitted more often with first person pronouns as subject than with other subjects. Finally, register can also be a contextual factor for individual linguistic features; thus, it might be that the complementizer *that* is omitted more often in conversation than in other registers. (Several other contextual factors turn out to be influential in this case. These include co-referential subjects in the matrix clause and *that*-clause; presence of an intervening noun phrase between the controlling verb and *that*-clause; and whether the controlling verb is in active or passive voice. For the sake of simplicity, the case study here is restricted to three variables.)

To investigate the relative influence of these contextual factors, we would code a large sample of *that*-clause constructions, where each occurrence of a clause constitutes a separate observation. The output of such an analysis might look like Table 61.1.

Tab. 61.1: Coded observations for the analysis of *that*-omission

Complementizer	Matrix verb	Subject	Register
<i>that</i>	indicate	noun	academic
<i>that</i>	suggest	noun	academic
<i>that</i>	imply	noun	academic
0	say	pro-he	newspaper
<i>that</i>	argue	noun	newspaper
<i>that</i>	say	pro-I	newspaper
0	think	pro-I	newspaper
<i>that</i>	report	pro-they	newspaper
0	think	pro-I	conversation
0	say	pro-I	conversation
<i>that</i>	feel	pro-I	conversation
0	think	pro-I	conversation
0	think	pro-he	conversation
0	know	pro-I	conversation

Each line represents information about a single observation, i. e., a single occurrence of a *that*-clause. Each column represents the values for a different variable. The first column shows whether the complementizer *that* is present or omitted; this is the linguistic choice that we are trying to predict. The other columns represent contextual factors: the second column records the matrix verb; the third column records the type of matrix-clause subject; and the fourth column records the register that the example was taken from. For example, the first line in the output above gives the codes for the following sentence:

This analysis indicates that plant growth and nutrient uptake are directly linked. (Academic)

In this case, the complementizer *that* is present; the matrix verb is *indicates*; the matrix clause subject is a full noun phrase (*this analysis*); and the sentence is taken from an academic text. Note that the values for matrix verb have been consolidated, to represent the different verb lemmas (e. g., INDICATE), rather than the individual inflected forms (e. g., *indicated*, *indicates*).

Data such as these allow quantitative analyses to determine the association of different contextual factors with each structural variant. The simplest kind of analysis is to compute simple frequency counts. In research designs of this type, where each occurrence of a linguistic feature represents an observation, the variables have nominal rather than quantitative values; that is, the values represent different categories. For example, the variable ‘matrix verb’ has nominal values, such as ‘indicate’, ‘suggest’, and ‘imply’.

The variable ‘complementizer’ has two values: ‘*that*’ and ‘0’. If we compute frequencies for these values from the data given in Table 61.1, we would obtain the results given in Table 61.2.

Tab. 61.2: Frequencies of complementizer variants from Table 61.1.

Complementizer	Frequency
0	7
<i>that</i>	7
Total:	14

We could similarly compute frequencies of the values for the variable ‘Matrix verb’, giving the results listed in Table 61.3.

Tab. 61.3: Frequencies of matrix verb variants from Table 61.1.

Matrix verb	Frequency
think	4
say	3
indicate	1
suggest	1
imply	1
argue	1
report	1
feel	1
know	1
Total:	14

In cases like this, we can simplify the results by grouping all values that occur only one time; see Table 61.4.

Tab. 61.4: Frequencies of matrix verb groups from Table 61.1.

Matrix verb	Frequency
think	4
say	3
other verbs	7
Total:	14

Simple frequency tables can be combined to produce ‘cross-tabulation’ tables (or ‘cross-tabs’), which display the frequency counts for each combination of values across variables. Cross-tabs show the extent to which contextual factors are associated with each linguistic variant. Table 61.5 is a cross-tabulation table for the observations coded above.

Tab. 61.5: Cross-tabulation frequencies of complementizer choice by matrix verb

Complementizer	Matrix verb			
	think	say	other	Total
0	4	2	1	7
<i>that</i>	0	1	6	7
total	4	3	7	14

This distribution indicates that matrix verb is an important factor influencing the choice of complementizer. Six of the seven clauses with an omitted complementizer have the matrix verbs *think* or *say*; in contrast, only one of the seven clauses with *that* have the matrix verb *think* or *say*. (Large-scale corpus analysis of these constructions shows that the verbs *think* and *say* are by far the most common verbs controlling *that*-clauses. Thus, the zero-complementizer is favored by the presence of the most common controlling verbs, while *that*-complementizer is favored by the presence of other controlling verbs; see Biber et al. 1999, 680–683.)

Register can also be used as a variable in a cross-tabulation table. Table 61.6 shows the findings compiled from Table 61.1 above.

Tab. 61.6: Cross-tabulation frequencies of complementizer choice by register.

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
that	3	3	1	7
total	3	5	6	14

From this distribution we can see that retention is favored in academic prose (all 3 clauses), but omission is favored in conversation (5 of 6 clauses). Of course, many more observations are needed as the basis for an actual study of this type.

Given a large enough data set, we could also consider the influence of multiple factors at the same time. For example, we could contrast the influence of the matrix verb for *that*-clauses in news reportage with the influence of this variable in conversation.

Inferential statistical techniques can be used to analyze the significance and strength of these associations. The chi-squared test is the simplest of these techniques, while VAR-BRUL analysis allows for much more sophisticated investigations of this type. We briefly discuss the use of statistical tests in section 7 below.

4.2. Corpus-based studies of the co-occurrence of words

One important application of corpus-based research has been to study the extended meanings and patterns of use associated with individual words (see article 58). For example, corpus-based research has shown that the most obvious meaning of a word often turns out to *not* be the most common meaning.

Corpus-based analyses of individual words often rely on the construct of ‘collocation’: how a word tends to occur together with other specific words. Often, words that are supposedly synonymous are shown to be strikingly different in terms of their collocations. For example, the verbs *turn*, *go*, and *come* can all be used as resulting copulas, with a meaning of ‘to become’ or ‘to change to a different state’. However, a consideration of their collocates immediately shows that these copular verbs are dramatically different in their extended meanings (see the *Longman Grammar of Spoken and Written English*, Biber et al. 1999, 444–445). The copular verb *turn* usually collocates with color adjectives (e.g., *turn white*) or adjectives that describe physical appearance (e.g., *turn pale*).

The copular verb *go* usually collocates with negative adjectives, like *bad*, *crazy*, and *nuts*. In contrast, the copular verb *come* collocates with *true* and with adjectives like *alive* and *clean*. By considering the set of common collocations, it is often easy to distinguish among related words that had previously been regarded as synonymous.

Corpus-based studies of collocation make use of a Type A design, where each occurrence of the target word is treated as an observation. The word immediately preceding the target word might be one variable, and the following word could be a second variable. For example, coded observations for the target word *blue* might look like those given in Table 61.7.

Tab. 61.7: Coded observations for an analysis of the collocates of *blue*.

Preceding word	Target word	Following word
your	blue	eyes
the	blue	sky
had	blue	and
her	blue	eyes
a	blue	napkin

Once a large number of observations have been coded, it is possible to identify important collocates of the target word by computing frequencies for each of the values of a variable. It is important to note such frequencies are not variables. Rather, frequencies are simply counts of how often each value occurs for a variable. The above example included two variables: the word positions immediately preceding and immediately following the target word. The actual words that precede or follow the target word are the values of those variables; *eyes*, *sky*, *and*, and *napkin* are values of the variable ‘following word’ in the example above. Frequency counts are a kind of descriptive statistic that tells us how often each of these values occurs.

To illustrate, consider the three most frequent ‘following words’ (or right collocates) of *blue* in a 1.67-million-word corpus of fiction; see Table 61.8.

Tab. 61.8: Frequencies of the three most frequent right collocates of *blue*. (Note the corpus is a sub-sample from the *Longman Spoken and Written English Corpus*.)

Target word	Following word	Frequency
blue	eyes	39
blue	and	25
blue	sky	11

Such frequency information identifies the most common combinations, which can be interpreted as collocations. Other combinations, such as *blue napkin*, rarely occur and so should not be regarded as collocations.

Simple frequency information can present a biased measure of the strength of a collocation, because very frequent words are likely to occur together simply by random chance. For example, the combination *and the* occurs 3,248 times in this fiction corpus, even though we would not normally consider this to be a strong collocation.

An alternative method to assess the strength of collocations is to compare the frequency of a combination with the likelihood that the words will occur together simply by chance. The most commonly used statistical measure for this purpose is the ‘mutual information score’: a ratio of the observed frequency (f_o) of the combination divided by the expected frequency (f_e) of the combination:

$$\text{Mutual Information Score} = f_o / f_e$$

(The original formula converts the result to a base-2 logarithm; see Church/Hanks 1990; Church et al. 1991. Many practitioners, however, use the simpler version given above, because the results are more easily interpretable; see Stubbs 1995; Barnbrook 1996. Article 58 provides a detailed discussion of collocation statistics.)

The expected frequency is the frequency if the combination were to occur merely by chance; it is computed by multiplying the total frequencies of the two words, divided by the corpus size (1,672,055 words in the present case):

$$\text{Expected frequency (}f_e\text{)} = (\text{Target word frequency} * \text{Collocate word frequency}) / \text{Total corpus size}$$

The observed frequencies of the individual words in the above example are:

Target word:

blue 371

Possible collocates:

eyes 1,649

and 49,598

sky 379

Using the formula above, we can compute the expected frequency (f_e) for *blue eyes*:

$$f_e(\text{blue eyes}) = (371 * 1649) / 1,672,055 = .37$$

The expected frequencies for the other two combinations are:

$$f_e(\text{blue and}) = (371 * 49598) / 1,672,055 = 11.0$$

$$f_e(\text{blue sky}) = (371 * 379) / 1,672,055 = .08$$

Notice that the expected frequency reflects the absolute frequencies of the individual words. For example, *and* is extremely common by itself, and therefore we can expect to find the combination *blue and* occurring together by random chance (shown by the expected frequency of 11.0). In contrast, both *blue* and *sky* rarely occur by themselves, and it is therefore extremely unlikely that we would find this combination by random chance (shown by the expected frequency of only .08).

The mutual information score compares the size of the actual observed frequency of a word combination to its expected frequency. For example, the mutual information score for *blue eyes* is:

$$\text{Mutual info (blue eyes)} = f_o(\text{blue eyes}) / f_e(\text{blue eyes}) = 39 / .37 = 105.4$$

This is a relatively large mutual information score, showing that this combination represents a strong collocation. In contrast, the mutual information score for *blue and* shows that this combination represents a much weaker collocation:

$$\text{Mutual info } (\textit{blue and}) = 25 / 11.0 = 2.3$$

At the other extreme, the mutual information score for *blue sky* indicates that this combination represents an even stronger collocation than *blue eyes*:

$$\text{Mutual info } (\textit{blue sky}) = 11 / .08 = 137.5$$

Although the mutual information index gives us a measure of the strength of association between two words, it can be misleading with particularly infrequent words. For example, imagine that the pair of words *blue marlin* appears once in our corpus, and that the word *marlin* appears only twice. We can then calculate the expected frequency of *blue marlin*:

$$f_e(\textit{blue marlin}) = (371 * 2) / 1,672,055 = .0004$$

Because the expected frequency of *blue marlin* is so low (due to the low frequency of the word *marlin*), the mutual information index for *blue marlin* is misleadingly high:

$$\text{Mutual info } (\textit{blue marlin}) = 1 / .0004 = 2500$$

The t-score is a second statistic used with collocations. While the mutual information index gives a measure of the strength of association between two words, the t-score is used to contrast the collocates of two supposedly synonymous words. Church et al. (1991) use t-scores to contrast the collocates of *strong* and *powerful*. For example, the words *showing*, *support*, *defense*, and *economy* are much more likely to occur as collocates of *strong* than with *powerful*. In contrast, the words *figure*, *minority*, *military*, and *presidency* are much more likely to occur as collocates of *powerful*.

Mutual information scores are the most commonly used measure of collocation, because they are easy to interpret: they simply measure the strength of association between a target word and a potential collocate. T-scores are more difficult to interpret because they are measures of dissimilarity, contrasting the possible collocates of two target words.

5. Type B designs: Corpus-based studies of texts and text categories

The second major type of research design in corpus linguistics examines differences between texts and text categories. Texts can have many nominal characteristics, such as the register category of the text, whether the text is spoken or written, whether the author is female or male, etc. However, texts also have quantitative characteristics: rates of occurrence for linguistic features. It is these quantitative characteristics that distin-

guish Type A and Type B research designs. In the following subsection, we discuss the methods for computing normed variables, which express the rates of occurrence for linguistic features in texts. Then, in section 5.2. we present a case study with a Type B research design. In section 5.3., we discuss a special kind of Type B design that treats subcorpora as units of analysis.

5.1. Normed rates of occurrence

When corpus-based studies examine counts of features across texts, it is important to make sure that the scores are comparable (see also articles 36, 37, and 41). In particular, if the texts in a corpus are not all the same length, counts from those texts are not directly comparable. For example, imagine that you analyzed two texts and found that each one has 20 modal verbs. It might be tempting to conclude that modals are equally common in the texts. However, further imagine that the first text has a total length of 750 words, and the second text is 1,200 words long. Because the second text is longer, there are more opportunities for modals to occur; therefore simply comparing the raw counts does not accurately represent the relative rate of occurrence of modals in the two texts.

‘Normalization’ is a way to convert raw counts into rates of occurrence, so that the scores from texts of different lengths can be compared. Normalization takes into account the total number of words in each text. Specifically, the raw counts are divided by the number of words in the text and then multiplied by whatever basis is chosen for norming. To continue with the example above, the counts in the two texts could be normed to a basis per 1,000 words of text as follows:

Text A:

$$(20 \text{ modals} / 750 \text{ words}) \times 1000 = 27.5 \text{ modals per 1,000 words}$$

Text B:

$$(20 \text{ modals} / 1200 \text{ words}) \times 1000 = 16.7 \text{ modals per 1,000 words}$$

You can see from these normed rates of occurrence that the raw counts are very misleading in this case; that is, modal verbs are actually considerably more common in Text A than in Text B.

In the above example, counts were normed to a basis of 1,000 words since both texts were approximately this long. In a corpus with shorter texts, counts might be normed to rates per 500 words of text. When working with very short texts, such as the writing of children, it might even be necessary to norm counts to rates per 100 words of text. If a higher basis is adopted, the counts for rare features can be artificially inflated – sometimes dramatically so. For example, if a student text of 80 words happened to have one passive construction (a generally rare feature in elementary student writing), and the texts were normed to a basis of 100 words, the text would have a normed score of 1.25 passives per 100 words. However, if that same count was normed to a basis of 1,000 words, the normed value would be 12.5 passives per 1,000 words, which represents a rate of occurrence unlikely to be achieved in any extended elementary student writing. Thus, counts should be normed to the typical text length in a corpus.

5.2. A case study: Newspaper and conversation texts

Normalized rates of occurrence for any linguistic feature can be analyzed as ‘interval’ variables (see section 3 above). Coded observations in this type of study would look like the display in Table 61.9. Each line in the display represents information about one text, and each column gives the values for a variable. In this case, two of these variables are nominal: text identification and register. The other four variables are interval, giving the total word count and quantitative rates of occurrence for past tense verbs, attributive adjectives, and first person pronouns. The scores for the three linguistic features have been normalized to a rate per 1,000 words of text.

Tab. 61.9: Linguistic data for twelve texts

Text ID	Register	Word count	Past tense	Attrib adjs	1st person pronouns
n1.txt	news	2743	47.4	68.1	3.1
n2.txt	news	1932	49.2	63.0	9.2
n3.txt	news	2218	42.2	74.8	7.1
n4.txt	news	2383	45.3	72.1	2.2
n5.txt	news	1731	47.1	67.3	5.4
n6.txt	news	2119	51.2	70.0	5.2
c1.txt	conv	2197	32.2	43.1	62.6
c2.txt	conv	2542	37.4	36.3	59.1
c3.txt	conv	2017	36.8	39.7	58.7
c4.txt	conv	1896	29.2	35.2	65.5
c5.txt	conv	1945	31.3	34.0	58.2
c6.txt	conv	2072	23.8	38.3	60.4

Because the observations in this type of study (i.e., the texts) are described with respect to quantitative variables, it is possible to compute descriptive statistics like a mean score (expressing the central tendency) and standard deviation (expressing dispersion). For example, the mean score for past tense verbs in the data above is:

$$(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2 + 32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 12 = 39.4$$

It is similarly possible to compute separate mean scores for each register, as in:

Mean score of past tense verbs for newspapers:
 $(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2) / 6 = 47.1$

Mean score of past tense verbs for conversations:
 $(32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 6 = 31.8$

A second descriptive statistic – the ‘standard deviation’ – indicates the extent to which texts are dispersed away from the mean score. Descriptive statistics like these can also be used for inferential statistical techniques like t-test or ANOVA, to test the statistical ‘significance’ of observed differences among registers (i.e., by comparing the size of the mean differences among registers relative to the size of the differences among the texts within a register). Inferential statistics are discussed further in section 7 below.

5.3. Type C designs: Corpus-based studies with subcorpora as the unit of analysis

It is also possible to use different subcorpora as observations, to compute rates of occurrence of linguistic features in each subcorpus (Research Design Type C). In this case, each subcorpus is treated as if it were a large text. This is the design used for the most part in the *Longman Grammar of Spoken and Written English*, where the rates of occurrence for grammatical features (per one million words) are compared across conversation, fiction, newspapers, and academic prose. For example, the following calculations compare the overall normed rates of occurrence for past tense verbs in the subcorpora of conversation and newspapers:

Conversation (corpus size = 3,929,500 words):

$$(113,170 \text{ past tense verbs} / 3,929,500) * 1,000,000 = 28,800 \text{ per million words}$$

Newspapers (corpus size = 5,432,800 words):

$$(204,273 \text{ past tense verbs} / 5,432,800) * 1,000,000 = 37,600 \text{ per million words}$$

Studies with this research design give similar kinds of findings to those where each text is treated as an observation. There is a major difference though: in this design, we compute a single, overall rate of occurrence for each register, based on the subcorpus for that register. In contrast, in studies with texts as the unit of analysis, we compute a rate of occurrence for each text, which provides the basis for computing mean scores and standard deviations. Thus, when the text is the unit of analysis, we can compute both the typical rate of occurrence (the mean score) and the extent to which individual texts vary away from that typical score (the standard deviation). This provides the basis for inferential statistical tests (such as ANOVA and correlational techniques); such tests are not possible with designs that use subcorpora as the units of analysis. Despite this fact, subcorpora are commonly used as the units of analysis in studies where the differences across registers are so large that inferential statistics are not essential (see section 7 below).

6. Comparing Type A and Type B designs for register analyses

Because the observations in Type B designs are texts, this kind of analysis is especially well suited to comparing the linguistic characteristics of registers (or other text categories), as in section 5.2. above. However, Type A designs can also be used to study register differences. For example, Table 61.6 above compares the use of *that* and 0 complementizers across three registers (academic prose, newspapers, and conversations).

Although both design types can be used to analyze register differences, there is a crucial distinction in the information that they provide: Type B designs are based on rates of occurrence, and so they tell us how common a linguistic feature is. In contrast, Type A designs tell us the relative preference for one variant over another, but we have no way of knowing how common the features actually are.

Tab. 61.10: Cross-tabulation frequencies of complementizer choice by register

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
<i>that</i>	3	3	1	7
total	3	5	6	14

For example, Table 61.6 in section 4.1. above (repeated here as Table 61.10) shows that *that*-retention is favored in academic prose (3 out of 3 clauses), while *that*-omission is favored in conversation (5 of 6 clauses). Based on this table, we might be tempted to conclude that *that*-retention occurs more commonly in academic prose than in conversation. However, a Type A design does not provide the basis for such conclusions: this study provides no information about the corpora used for academic prose and conversation, and so we are unable to determine the actual rates of occurrence. That is, Table 61.6 only gives us proportional information: when a *that*-clause occurs in conversation, it is likely to omit the complementizer. When a *that*-clause occurs in academic prose, it is likely to retain the complementizer. However, Table 61.6 does not tell us the actual rates of occurrence for *that*-clauses in each register.

In fact, it is possible to imagine a scenario where *that*-retention occurs more commonly in conversation, even though it is the dispreferred variant in that register.

Imagine, for example, that the data in Table 61.6 are based on a 50,000-word corpus of academic prose and a 10,000-word corpus of conversation. With this background information, we can compute rates of occurrence for each corpus:

$$\text{Normed rate of occurrence of } \textit{that}-\text{retention in academic prose:} \\ (3 \text{ clauses} / 50,000 \text{ words}) \times 100,000 = 6 \text{ clauses per 100,000 words}$$

$$\text{Normed rate of occurrence of } \textit{that}-\text{retention in conversation:} \\ (1 \text{ clause} / 10,000 \text{ words}) \times 100,000 = 10 \text{ clauses per 100,000 words}$$

In this case, the actual distribution turns out to be the exact opposite of the apparent difference seen in Table 61.6: the rate of occurrence for clauses with *that*-retention is actually higher in conversation than in academic prose. This happens because *that*-clauses overall are many times more common in conversation than in academic prose. As a result, even the dispreferred variant in conversation (with *that*-retention) has a relatively high rate of occurrence.

The important point here is that Type A research designs do not provide the basis for determining rates of occurrence, so they cannot be used to determine if a feature or variant occurs more commonly in one register or another. This is potentially confusing, and even published research studies sometimes make this mistake. Type A studies do tell us what the preferred variant is in a register, and how registers differ in their reliance on a particular variant. For example, Table 61.6 above shows that when a *that*-clause is used in academic prose, it will usually retain the complementizer. When a *that*-clause is used in conversation, it will usually omit the complementizer. This is a genuine register difference. However, it would be incorrect to therefore conclude that *that*-retention is

more common in academic prose. Even though a much higher proportion of *that*-clauses retain the complementizer in academic prose, the actual rate of occurrence for this variant could be higher in conversation – Table 61.6 does not tell us one way or the other.

7. The role of inferential statistics in corpus linguistics

Inferential statistics can be used with all corpus-based research designs to assess whether observed differences might have occurred simply due to chance (see articles 36 and 37 for more detailed discussion). In the case of a Type A study, only non-parametric techniques can be applied, because all variables are nominal. The most commonly used non-parametric technique in this type of study is the chi-squared test. It is also possible to use multivariate statistical techniques for this type of design, such as loglinear regression and VARBRUL.

In contrast, it was pointed out in section 5 that Type B research designs have true numeric variables, which permit descriptive statistics such as mean scores and standard deviations. These designs allow the use of parametric statistical techniques, such as t-test and ANOVA to test for differences across categories, and Pearson correlations to test for relationships among linguistic variables. Multivariate statistical techniques used with these designs include multiple regression, factor analysis, and discriminant analysis.

Inferential statistical tests help to identify meaningful differences, as opposed to differences that occur just due to random chance. However, we would argue that inferential statistics should be used and interpreted with caution in corpus-based research. Tests of statistical significance depend on the sample size (N): as the sample size becomes larger, the difference among groups required to achieve significance becomes smaller. For very large samples – the normal case in corpus-based research studies – relatively small differences between groups are considered significant.

A complementary statistic is the measure of strength, which indicates the importance of a quantitative difference or relationship. With very large samples, it is easy to find small linguistic differences that are statistically significant but not strong; we would argue that these differences are often not interesting, because they do not reflect the important differences across text categories. By also considering measures of strength, researchers can identify the linguistic differences that are important and therefore more interesting for interpretation.

8. Conclusion

In this article, we have attempted to provide an overview of some methodological issues for doing quantitative corpus-based research. Arguably, the most important of these is also the one least often recognized: determining the ‘unit of analysis’ and the appropriate research design required for a particular research question. Thus, most of the article has focused on those considerations. By considering the design requirements of a research project before embarking on corpus construction, data collection, and/or linguistic coding and analysis, the researcher can ensure that the effort invested in a major research project will in fact result in the intended outcomes.

9. Literature

- Barnbrook, G. (1996), *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.
- Biber, D. (1990), Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. In: *Literary and Linguistic Computing* 5, 257–269.
- Biber, D. (1993), Co-occurrence Patterns among Collocations: A Tool for Corpus-based Lexical Knowledge Acquisition. In: *Computational Linguistics* 19, 549–556.
- Biber, D. (1993), Representativeness in Corpus Design. In: *Literary and Linguistic Computing* 8(4), 243–257.
- Biber, D./Conrad, S./Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Church, K. W./Hanks, P. (1990), Word Association Norms, Mutual Information, and Lexicography. In: *Computational Linguistics* 16, 22–29.
- Church, K. W./Gale, W. A./Hanks, P./Hindle, D. (1991), Using Statistics in Lexical Analysis. In: Zernick, U. (ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Erlbaum, 115–164.
- Collins, P. (1991), *Cleft and Pseudo-cleft Constructions in English*. London: Routledge.
- Nakamura, J./Sinclair, J. (1995), The World of *Woman* in the Bank of English: Internal Criteria for the Classification of Corpora. In: *Literary and Linguistic Computing* 10, 99–110.
- Oakes, M. P. (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Oh, S-Y. (2000), *Actually and in fact* in American English: A Data-based Analysis. In: *English Language and Linguistics* 4, 243–268.
- Prince, E. F. (1978), A Comparison of *wh*-clefts and *it*-clefts in Discourse. In: *Language* 54, 883–906.
- Schiffrin, D./Tannen, D./Hamilton, H. E. (eds.) (2001), *The Handbook of Discourse Analysis*. Oxford: Blackwell.
- Stubbs, M. (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.
- Thompson, S. A./Mulac, A. (1991), The Discourse Conditions for the Use of the Complementizer *that* in Conversational English. In: *Journal of Pragmatics* 15, 237–251.

Douglas Biber and James K. Jones, Flagstaff, AZ (USA)

Indexes

Index of names

A

Aarts, Bas 155, 165, 193, 424, 505, 739, 742, 747–748, 750–751, 995, 1011, 1057
Aarts, Jan 34, 46, 995, 1057
Abbs, Brian 116
Abdullah, Hasan 521
Abe, Naoki 948, 961–963
Abeillé, Anne 226, 230, 235
Abidi, Syed Sibde Raza 399
Abney, Steven 76, 230, 604, 803, 869, 961, 988
Abresch, Julia 1038
Adamic, Lada A. 343, 350, 361
Adams, Anthony 1082
Adamson, George W. 693
Adar, Eytan 367–368
Adda-Decker, Martine 1038–1039
Ädel, Annelie 112
Ades, Anthony 70
Adger, David 990
Adler, Manuela 1038
Aduriz, Itziar 231, 394
Afonso, Susana 231
Agirre, Eneko 322, 569
Agrawal, Rakesh 365
Agresti, Alan 792–793, 801, 1235, 1245
Agustini, Alexandre 895
Aha, David W. 858
Ahlemeyer, Birgit 281, 1150
Ahmad, Khurshid 399
Aijmer, Karin 48, 98, 117, 160, 196, 204, 277–278, 284–287, 433, 616, 1016–1017, 1053–1054, 1142–1143, 1145, 1163
Aiken, A. 1263–1264
Aitchison, Jean 687
Aizawa, Akiko 1249, 1252
Ajdukiewicz, Kazimierz 70
Ajiferuke, Isola 360
Alarcón, Enrique 707
Albert, Mark C. 858
Albert, Réka 330, 335, 341–344, 354–355, 361
Albright, Adam 901, 904, 911
Alderson, Peter 654, 1009
Algeo, John 1120
Alkula, Riitta 555
Allan, James 1281
Allan, Quentin Grant 271

Allen, James F. 614, 622–623, 637
Allen, Will 896
Allopenna, Paul 1207
Allwood, Jens 209, 211, 217, 220–221, 1026, 1034
Al-Onizan, Yaser 1192
Alshawi, Hiyan 1191
Al-Sulaiti, Latifa 517
Altenberg, Bengt 47–48, 117, 281, 284–285, 287, 433, 1053–1054, 1142–1143, 1145, 1163
Altmann, Gabriel 331, 333, 344, 365, 820
Amerl, V. 535
Andersen, Gisle 108, 194, 503, 617, 626, 1010, 1017, 1055, 1130
Andersen, Mette Skovgaard 146
Anderson, Anne H. 189
Anderson, James A. 506, 541
Andersson, Lars-Gunnar 1034–1035
Anderwald, Lieselotte 1128, 1130, 1135
Ando, Rie Kubota 540
Andor, József 987
Andrew, Galen 765
Androutsopoulos, Jannis K. 300
Ang, Chee Siang 293
Anshen, Frank 906, 909–911
Antosch, Friederike 1071
Antunović, Goranka 1028
Antworth, Evan L. 74, 560
Aone, Chinatsu 590
Apresjan, Jurij D. 963
Aquinas, Thomas 34, 707
Arabie, Phipps 891, 968
Archer, Dawn 567, 617–618, 626, 629–632, 634–635, 637
Argamon, Shlomo 846, 1072–1073, 1084
Argyle, Michael 211
Aristonicus of Alexandria 3
Armstrong, Susan 437, 541
Armstrong-Warwick, Susan 435
Arnold, Jennifer 989–990
Aronoff, Mark 554, 906, 909–911
Arthern, Peter J. 1177–1178
Asahara, Masayuki 538–539
Asher, Nicholas 973
Aske, Jon 1018
Asmah, Haji Omar 521
Asmussen, Helle 146
Asmussen, Jørg 135, 146–147, 1111

- Aston, Guy 98, 112, 115, 117, 119, 124, 155, 159, 165, 384, 738, 751, 1012, 1218
 Atkins, Beryl T. Sue 134, 139–140, 142, 144, 264, 458, 1012
 Atkinson, Dwight 403, 845, 1103
 Atwell, Eric 45–46, 90–91, 502–503, 505–506, 510, 512, 514, 517, 522
 Auran, Cyril 197, 409
 Austin, John Langshaw 614, 616–617
 Aycock, Joanna 906
- B**
- Baayen, R. Harald 331, 553, 787, 802, 804, 818–821, 882, 901–902, 904–914, 916, 1027, 1071–1072, 1122
 Babitch, Rose Mary 896
 Babko-Malaya, Olga 765
 Bacelar do Nascimento, Maria Fernanda 1017
 Bach, Kent 617
 Bächle, Michael 363–364, 366
 Baecker, Ronald M. 614
 Baeza-Yates, Ricardo 329, 707, 1256
 Bahl, Lalit R. 541
 Bai, Xiaojing 438
 Baker, Collin F. 147, 762, 962, 965, 967
 Baker, James K. 79
 Baker, Mona 277–278, 280, 286, 1141, 1154, 1162–1163, 1169–1170
 Baker, Paul 437, 515–517, 701
 Bakiri, Ghulum 863
 Balashek, S. 77
 Balasubrahmanyam, Vriddhachalam K. 816
 Baldi, Pierre 342, 361–362, 820
 Baldridge, Jason 618
 Baldry, Anthony 160
 Ballester, Almudena 1018
 Ballmer, Thomas T. 617
 Banerjee, Jayeeta 709
 Banerjee, Satanjeev 569
 Bangalore, Srinivas 89, 608
 Banjo, Ayo 1011
 Banko, Michele 870–871
 Barabási, Albert-László 330, 335, 341–344, 346, 354–355, 361
 Barbiers, Sjef 1134–1135
 Barbu, Catalina 583, 592
 Bar-Hillel, Yehoshua 16, 70, 565–566
 Bar-Ilan, Judit 356, 364
 Barker, Fiona 271
- Barlow, Michael 114–116, 120, 122, 283, 420, 686, 711, 1012, 1143, 1154, 1160, 1167–1168, 1178
 Barnbrook, Geoff 120, 1297
 Barnes, Helen 1198
 Barnett, Ros 510
 Baroni, Marco 154, 158, 318–319, 322, 328, 812, 820, 879, 882, 922, 982, 1088, 1111, 1142, 1162
 Barr, George K. 155, 160, 1080
 Barras, Claude 645, 658
 Barrett, Rusty 481
 Barry, Michael V. 1128
 Bartning, Inge 262
 Bartsch, Sabine 1213–1215, 1223–1224, 1245
 Barzilay, Regina 1261, 1281
 Bassecoulard, Elise 356, 359
 Bates, Elizabeth 1198, 1206
 Bauer, Laurie 42, 108, 185, 554, 901, 907, 911, 916–917, 1110, 1120
 Baum, Leonard E. 79
 Baumann, Tanja 493–494
 Baumgarten, Nicole 1161
 Bawcom, Linda 974
 Bayerl, Petra S. 766–767
 Bazzanella, Carla 286
 Beal, Joan C. 1127, 1130
 Bean, David 628
 Beare, Judith 411
 Beattie, Geoffrey 220
 Beaudouin, Valérie 846
 Beesley, Kenneth R. 74, 560
 Beier, Ernst G. 1008
 Beißwenger, Michael 293, 299, 301, 303–305
 Belew, Richard K. 877, 880, 884, 886, 896
 Belica, Cyril 1111
 Bell, Allan 1250–1251
 Bell, Timothy 820
 Bellugi, Ursula 217, 1199
 Belmore, Nancy 505
 Belz, Julie A. 262
 Benello, Julian 506, 541
 Bennett, Rick 310, 313
 Bennett, Scott W. 590
 Bennett, Tony 730
 Bennison, Peter 1168
 Bense, Max 333
 Berdan, Robert 895
 Berg, Celeste A. 364
 Bergenholz, Henning 33, 40, 132–133, 136, 146

- Bergh, Gunnar 309–313, 316, 322–323
Berglund, Ylva 109–110
Berjaoui, Nasser 295
Bernardini, Silvia 112, 118, 120, 122, 154, 158, 280, 318, 322, 324, 328, 1088, 1142–1143, 1162, 1168, 1170
Berners-Lee, Tim 492, 567
Bernot, Eduardo 707
Bernsen, Niels Ole 664
Berré, Michel 140, 147
Berry-Rogghe, Godelieve L. M. 1227
Bertagna, Francesca 146
Bertram, Raymond 914
Bertuglia, Cristoforo Sergio 889
Besnier, Niko 850
Bever, Thomas G. 1207
Bex, Tony 1095
Bharat, Krishna 1257
Bhatia, Vijay K. 1062
Biber, Douglas 13, 48, 116, 132, 134, 155–157, 161–162, 166, 169–172, 185, 188, 247–248, 317, 323, 331, 386, 403, 635–636, 778, 791, 803, 824, 826–827, 830, 832–834, 836, 843–851, 896, 907, 948, 978, 1002, 1012, 1015, 1052–1053, 1056, 1060–1061, 1082, 1092, 1094, 1102, 1154, 1163, 1254, 1289, 1292, 1295
Bick, Eckhard 231, 234
Biddulph, R. 77
Bierwisch, Manfred 973, 979
Bies, Ann 231, 765
Bilenko, Mikhail 1259
Billi, R. 1028
Bilmes, Jack 1055
Binnenpoorte, Diana 659
Binongo, J. N. G. 1072, 1085
Bird, Steven 592, 644, 674–675, 767, 772
Birdwhistell, Ray L. 209, 217
Bishop, Christopher M. 882, 888, 895
Björneborn, Lennart 339, 356, 358, 360–361
Black, Ezra 79, 237, 584, 608
Black, William 620
Blaheta, Don 1234
Blanche-Benveniste, Claire 1014–1015, 1053, 1111
Bledsoe, Woodrow W. 77
Blei, David M. 861, 871
Blockeel, Hendrik 356, 361
Bloedorn, Eric 1261, 1277
Bloom, Lois 1199
Bloomfield, Leonard 480
Blum, Avrim 869
Blum-Kulka, Shoshana 616, 634
Bobrow, Daniel G. 74
Bock, Hans Hermann 340
Bod, Rens 77, 79, 236–237, 602
Boersma, Paul 197
Boguraev, Branimir 571, 958, 973
Böhmová, Alena 230, 234, 443, 763
Bolinger, Dwight 200
Bollacker, Kurt 332
Bollobás, Bela 330, 336, 340–344, 346–347, 353
Bolozky, Shmuel 909
Bolton, Kingsley 1011
Bondi, Marina 120, 1053
Booij, Geert 554
Booth, Barbara 505
Booth, Taylor L. 600–601
Bordag, Stefan 351–352
Boreham, Jillian 693
Borg, Ingwer 894
Borin, Lars 762, 1143
Börner, Katy 369
Bornholdt, Stefan 348
Bosch, Peter 1244
Bosco, Cristina 231–232
Bosveld, Leonie 896
Botafogo, Rodrigo A. 361
Botley, Simon P. 112, 581, 583–584, 589
Bouayad-Agha, Nadjet 1252
Boucher, Victor J. 1016
Bouillon, Pierette 973
Boulton, Alex 123
Bouzon, Caroline 409
Boves, Lou (Louis) 195–196, 202, 659
Bowerman, Melissa 1206
Bowker, Lynne 117, 122, 280, 1168, 1179–1180
Boxer, Diana 1017
Božičević, Miran 369
Braine, Martin 1199
Brainerd, Barron 333
Brandt-Pook, Hans 614, 616
Brants, Sabine 231, 609, 763, 924
Brants, Thorsten 234–235, 541, 543, 545, 547, 605, 607, 766, 866
Braun, Sabine 112
Brazil, David 209
Breiman, Leo 535, 858
Breivik, Leiv Egil 48, 1010
Brekke, Magnar 321
Brennenstuhl, Waltraud 617
Brent, Michael R. 955–958, 965–966

- Bresnan, Joan 70, 599
 Breuer, Stefan 1038
 Brew, Chris 964, 968
 Brill, Eric 510, 541, 866, 870–871, 948, 1256
 Brin, Sergey 1253, 1257, 1263–1264
 Brinker, Klaus 331
 Brinton, Laurel J. 1004, 1055
 Briscoe, Edward J. (Ted) 238, 507, 571, 608,
 738, 954–958, 963, 973
 Brockett, Chris 1259
 Brockmann, Carsten 962
 Brodda, Benny 1049
 Broder, Andrei Z. 315, 362, 371, 1252,
 1256–1257, 1264–1265
 Brodine, Ruey 119–120
 Broeder, Daan 394, 497, 678
 Brom, Niek 843, 846–847
 Bronstein, Ilja N. 335
 Brosens, Veerle 193
 Brown, Gillian 636
 Brown, Keith 1136
 Brown, Penelope 619, 824
 Brown, Peter F. 279, 569, 574, 687, 689, 695,
 699, 701, 1183–1184, 1186, 1253, 1260
 Brown, Ralf D. 1182
 Brown, Roger 1199, 1205, 1209
 Brown, Steven D. 876, 891
 Browning, Iben 77
 Bruce, Rebecca 575
 Bruckman, Amy 298, 301
 Brugman, Hennie 394, 674–675
 Bruneseaux, Florence 583–584, 587
 Brunet, Etienne 1073–1074
 Brunswick, Egon 216
 Brusilowsky, Peter 146
 Buchholz, Sabine 870
 Bucholtz, Mary 1013
 Buckley, Chris 886
 Budzikowska, Małgorzata 1272, 1278
 Buhmann, Jeska 655
 Bunt, Harry 620, 623
 Burchardt, Aljoscha 763
 Burdine, Stephanie 116, 122
 Burger, Harald 1245
 Burges, Jená 845
 Burnard, Lou 47, 98, 112, 155, 159, 165,
 197, 282, 384–385, 487, 726, 738, 751, 1218
 Burrows, John F. 907, 1072, 1078, 1081,
 1084
 Burston, Jack 121
 Busa, Roberto 34, 208, 707
 Bußmann, Hadumod 929
 Butler, Christopher 801
 Buxton, William A. S. 614
 Bybee, Joan L. 189, 901, 904, 911, 916, 1096
- C**
- Cahill, Aoife 232
 Califf, Mary Elaine 868, 869
 Cameron, Lynne 1012, 1016
 Candlin, Christopher 120
 Canisius, Sander 863
 Canter, David 1079
 Cantos-Gómez, Pascual 271
 Capel, Annette 116
 Capocci, Andrea 370–373
 Carbonell, Jaime 687
 Cardey, Sylviane 282
 Cardinal Hugo 2
 Carl, Michael 1181
 Carletta, Jean 594, 614, 618, 624–625, 631,
 637, 766, 923
 Carlson, Lynn 232, 618, 621, 762
 Carlson, Rolf 1028
 Carpenter, Bob 70
 Carpenter, Edwin 116
 Carpuat, Marine 576
 Carr, Les 357
 Carreras, Xavier 863, 871
 Carroll, Glenn 955–959
 Carroll, John A. 237–238, 507, 560, 608,
 954–958, 963
 Carroll, John B. 902
 Carstairs-McCarthy, Andrew 555
 Carter, David 234, 765
 Carter, Ronald 116, 165, 411, 973, 975, 994,
 1011–1012, 1015–1017
 Carterette, Edward C. 1008
 Caruana, Rich 865
 Casacuberta, Francisco 319, 1188
 Cassell, Justine 213
 Castellan, N. John 1276
 Catizone, Roberta 689, 1260
 Cauldwell, Richard 1018
 Cavaglià, Gabriela 572
 Cavar, Damir 391
 Cecchi, Guillermo A. 344, 354–355
 Čermák, František 136, 1018
 Čermáková, Anna 979
 Cerrato, Loredana 220
 Chafe, Wallace 169, 1009, 1015, 1056
 Chakrabarti, Soumen 361–362, 1257

- Chalkley, Mary Anne 1206
Chambers, Angela 119–120, 123
Chambers, Jack K. 896, 1122, 1126, 1129
Champollion, Jean-François 275
Chang, Baobao 438
Chang, Jason S. 695, 698, 702
Chang-Rodriquez, Eugenio 148
Chapelle, Carol A. 262
Chapman, Robert L. 573
Chapman, Robin S. 1200
Charniak, Eugene 79–80, 230, 237, 600, 602, 1191
Chater, Nick 1206–1207
Chaucer, Geoffrey 2, 107, 242
Chen, Chaomei 360
Chen, Francine 1275
Chen, Hsin-hsi 693, 701
Chen, Keh-Jiann 232, 443, 536–537
Chen, Kuang-hua 693, 701
Chen, Mathis H. M. 695
Chen, Stanley F. 537, 543, 1260, 1262
Cheng, Peter C.-H. 749
Cheng, Winnie 417, 1011, 1014
Chesley, Paula 959
Chinchor, Nancy 583, 585–586
Choi, Hyung Eun 271
Chomsky, Noam 15–30, 70–71, 77–78, 230, 708, 777, 925, 988–991
Choueka, Yaakov 283, 1213, 1221, 1224
Choukri, Khalid 394, 643
Chowdhury, Abdur 1252, 1263
Chowdhury, Nandini 266, 430
Christ, Oliver 1167
Christiansen, Morten H. 1207
Chun, Lan 282
Church, Kenneth W. 49, 283, 541, 564, 570, 574, 687–691, 694–695, 697–699, 701–702, 797, 886, 1177, 1226–1227, 1259–1260, 1263, 1265, 1297–1298
Ciaramita, Massimiliano 318, 572
Cicekli, Ilyas 1182
Citron, Sabine 137
Clark, Eve V. 1197, 1209
Clark, Herbert H. 908, 917
Clark, Peter 858
Clark, Steven 961–962
Clarke, Charles L. A. 319, 1220
Clear, Jeremy 134–136, 264, 458, 567, 1012
Clément, Lionel 145
Clift, Rebecca 1016
Clough, Paul 567, 1251–1253, 1262, 1265
Čmejrek, Martin 227, 1191
Coates, Jennifer 48
Cobb, Thomas (Tom) 120, 269, 271
Coffey, Stephen 119
Coggins, James M. 1259
Cohen, Ira 680
Cohen, Jacob 766
Cohen, William W. 858
Collberg, Christian 1250
Collier, René 189
Collins, Bróna 1182
Collins, Jeff 1076
Collins, Michael 235, 237, 602, 604
Collins, Peter 48, 1292
Comrie, Bernard 999, 1003–1004, 1131
Condamines, Anne 137
Conley, Ehud S. 283
Connine, Cynthia 568
Connor, Ulla 112, 845, 847
Connor-Linton, Jeffrey 845, 847
Conrad, Susan 115–116, 132, 134, 791, 803, 827, 845, 948, 1012, 1015, 1092, 1289
Conti-Ramsden, Gina 1200
Cook, Guy 1013
Cooper, Matthew 1261
Copestake, Ann 74, 973
Corbin, Danielle 901–902
Core, Mark G. 614, 623, 637
Cornips, Leonie 1134–1135
Corns, Thomas N. 1082
Corrigan, Karen 1130, 1134
Cortes, Corinna 858
Cotterill, Janet 1016
Coulthard, Malcolm 618, 624, 631, 1263
Couper-Kuhlen, Elizabeth 202
Cover, Thomas M. 858
Cowan, Ron 271
Cowie, James R. 569
Coxhead, Averil 117, 419–420
Cranias, Lambros 694, 1179
Creer, Sarah 1018
Cremelie, Nick 1039
Cresswell, Andy 120
Cresti, Emanuela 1018
Creutz, Mathias 561, 820
Croft, Anthony 890
Croft, W. Bruce 566
Cronin, Beau 147
Crowdy, Steve 11, 412, 1009, 1012
Cruden, Alexander 1–2, 13
Cruse, David Alan 974, 977
Cruttenden, Alain 200

- Crystal, David 200–201, 293, 301, 303, 920, 1012
 Csomay, Eniko 636, 845, 847
 Cucchiarini, Catia 651, 659
 Culicover, Peter 778
 Culpeper, Jonathan 244, 246–247, 406, 619, 629–630, 634–635, 637, 733, 1055, 1061
 Culwin, Fintan 1261, 1263
 Cuřín, Jan 1191
 Curme, George O. 4, 23, 28–29
 Curzan, Anne 5, 1092, 1096, 1101
 Cutting, Douglas R. 541, 545, 846, 1253
 Cyrus, Lea 227
 Czerwinski, Mary 360
 Czerwon, Hans-Jürgen 357, 360
- D**
- da Silva, Joaquim Ferreira 1244
 Dabrowska, Ewa 901, 911
 Daelemans, Walter 571, 603, 861, 865–866
 Dagan, Ido 283, 581
 Dagneaux, Estelle 262, 266
 Daille, Béatrice 696–697, 1224, 1226, 1238, 1245
 Dale, Philip S. 1198
 Daley, Robert 49
 Dalgaard, Peter 801, 815
 Dalli, Angelo 394
 Dalrymple, Mary 74
 Dalton-Puffer, Christiane 908
 Damashek, Marc 1256–1257
 Damerau, Fred J. 693
 Danet, Brenda 303
 Dang, Hoa 766
 Danielsson, Pernilla 277, 1168
 Darbelnet, Jean 1160
 Dasarathy, Belur V. 858
 Dasher, Richard B. 60
 Daumé III, Hal 871
 Dave, Kushal 293
 Davidson, Lisa 1037
 Davies, Mark 119
 Davies, Sarah 584, 587, 589, 594
 Davis, James 1253, 1257, 1263–1264
 Davis, K. H. 77
 Davis, Randall 1253–1254
 Davison, Mark L. 894
 Davison, Robert 890
 Day, David 590, 592
 Day, Michael 492
 de Beaugrande, Robert 116, 330, 988–989, 991, 1057, 1059–1060
 De Cock, Sylvie 269–270, 1012
 de Houwer, Annick 1200
 De Kuthy, Kordula 921
 de la Clergerie, Éric 145, 772
 de Lima, Erika 959
 de Mönnink, Inge 843, 846–847
 de Moya-Anegón, Félix 360
 de Ruiter, Jan Peter 670
 de Santis, Cristiana 390
 de Saussure, Ferdinand 974, 988
 De Schryver, Gilles-Maurice 135, 141, 144, 315, 472
 de Silva, Vin 895
 de Soete, Geert 891
 de Solla Price, Derek John 356
 Debrock, Mark 262
 Declerck, Thierry 295, 491
 DeCristofaro, Jonathan 590
 DeForest, Mary 107, 1072, 1076
 Degand, Liesbeth 262
 DeGroot, Morris H. 780, 783–784, 786–787, 789, 792, 801, 1235, 1245
 Deignan, Alice 1012, 1016
 Demers, Alan J. 601
 Demeyer, Serge 1261, 1263
 Dempster, Arthur P. 79, 869, 1185
 Demuynck, Kris 650, 652
 den Os, Els 643
 Denison, David 407, 1121
 Denness, Sharon 266
 Dennis, Sally F. 1227
 DeRose, Steven J. 541, 769
 Derouault, Anne-Marie 541
 Di Eugenio, Barbara 766
 Diamantaras, Kostas 888, 895
 Dias, Gaël 1220, 1235
 Dice, Lee R. 693–694
 Dickerson, Robert (Bob) 1256–1257, 1263–1264
 Dickinson, Markus 235, 765, 923
 Diessel, Holger 1208–1209
 Diestel, Reinhard 335
 Dietrich, Rainer 1055
 Dietterich, Thomas 857, 863
 Dill, Ken A. 693
 Dillon, Andrew 362
 Dinesh, Nikhil 762
 Dipper, Stefanie 609, 763, 765, 771
 Dix, Alan J. 614
 Dixon, Peter 1072

Docherty, Gerry 1128, 1130
 Docherty, Vincent T. 145
 Dodd, Bill 119
 Doddington, George 1256
 Doherty, Monika 1170
 Dolan, William B. 1259
 Domashnev, Constantine 1182
 Don, Zuraidah Mohd 511, 520–523
 Donegan, Patricia J. 1028
 Doreian, Patrick 356, 359
 Dorgeloh, Heidrun 1053, 1062
 Döring, Nicola 301
 Dorogovtsev, Serguei N. 351
 Dorow, Beate 351
 Dorr, Bonnie J. 965
 Douglas, Shona 1191
 Douglas-Cowie, Ellen 195
 Doval Suárez, Susana María 1143
 Dowty, David 954, 973–974
 Drakos, Nicos 522
 Dressler, Wolfgang 900–901, 911, 916
 Drucker, Steven 304
 Du Bois, John W. 194, 413, 1009, 1055
 Dubes, Richard C. 891
 Ducasse, Stéphane 1261, 1263
 Duda, Richard O. 891, 894–895
 Dudley-Evans, Tony 118
 Duit, Charlotta 1039
 Dumais, Susan T. 360
 Dunn, Graham 876, 888, 891, 894
 Dunning, Ted E. 697, 1227, 1235
 Duranti, Alessandro 1013
 Dutilh, Tilly 440
 Dykins, Jane 1200
 Dyvik, Helge 284

E

Earley, Jay 72
 Ebeling, Jarle 278, 282–283, 433, 1143,
 1163, 1167
 Ebersbach, Anja 369
 Echermane, Ana 364
 Eckle-Kohler, Judith 959
 Edmonds, Philip 564, 575
 Edmont, Edmond 1127
 Edmundson, H. P. 1274
 Edwards, Jane A. 1009, 1013
 Eeg-Olofsson, Mats 46, 503
 Egghe, Leo 339
 Eggins, Suzanne 1009

Einarsson, Jan 1034–1035
 Eiron, Nadav 362
 Eklund, Robert 1025, 1029–1031, 1038–
 1039
 Ekman, Paul 220
 Elhadad, Michael 1281
 Elkan, Charles 861, 871
 Elkiss, Aaron 923
 Ellegrård, Alvar 46, 760, 910
 Elliot, Dale 1002
 Elliott, John 505
 Ellis, Nick C. 441
 Ellis, Rod 264
 Elsig, Martin 1134
 Elworthy, David 542, 545
 Emerson, Thomas 701
 Emmons, Kimberley 1092
 Endres-Niggemeyer, Brigitte 1276
 Engel, Dulcie M. 278
 Enkvist, Nils Erik 1046, 1052, 1061
 Erdmann, Oskar 929
 Erdős, Paul 342
 Erjavec, Tomaz 287, 434, 436, 1165
 Erk, Katrin 962
 Erman, Britt 48
 Ernestus, Mirjam 911
 Ervin-Tripp, Susan 636, 824
 Eskin, Eleazar 1253, 1261
 Eskola, Sari 1169
 Esteve Ferrer, Eva 966–968
 Estling, Maria 236
 Evans, Richard 585
 Everitt, Brian 876, 888, 891, 894, 895
 Evert, Stefan 135, 143, 298, 385, 766, 799,
 820, 907–910, 922, 935–936, 977, 1111,
 1215, 1218, 1220, 1223–1224, 1226–1227,
 1232, 1234–1236, 1238, 1240, 1242–1245
 Eye, Alexander von 949

F

Faba-Pérez, Cristina 360
 Fabricius-Hansen, Catherine 1150, 1160,
 1170
 Facchinetti, Roberta 1056, 1100
 Fager, Edward William 697
 Fairclough, Norman 329–330, 1062
 Falk, Johan 1111
 Faloutsos, Christos 347–348
 Fang, Alex 507, 958, 1009
 Fang, Yong 357

- Fankhauser, Peter 677, 1167
 Farr, Fiona 415, 1011–1012, 1016
 Farrar, Scott 770
 Faulstich, Lukas C. 252
 Faust, Katherine 334, 339–340
 Fazly, Afsaneh 1244
 Feldweg, Helmut 295
 Fellbaum, Christiane 85–86, 565, 571–572,
 766, 960, 967
 Feng, Jiang 539
 Ferguson, Charles A. 823
 Fernandez, M. M. Jocelyne 1055
 Fernández Díaz, Gabriela 1178, 1180
 Ferrer i Cancho, Ramon 330, 347, 351–353
 Fidell, Linda S. 876, 888, 890–891
 Filipović, Rudolf 1028
 Fillmore, Charles J. 33, 139–140, 142, 147,
 762, 962, 965, 967, 987, 996, 1005, 1050
 Finch, Steven 1206–1207
 Finegan, Edward 157, 248, 403, 845, 1094,
 1102
 Fink, Barbara 1154
 Firth, David R. 362
 Firth, John Rupert 11, 55, 83, 116, 716, 938,
 940, 996, 1057, 1212–1213, 1217, 1219,
 1242
 Fischer, Kerstin 286, 614, 616
 Fischer, Klaus 954
 Fischer, Roswitha 1122
 Fisher, Danyel 364
 Fisher, John Hurt 1104
 Fisher, Michael J. 1259
 Fitschen, Arne 298, 560
 Fitt, Susan 1026, 1038
 Fitzmaurice, Susan 1100
 Fix, Ulla 330
 Flake, Gary W. 362
 Flament-Boistrancourt, Danièle 262
 Fletcher, Paul 801, 888, 891, 895, 1078
 Fletcher, William 310–316, 318, 324
 Fligelstone, Steven 35, 40, 113, 119, 123,
 583, 590
 Florio, John 7
 Fløttum, Kjersti 1056
 Flowerdew, John 117
 Flowerdew, Lynne 117
 Flynn, Patrick J. 891, 894
 Fodor, Jerry A. 27
 Folch, Helka 846
 Foltz, Peter 896
 Fonseca-Greber, Bonnie 1014
 Foote, Jonathan 1261
 Ford, Cecilia E. 202
 Forsberg, Karin 1036
 Forst, Martin 609, 763
 Forsyth, Richard S. 1072, 1083–1084, 1086,
 1088
 Foster, George F. 283, 693, 1178, 1192, 1260
 Foth, Kilian 601
 Fourcin, Adrian 643
 Fowler, Roger 169
 Fox, Gwyneth 116
 Frakes, William B. 883
 France, Ilene M. 565
 Francis, Gill 114, 116, 120, 937–938, 940,
 990, 994
 Francis, Ivor S. 1077
 Francis, W. Nelson 1, 7–8, 10, 35–38, 40,
 44, 48, 90, 227, 395, 575, 760–761, 1012
 Frankenberg-Gracia, Ana 1143
 Frasconi, Paolo 342, 361–362, 820
 Fraser, Colin 824
 Fraser, J. Bruce 28, 74
 Fraser, Michael 2
 Fredriksson, Anna-Lena 282
 Freebairn, Ingrid 116
 Freeman, Andrew 510
 Fries, Charles Carpenter 4–6
 Fries, Udo 405, 1251
 Friesen, Wallace V. 220
 Fu, Guohong 539
 Fukumoto, Fumiyo 1281
 Fung, Pascale 688, 697–698, 702, 1262
 Furnival, Frederick James 9
 Fürnkranz, Johannes 858
 Furuse, Osamu 1179

G

- Gadsby, Adam 429
 Gaeta, Davide 904–905
 Gaizauskas, Robert J. 585, 1252, 1265
 Gale, William A. 283, 564, 570, 574, 687,
 689–691, 694–695, 699, 701–702, 886,
 1177, 1260
 Gamallo, Pablo 895
 Gamper, Johann 146
 Gao, Yanjie 121
 Garabik, Radovan 392
 Garcia, Angela Cora 301, 304
 Garcia-Molina, Hector 1252–1253, 1255,
 1257, 1263–1264
 García-Varea, Ismael 1188

- Garfield, Eugene 356–357, 360
Garlaschelli, Diego 372
Garofolo, John S. 652
Garside, Roger 45–46, 90, 229, 386, 422, 434, 503, 505, 510–511, 517, 541, 582–583, 590, 760, 1144
Gaskell, Delian 120, 271
Gass, Susan M. 264
Gast, Volker 990, 993–994
Gaussier, Éric 283, 686, 697–698, 1179
Gautier, Gérard 394
Gavioli, Laura 112, 117, 119–122, 124
Ge, Niyu 582, 589
Gee, James Paul 604
Geisler, Christer 845
Gellerstam, Martin 161–163, 278, 281
Gentner, Dedre 565
Georgala, Effi 959
Germann, Ulrich 287, 1163, 1188, 1192
Geyer, Klaus 1131
Geyken, Alexander 135, 140, 391
Ghadessy, Mohsen 112, 121
Ghani, Rayid 318, 462
Gianitsová, Lucia 392
Gibbon, Dafydd 196, 628, 643
Gibson, David 362, 368
Giddens, Anthony 102
Gil, Alexandre 1220
Gildea, Daniel 700, 761, 768, 962, 965, 1191
Giles, C. Lee 312, 314, 332, 362
Gill, Kathy E. 366
Gill, Martin 1045
Gillard, Patrick 429
Gilliéron, Jules 1127
Gillis, Steven 650, 652
Gilliver, Peter 8–9
Gilmore, Alex 1017
Gilquin, Gaëtanelle 1141
Giménez, Jesus 541
Gippert, Jost 772
Gitchell, David 1259
Givón, Támy 82, 1052
Glance, Natalie 366–367
Glänzel, Wolfgang 357, 360
Glaser, Markus 369
Glass, Michael 766
Gleim, Rüdiger 296, 328–329, 363
Glenisson, Patrick 357
Glover, Angela 1263
Glover, Warren W. 394
Goddijn, Simo 659
Godfrey, John J. 643
Goh, Chooi-Ling 539
Gold, Bernhard 666
Goldberg, Adele E. 941
Goldman, Jean-Philippe 1215, 1224
Goldman-Eisler, Frieda 220
Goldschen, Alan 590, 592
Goldstein, Jade 1277
Gollhardt, Kurt D. 1028
Gómez Guinovart, Xavier 435
Gonzalo, Julio 328
Good, Irving J. 1026, 1037
Goodale, Malcolm 116
Goodluck, Helen 38, 41, 47, 395
Goodman, Joshua T. 537, 543
Goodwin, Charles 211, 622
Gordon, A. D. 876, 891, 894–895
Gordon, Lesley 1082
Gordon, Matthew 98
Gordon, Raymond G. Jr. 461
Gore, Paul A. 891, 894
Gorin, Allen 1028
Görlach, Manfred 1027
Goronzy, Silke 1039
Gorrell, Genevieve 957
Gosset, William Sealey 783
Götze, Michael 763, 765
Gough, Nano 1191
Goutte, Cyril 698
Gouverneur, Céline 116
Gouws, Liezl 135
Gouws, Rufus H. 132, 135, 139c140
Grabe, Esther 193, 204, 412
Grabe, William 847, 849
Grabowski, Eva 115
Graddol, David 1027–1029
Granger, Sylviane 48, 112, 117, 121, 160, 262, 265–267, 270–271, 281, 427–428, 1012, 1053, 1057, 1063, 1141, 1145
Grannes, Dean J. 1182
Granström, Björn 1028
Gratch, Jonathan 213
Gray, Bethany E. 120
Gray, Charles 9
Greasely, Peter 195
Greaves, Christopher 417
Greenbaum, Sidney 11–12, 39, 42, 408, 503, 1011
Greenberg, Steven 652
Greene, Barbara B. 43–44, 90, 503, 541
Greenfield, Peter 282
Grefenstette, Gregory 138, 310, 312–313, 316–317, 324, 328, 507, 533–534, 536, 540, 922, 1252

- Grégoire, Nicole 1245
 Grewendorf, Günther 925
 Grice, Herbert Paul 614, 617, 626–627
 Gries, Stefan Th. 787, 802, 895, 935, 941–943, 948–949, 992
 Grimes, Carrie 895
 Grimm, Jacob (Jakob) 3, 255
 Grimm, Laurence G. 876, 888, 891, 895
 Grimm, Wilhelm 255
 Grimshaw, Jane B. 954
 Grinter, Rebecca E. 293
 Grishman, Ralph 699, 953, 958
 Groenen, Patrick 894
 Gronostaj, Maria Toporowska 1034
 Groom, Nicholas 165
 Grosjean, François 604
 Grossberg, Lawrence 730
 Grossé, Siegfried 1134
 Grossmann, Francis 1213–1214, 1245
 Gruhl, Daniel 366–368
 Grundy, Valerie 137, 140
 Gu, Yueguo 445
 Guerrero-Bote, Vicente P. 360
 Guilloré, Sylvie 1235
 Gulikers, Leon 553, 904
 Gulli, Antonio 312
 Gushrowski, Barbara A. 362
 Gustafsson, Leift 666
 Gut, Ulrike 1017
 Guthrie, Joe A. 569
 Guthrie, Louise 569
 Guttropf, Anja 910
 Güvenir, H. Altay 1182
- H**
- Ha, Le Quan 815–816
 Haarman, Louann 1053
 Häckl Buhofer, Annelie 169, 1245
 Hadenius, Patric 312, 315
 Hadley, Gregory 119
 Hahn, E. 26
 Hahn, Udo 620
 Haider, Hubert 925, 928
 Hair, Joseph F. 876, 878, 888, 890–891, 894–895
 Hajíč, Jan 230, 443, 555
 Hajíčová, Eva 232, 763
 Hakkani-Tür, Dilek Z. 561
 Hakulinen, Auli 469, 1055
 Halavais, Alexander 368
 Hall, Joan H. 1128
 Hall, Johan 763
 Halle, Morris 21, 71
 Halliday, Michael A. K. 26, 55, 221, 331, 351, 353, 580, 584, 589, 1015, 1163
 Halliday, Wilfred J. 1128
 Hamaker, Jonathan 655
 Hamilton, Heidi E. 1287
 Hammarberg, Björn 262
 Hampe, Beate 949
 Han, Chung-hye 231
 Handford, Michael 1011
 Hanks, Patrick 139, 503, 1226, 1297
 Hannabuss, Stuart 1250
 Hannah, Jean 1112
 Hansen, Silvia 677, 1154, 1161–1164, 1167–1170
 Hansen-Schirra, Silvia 232, 1160–1161, 1164, 1167, 1169, 1170
 Harabagiu, Sandra M. 583
 Hardie, Andrew 510–511, 515–518, 520
 Hared, Mohamed 850
 Hargreaves, Martin 890
 Harley, Andrew 143
 Harnad, Stevan 357
 Harnish, Robert M. 617
 Harnly, Aaron 1279
 Harrington, Jonathan 644
 Harris, Brian 1177
 Harris, Kathryn 263, 1018
 Harris, Zellig S. 16, 17, 23, 88
 Hart, Peter E. 858, 891, 894–895
 Haruno, Masahiko 699
 Harry, Benjamin 194
 Hasan, Ruqaiya 353, 580, 584, 589
 Hasler, Laura 582, 591, 593, 1273–1274, 1276, 1282, 1284
 Haslerud, Vibeke 410, 1016
 Hasselgård, Hilde 284, 1143, 1149, 1170
 Hasselgren, Angela 262, 1010
 Hasund, Kristine 97, 108, 1010, 1017, 1130
 Hatzigeorgiou, Nick 391
 Hatzivassiloglou, Vasileios 699, 968, 1234–1235, 1253, 1261
 Hauge, Jostein 38
 Hausmann, Franz Josef 7, 132, 134, 1213
 Hausser, Roland 557
 Havelka, Jiri 1191
 Hawkins, Donald T. 315
 Hawkins, John A. 1148
 Hay, Jennifer B. 77, 901, 906, 911–916, 982
 Hayes, Bruce 901, 904, 911

- Hayes, Patrick J. 745
Hays, William 801
Hearne, Mary 1191
Hearst, Marti A. 535, 540, 636, 687
Heckel, Paul 1259
Hedberg, Nancy 204
Heenan, Charles H. 2
Heeringa, Wilbert 896
Heid, Ulrich 139–140, 145, 491, 560, 608, 959, 1217, 1244
Heigl, Richard 369
Heikkinen, Kanerva 1100
Heikkonen, Kirsi 1104
Heinemann, Wolfgang 330, 332, 357
Heintze, Nevin 1256, 1263–1264
Heintze, Silvan 620
Heinz, Michaela 142
Helbig, Gerhard 954
Helfman, Jonathan 1259, 1263, 1265
Helt, Marie E. 845
Helwigh, Harwig Richard 9
Henderson, John 590, 592
Hendler, James 567
Hendriks, Henriette 997
Henry, Alex 112, 117
Hepburn, Alexa 1013–1014
Hermerén, Lars 48
Hernández, Nuria 1131
Hernandez-Borges, Angel A. 365
Herring, Susan C. 292–293, 295, 297, 301, 303, 367
Herrmann, Tanja 1131
Herzog, Marvin I. 56, 1092–1093
Hewings, Martin 271
Heyer, Gerhard 351–352
Heylen, Dirk 213
Heyn, Matthias 1164
Heyvaert, Liesbet 277
Hickey, Leo 1056
Hickman, Maya 1209
Hidalgo, Encarnación 112
Hilbert, Mirco 769
Hill, Jimmie 116
Hillmann, Diane 494
Hilton, Michael L. 1071, 1079–1080
Himmelmann, Nikolaus P. 772
Hindle, Donald 235, 948
Hinrichs, Erhard W. 226–227, 230, 608, 763, 766, 768
Hinton, Geoffrey E. 858
Hirschman, Lynette 583, 585–586, 595
Hirst, Daniel 409
Hirst, Graeme 568, 571, 973, 1263
Hislop, Gregory W. 1262
Hjortsjö, Carl-Hermann 220
Hoad, Timothy C. 1264
Hobbs, Jerry R. 620
Hockenmaier, Julia 232, 693, 762
Hockett, Charles F. 21
Hockey, Susan 329
Hodges, Mary 104, 108, 726, 729, 1072, 1118–1120
Hoey, Michael 114, 330, 352, 712, 938–940, 977, 979, 982, 996, 1145
Hoff-Ginsberg, Erika 1198
Hoffmann, Sebastian 143, 180, 255–256, 385, 1101
Hofland, Knut 37, 41, 48, 172, 177, 262, 282–283, 694, 1142–1144, 1163–1164, 1168
Hofmann, Walter 1
Höge, Harald 643
Holanda, Adriano de Jesus 354
Hollan, James 333, 363
Holler, Anke 581
Holliman, Edward C. 643
Hollmann, Willem B. 1132, 1136–1137
Holloway, Todd 369
Holme, Petter 355
Holmes, David I. 1071, 1074–1076, 1079–1078, 1082–1084, 1088, 1253
Holmes, James S. 280
Holmes, Janet 108, 414, 1011, 1013, 1017, 1055
Holmes-Higgin, Paul 399, 1253
Holthuis, Susanne 330
Honoré, Antony 1073–1074
Hoorn, Johan F. 1082
Hoover, David 896, 1071, 1075
Hopcroft, John E. 558
Hopey, Philip D. 89
Hopper, Paul 58, 1001, 1096
Horn, Ernest 10
Horn, Laurence R. 620
Hornby, Albert S. 116, 958
Horvath, Barbara 896
Horváth, József 431
Houghton, George 624
House, Jill 188, 190
House, Juliane 616, 634, 1161, 1170
Housen, Alex 269
Hovy, Eduard 1271, 1280
Howes, D. H. 1009
Hu, Xiaoling 1088

- Huang, Chu-Ren 443–444
 Huang, Ping-Yu 269
 Huang, Renje 505
 Huber, Magnus 1129
 Huberman, Bernardo A. 361
 Hubert, Lawrence 891, 968
 Huddleston, Rodney 29, 1146
 Hudson-Ettle, Diana 1011
 Hughes, Arthur 801, 888, 891, 895, 1078
 Hughes, John 512, 522
 Hughes, Rebecca 116, 1017
 Huh, Myung-hoe 169
 Hull, David A. 883
 Hummon, Norman P. 356, 359
 Humphreys, Kevin 585
 Hundt, Marianne 182–184, 256, 395, 1116, 1121
 Hung, Josef 112, 267, 1012
 Hunston, Susan 112, 114, 116, 120, 269, 937–938, 940, 990, 994, 1012, 1056, 1063
 Hurst, Matthew 366–367
 Hutchins, Edwin 333, 363
 Hutchins, John W. 87, 566
 Hymes, Dell H. 114, 619, 629, 823
 Hyönä, Jukka 914
- I**
- Ide, Nancy 235, 386, 486, 564–566, 674, 680, 769, 772, 1010, 1144
 Ife, Anne 262
 Ihälainen, Ossi 43, 1016, 1126, 1131
 Iida, Hitoshi 1181
 Ikehara, Satoru 699
 Imbs, Paul 135
 Ingwersen, Peter 356, 360–361
 Inui, Kentaro 601
 Isabelle, Pierre 283, 693, 1176, 1177, 1260
 Isahara, Hitoshi 430, 702, 1259
 Ishimoto, Hiroyuki 699
 Itai, Alon 581
 Itkonen, Esa 22
 Itô, Junko 1028
 Itzkovitz, Shalev 346
 Iwasaki, Hideya 698
 Izre'el, Shlomo 194
 Izumi, Emi 430
- J**
- Jackendoff, Ray S. 778, 900
 Jackson, Howard 567
 Jackson, MacDonald P. 1071
 Jacobs, Andreas 56, 1097
 Jacobs, Jennifer Baker 301, 304
 Jain, Anil K. 891, 894
 Jakobs, Eva-Maria 330
 Jakobson, Roman 21, 707
 Jang, Jyh-Shing R. 698, 702
 Jang, Shyue-Chian 850
 Janich, Nina 295
 Jannedy, Stefanie 77
 Järborg, Jerker 233
 Järvinen, Timo 234, 605
 Jassem, Wiktor 895
 Jefferson, Gail 202, 218, 618, 623, 1014
 Jekosch, Ute 1038–1039
 Jelinek, Frederick 601, 903, 1176
 Jenkins, Jennifer 1112
 Jeong, Hawoong 330, 341, 343, 354, 361
 Jespersen, Otto 4–5, 10, 13, 23, 28, 1027, 1113
 Jing, Hongyan 1261, 1272, 1274, 1276–1278
 Joanis, Eric 966–968
 Johansson, Mats 281, 283
 Johansson, Stig 33, 38–39, 41, 46–48, 112, 122–123, 172, 177, 277–279, 281–284, 287, 395, 433, 503, 990, 1142–1144, 1150–1151, 1161, 1163, 1168, 1170
 Johns, Tim 118–121
 Johnson, C. Douglas 71
 Johnson, Christopher R. 140
 Johnson, Eric 107, 1072, 1076
 Johnson, Gary 414, 1013
 Johnson, Mark 73, 572, 870, 1234
 Johnson, Samuel 6–9, 255
 Jolliffe, Ian T. 888
 Jones, Danny 1181
 Jones, Douglas 965
 Jones, Glyn 119, 124
 Jones, James K. 847
 Jones, Margaret H. 1008
 Jones, Mark J. 1128
 Jones, Megan 1133–1134
 Jones, Paul A. 1283
 Jones, Randall L. 115, 119
 Jones, Rodney 304
 Jones, Rosie 318, 462
 Jones, Steven 974–975
 Jones, Susan 49
 Jones, Valerie 1130
 Jones-Sargent, Valerie 896
 Jordan, Michael I. 861, 871
 Jorgensen, Julia 565

Joseph, Kate 1200
Josey, Meredith Pugh 1113
Joshi, Aravind K. 70–71, 74, 89, 693
Jucker, Andreas H. 56, 1097, 1102
Juillard, Alphonse 34, 148
Jungbluth, Konstanze 1025
Junqua, Jean-Claude 1039
Jurafsky, Daniel S. 88, 560, 619–620, 622, 949, 955, 1187, 1221, 1238, 1256
Juszczyk, Peter W. 1207
Justeson, John S. 935

K

Kachigan, Sam Kash 876, 891, 894–895
Kachru, Braj 260
Käding, Friedrich Wilhelm 147
Kageura, Kyo 820
Kahlas-Tarkka, Leena 1099
Kahrel, Peter 510
Kallen, Jeffrey L. 1011, 1016
Kallmeyer, Laura 747, 923
Kaltenböck, Gunther 1053
Kalton, Graham 7
Kameyama, Megumi 621
Kan, Min-Yen 965
Kang, Beom-mo 169, 393
Kang, Jian 1000
Kangas, Jari 895
Kaplan, Abraham 77
Kaplan, Ronald M. 70, 71
Kärkkäinen, Elise 1055
Karlgren, Jussi 846, 1253
Karsson, Fred 20, 46, 282, 469, 541, 601, 605, 739
Karp, Daniel 560
Karrass, Jan 1198
Karttunen, Lauri 74, 560
Kasami, Tadao 72
Kaski, Sami 895
Kasper, Gabriele 616, 634
Kasper, Robert T. 70
Kassarjian, Harold H. 567
Kastovsky, Dieter 908, 910, 914
Katz, Boris 1253–1254
Katz, Jerrold J. 26–27
Katz, Slava M. 935
Kauchak, David 861, 871
Kaufman, Leonard 967
Kaufman, Terrence 1028
Kaufmann, Uwe 136

Kautz, Henry 334
Kay, Martin 71, 92, 689, 696, 1177–1178, 1260
Keay, Julia 2
Keeble, Richard 1251
Keenan, Edward L. 1131
Keene, Derek 1105
Kehler, Andrew 620
Kehoe, Andrew 316–317, 707
Keller, Frank 328
Kelly, Michael H. 1207
Kendon, Adam 211
Kennedy, Arthur 3
Kennedy, Claire 119
Kennedy, Graeme 1, 10, 35, 115–116, 170, 178, 989–990, 993, 1116
Kenny, Anthony 1071
Kenny, Dorothy 280, 1143, 1162, 1170
Kepser, Stefan 746–747, 923
Kermes, Hannah 605, 935, 1224
Kessler, Brett 846, 896, 1253
Kessler, Michael M. 357
Kettemann, Bernhard 112, 118, 256
Keune, Karen 910, 912
Khoja, Shereen 510–511, 517–520
Kibble, Rodger 586, 593
Kibiger, Ralf 295
Kibler, Dennis F. 858
Kiefer, Ferenc 148
Kilgarriff, Adam 138–139, 144, 310–312, 316–317, 319, 328, 505, 564–565, 573, 575, 608, 766, 922, 949, 1169, 1217–1218, 1224, 1252
Kim, Doe Hyung 271
Kim, Heunggyu 169, 393–394
Kim, Su Nam 1245
Kim, Yong-Jin 850
Kimbrell, Roy E. 1256
King, Gil W. 1175
King, Paul John 70
King, Philip 118
King, Tracy Holloway 762
Kingsbury, Paul 232, 761, 768, 864, 962, 965
Kinouchi, Osame 353
Kintsch, Walter 331, 353
Kinyon, Alexandra 955, 957
Kirk, John M. 1011, 1016, 1129
Kirriemuir, John 1265
Kirsch, David 333, 363
Kiss, Gábor 148
Kiss, Tibor 687
Kita, Kenji 895

- Kita, Sotaro 217
 Kitamura, Mihoko 699
 Kittredge, Richard I. 686
 Kjell, Bradley 1071, 1082, 1084
 Kjellmer, Göran 116, 118, 1057
 Klarskov Mortensen, Hans Jørgen 712
 Klatt, Stefan 540
 Klavans, Judith L. 77, 965, 1253, 1261
 Kleene, Stephen Cole 71
 Klein, Dan 870
 Klein, Judith 295
 Klein, Sheldon 89–90, 541
 Klein, Wolfgang 997
 Kleinberg, Jon M. 347–348, 351, 361–362, 368
 Kleiweg, Peter 896
 Klemola, Juhani 1128–1129, 1133
 Klima, Edward S. 24, 27
 Klosa, Annette 135
 Knapp, Judith 146
 Knight, Kevin 1185, 1188–1189, 1191
 Knowles, Gerry (Gerald) 192, 200, 408, 503, 510–511, 517, 520–523, 554 654, 1009–1010
 Kobourov, Stephen G. 1250
 Koch, Heinz-Detlev 503
 Koch, Peter 169
 Koehn, Philipp 287, 1163, 1185, 1188–1190
 Köhler, Reinhard 347, 352–353, 820
 Køhler Simonsen, Henrik 146–147
 Kohlhof, Inga 281, 1150
 Kohn, Kurt 112
 Kohonen, Teuvo 895
 Kohyama, Hideo 1181
 Koiso, Hanae 199
 Koivisto-Alanko, Päivi 1101–1102
 Kompe, Ralf 1039
 König, Esther 235, 765
 Koppel, Moshe 846, 1072–1073, 1084
 Korfhage, Robert R. 1253–1254
 Korhonen, Anna 957, 965–968
 Kornai, András 1026–1027
 Kortmann, Bernd 1128, 1130
 Kosala, Raymond 356, 361
 Koskenniemi, Kimmo 71, 74, 559
 Kot, Mark 364
 Kraaij, Wessel 319, 324,
 Krenn, Brigitte 766, 935–936, 1214, 1238–1239
 Kress, Gunther 221
 Krippendorff, Klaus 766
 Krishnamurthy, Ramesh 989–990, 994, 996
 Krishnamurthy, Sandeep 366
 Kristoffersen, Gjert 644
 Kroch, Anthony S. 70, 227, 426
 Kromann, Matthias T. 231
 Krott, Andrea 916
 Krovetz, Robert 566
 Krug, Manfred 1101
 Kruisinga, Etsko 10
 Kruskal, Joseph 693, 1258–1259
 Kruyt, Truus 440
 Krymolowski, Yuval 967–968
 Kübler, Sandra 226, 608, 766, 768
 Kučera, Henry 1, 35–38, 48, 90, 227, 395, 575, 760–761, 990
 Kučera, Karel 388
 Kuhlen, Rainer 331, 368–369
 Kuhn, Jonas 609
 Kujamäki, Pekka 1160, 1163, 1170
 Kukich, Karen 1259
 Kulikowski, Casimir A. 869
 Kumar, Ravi 366–368
 Kumar, Vipin 891, 895–896
 Kumari, Lalita 364
 Kung, Sun-Yuan 888, 895
 Kunz, Kerstin 232
 Kuperman, Victor 365
 Kupiec, Julian 1275
 Kurath, Hans 10
 Kurimo, Mikko 561
 Kurohashi, Sadao 231
 Kurokawa, Kazuya 1283
 Květoň, Pavel 922
 Kytö, Merja 43, 60, 66, 243–244, 246–247, 250, 255, 401, 406, 505, 1045, 1055, 1061, 1095, 1100
- L**
- Laan, Nancy M. 1072
 Laban, Rudolf 209, 217
 Labov, William 55–57, 98, 243, 248, 990–991, 1092–1093, 1097, 1112–1113, 1118, 1129
 Ladefoged, Peter 1024
 Lafferty, John D. 79, 601, 603, 863
 Laham, Darrell 896
 Lai, Jennifer C. 689, 701, 1253, 1260
 Laird, Nan 79, 869, 1185
 Laitinen, Mikko 404
 Lakoff, George 29
 Lakoff, Robin T. 104, 108

- Lam, Peter Y. W. 117
Lambert, Sylvie 1164
Lambrecht, Knud 1209
Lamel, Lori F. 644, 1028
Lampert, Martin D. 636, 1009, 1013
Lamy, Marie-Noëlle 712
Lancashire, Ian 33–34, 43, 255
Lancaster, Thomas 1261, 1263
Landau, Sabine 891, 894–896
Landau, Sidney I. 6, 8, 10
Landauer, Thomas K. 360, 896
Landes, Shari 575
Landini, Gabriel 820
Langé, Jean-Marc 283, 686, 697–698, 1179
Langendoen, D. Terrence 28, 770
Langer, Hagen 599
Langford, John C. 895
Langlais, Philippe 1178
Lanshammar, Håkan 666
Lapalme, Guy 1282
Lapata, Mirella 328, 962, 964
Lapidus, Naomi 1017
Larsen-Freeman, Diane 269
Larson, Ray R. 356, 359–360
Lascarides, Alex 618
Lassila, Ora 567
Lau, Rynson W. H. 1264
Laureys, Tom 650, 652
Lauridsen, Karen 278, 282
Laviosa, Sara 280
Laviosa-Braithwaite, Sara 1162, 1169
Lavoie, Brian F. 310, 313
Lawrence, Cameron 362
Lawrence, Helen 1133
Lawrence, Steve 312, 314, 332, 362
Lawrence, Valerie W. 1198
Lawson, Ann 287
Lea, Diana 1217, 1220
Leacock, Claudia 575
Lebart, Ludovic 877, 880, 888, 896
LeBrun, Eric 896
Lecomte, Josette 510
Ledger, Gerard 896
Ledgeway, Adam 1134
Lee, Chun-Jen 698, 702
Lee, David 385, 844, 1012
Lee, Hiang Beng 572, 575
Lee, John A. 882, 887–888
Lee, Lillian 540, 948
Leech, Geoffrey 33, 38, 40–41, 45–47, 90–91, 104, 106, 108, 113–116, 145, 158, 161–162, 166, 225, 256, 270, 386, 395–396, 422, 484, 505, 510, 582, 584, 613–614, 617, 619, 624–627, 726, 729, 760, 763, 987–988, 991–992, 994, 1009–1010, 1012, 1014–1016, 1072, 1118–1121, 1147
Lees, Robert B. 18, 21–24, 27, 37
Leese, Morven 891, 894–895
Legum, Stanley 1002
Lehiste, Ilse 204
Lehmann, Christian 554
Lehmann, Sabine 237
Lehrberger, John 1176
Leicht, Elizabeth A. 355
Lenci, Alessandro 1245
Léon, Jacqueline 989
Leonard, Rosemary 38
Leong, Hong Va 1264
Leopold, Edda 356
Leopold, Werner 1199
Leriko-Szymanska, Agnieszka 117
Lesk, Michael 569
Leskovec, Jure 347–348
Leung, Louis 293
Levelt, Willem J. M. 664, 670
Levin, Beth 954, 957, 963–966, 968
Levin, Lori 1245
Levin, Magnus 184
Levinson, Stephen C. 614–615, 619, 629, 633
Levy, Leon S. 70
Levy, Roger 765
Lewandowska-Tomaszczyk, Barbara 387, 431, 694
Lewis, Charlton T. 208
Lewis, M. Blanche 521
Lewis, Michael 116
Lewis, Philip M. 601
Leydesdorff, Loet 356, 365–366
Lezius, Wolfgang 235, 557, 747, 765, 923–924
Li, Hang 948, 961–963
Li, Mei 1000
Li, Ping 997
Li, Wen Syan 363
Li, Wentian 817, 820
Li, Xuemei 360
Liberman, Mark 674–675, 767
Liddy, Elizabeth 1282
Liebetrau, Albert M. 1234, 1237
Lieven, Elena V. M. 1198
Light, Marc 961
Lightfoot, David 1118
Lin, Chen-Yew 1278–1279

- Lin, Dekang 237, 507, 1244
 Lin, Jia 368
 Lindquist, Hans 61, 1101
 Lindström, Anders 1028–1031, 1033, 1037–1039
 Limmans, Adrianus 896
 Lippi-Green, Rosina 1094–1095
 Litman, Diane 584, 586
 Littlestone, Nicholas 858
 Ljung, Magnus 1029–1031
 Lloret, Jaime 319
 Lobacz, Piotra 895
 Loffredo, Maria I. 372
 Løken, Berit 278, 281, 286,
 Lombardo, Vincenzo 231–232
 Long, Michael H. 269
 Longacre, Robert E. 1059
 Lopes, Gabriel Pereira 895, 1244
 Lopes, José G. P. 1235
 Lopresti, Daniel P. 1255, 1259
 Lorentzen, Henrik 147
 Lorenz, Gunter 115, 117
 Lorge, Irving 10
 Lörscher, Wolfgang 1160
 Lounsbury, Floyd G. 17
 Louw, William E. (Bill) 712, 993–994
 Louwerse, Max M. 846
 Love, Alison 120
 Lowe, John 762, 962, 965, 967
 Lowenberg, Peter H. 260
 Lowth, Robert 3–4
 Lucka, Bret 1011
 Lüdeling, Anke 117, 252, 262, 266, 311, 907–910, 922, 1028, 1111, 1244
 Luhn, Hans Peter 884
 Luke, Kang Kwong 539
 Luong, Xuan 1061
 Luschützky, Hans Christian 554
 Lux, Paul 849
 Lyon, Caroline 1256–1257, 1263–1264
 Lyons, John 974, 976, 980
- M**
- Ma, Wei-Yun 536
 Maamouri, Mohamed 231
 Maas, Jan Frederik 581
 Macchi, Marian J. 1028
 Macdonald, K. 1180
 Macias, Pablo 365
 Mackey, William Francis 115
 Mackie, Andrew W. 506, 541
 Macklovitch, Elliott 1179
 MacLeod, Anna 1261
 Macleod, Catherine 953, 958, 1010
 MacWhinney, Brian 427, 467, 1199–1200, 1203–1206
 Maddieson, Ian 1024
 Madsen, Rasmus E. 861, 871
 Magerman, David 608
 Magnini, Bernardo 572
 Maguire, Warren 896
 Maia, Belinda 1143, 1168
 Maier, Wolfgang 766
 Maiorano, Steven J. 583
 Mair, Christian 61, 183, 939, 1049, 1094, 1101, 1113, 1115, 1117, 1121, 1147
 Malcolm, James 1256–1257, 1263–1264
 Malczewski, Bonnie 1053, 1063
 Malinowski, Bronislaw 55
 Mallon, Thomas 1250
 Mallory, Jim 482
 Malmkjær, Kirsten 1168
 Malpohl, Guido 1263
 Malten, Thomas 394
 Manber, Udi 1264
 Manca, Elena 278
 Mandelbrot, Benoit 803–804, 815, 817
 Mani, Inderjeet 1253, 1261, 1272, 1277–1278
 Mann, William C. 762–763, 1281
 Manning, Christopher D. 77, 116, 351, 506, 537, 543, 545, 600–601, 608, 707, 801, 819, 870, 877, 879, 882, 888, 891, 896, 955–958, 1187, 1213, 1256, 1260
 Manning, Elizabeth 990, 994
 Mannion, David 1072
 Månnsson, Ann-Christin 217
 Mao, Wenji 213
 Maratsos, Michael 1206
 Marbeck, John 2
 Marcella, Stacy 213
 Marcinkiewicz, Mary Ann 84, 386, 425, 540, 760, 857, 864, 923, 955
 Marcos-Marin, Francisco A. 1018
 Marcu, Daniel 618, 620–621, 762, 871, 1189–1190, 1275–1277
 Marcus, Gary F. 1209
 Marcus, Manfred 406
 Marcus, Mitchell P. 84, 227, 229, 231, 386, 425–426, 540, 605, 739, 752, 754, 760–761, 857, 864, 923, 955
 Marín-Arrese, Juana 1150

- Markert, Katja 620
Markie, Peter 920
Marko, Georg 112
Márquez, Lluís 541, 871
Marshall, Ian 45, 82, 91
Marsi, Erwin 862
Martin, Brian 1250
Martin, James H. 88, 560, 622, 1187, 1256
Martin, James R. 346
Martin, Jean-Claude 213
Martínez-Barco, Patricio 620
Marx, Zvika 967–968
Matsumoto, Yuji 539–540, 699, 812
Matthews, Peter 23, 554
Matthews, Robert A. J. 1076, 1082, 1088
Matthews, Stephen 1200
Matthiessen, Christian 1163
Mauranen, Anna 260, 287, 1053, 1055–
1056, 1063, 1142, 1160, 1162–1163, 1169–
1170
Maybury, Mark 1253
Mayfield Tomokiyo, Laura 1038
Maynor, Natalie 292
Mazaud, Carolin 179, 185
McCain, Katherine W. 356, 357, 360
McCallum, Andrew 863
McCarten, Jeanne 116
McCarthy, Diana 957, 963–964
McCarthy, Michael J. 116, 165, 193, 411,
994, 1011–1013, 1015–1017, 1055
McCoy, Kathleen F. 590
McCurley, Kevin 362
McDaniel, Dana 1197
McDaniel, Jan 643
McElligott, Annette 503
McEnery, Anthony M. (Tony) 34, 110, 112,
116, 132, 170, 208, 279, 286–287, 396, 515,
584, 592, 614, 616–617, 619–620, 626–
627, 693–694, 699, 701, 739, 803, 896,
990–993, 997–1001, 1003, 1071, 1101,
1144, 1253–1254, 1257
McGill, Michael J. 877, 884, 886, 896, 1276
McGowan, John A. 697
McIntosh, Angus 1098
McKee, Cecile 1197
McKelvie, David 769, 923
McKeown, Kathleen R. 699, 702, 968,
1234–1235, 1261, 1265, 1277, 1281
McLaren, Yvonne 278
McLaughlin, Jamie 1088
McNeill, David 217, 220
McRoy, Susan 571
McTait, Kevin 1182
Mealand, David L. 896, 1076, 1082
Medori, Julia 1263
Mehl, Nathali 1178
Mehler, Alexander 296, 328–329, 331–332,
349–350, 363, 370, 372–373
Mehrabian, Albert 221
Meibauer, Jörg 910
Meijs, Willem 34
Meinel, Christoph 364
Mel'čuk, Igor 230
Melamed, Dan I. 695, 1260, 1262
Melander Marttala, Ulla 256
Melby, Alan 1177–1178
Melchers, Gunnel 1027
Melnikov, Oleg 335
Mencken, Henry Louis 1112
Menczer, Filippo 329–330
Mendenhall, T. C. 1071
Mendes, Jose Fernando Ferreira 351–352
Menzel, Wolfgang 601
Mercer, Robert L. 541, 601, 689, 701, 903,
1253, 1260
Merelo, Juan 367
Merialdo, Bernard 541
Merlo, Paola 965–967
Merriam, Thomas V. N. 1076, 1082
Mester, Arnim 1028
Metzler, Donald 1252–1255
Meunier, Fanny 116–117, 121, 262, 265, 283
Meunier, Frederic 697, 1179
Meurer, Paul 262
Meurers, Walt Detmar 235, 765, 921–923
Meurman-Solin, Anneli 248, 847, 1100, 1104
Meyer, Bernd 1164
Meyer, Charles 1, 10, 12, 157–159, 161–
162, 164–165, 310, 312–313, 321, 323, 987,
989, 1002
Meyers, Adam 761, 767, 953, 958
Mezquiriz, David 316–317
Miceli, Tiziana 119
Michaelson, Sidney 1071, 1078
Michelbacher, Lukas 1245
Mihailov, Mikhail 1164
Mika, Peter 369
Mikheev, Andrei 533–536, 540
Milgram, Stanley 333, 339
Milić, Louis 244, 251–252, 1071
Miller, Don E. 1008
Miller, George A. 24–25, 71, 354, 803, 817,
895
Miller, James D. 653

- Miller, Jim 1053, 1061
 Milo, Ron 341, 343, 346, 361
 Milroy, James 56, 57, 60, 1094, 1128, 1130
 Milroy, Lesley 98, 1094, 1128, 1130
 Milton, John 262, 266, 271, 430, 1088
 Miltakaki, Eleni 232, 618, 762
 Mindt, Dieter 114–115, 980
 Minkova, Donka 1092
 Minnen, Guido 560, 958
 Mintz, Toben H. 1207
 Mírovský, Jiri 747
 Mitchell, Rosamond 262
 Mitchell, Tom 855, 869
 Mitkov, Ruslan 580–583, 585, 591–594,
 1273–1274, 1276, 1284
 Mladenić, Dunja 318, 462
 Mo, Lili 396, 1001
 Möbius, Bernd 818
 Moens, Marc 1275, 1283
 Moffat, Alistair 820
 Mohri, Mehryar 558
 Moisl, Hermann 890, 896
 Mol, Susan 282
 Mollin, Sandra 172
 Monaghan, Padraig 1207
 Mondorf, Britta 1017
 Monostori, Krisztián 1263–1264
 Montemagni, Simonetta 231–232
 Montgomery, Michael B. 1037
 Moon, Rosamund 1009
 Mooney, Raymond J. 868–869
 Moore, Colette 1101–1102
 Moore, Roger 628, 643
 Moreno, Antonio 231
 Morgan, James L. 1207
 Morley, John 1053
 Morra, Lucia 286
 Morris, Charles 614
 Morris, Meaghan 730
 Morrison, Andrew 120
 Morset Størseth, Torbjørn 1038
 Morton, Andrew Queen 1071, 1078
 Moscoso del Prado Martín, Fermín 904
 Mosel, Ulrike 772
 Moser-Mercer, Barbara 1164
 Mosteller, Frederick 77, 1076–1077, 1083
 Motter, Adilson E. 354
 Mparutsa, Cynthia 120
 Mugdan, Joachim 554
 Mukherjea, Sougata 361–362
 Mukherjee, Joybrato 112, 121, 123, 189
 Mulac, Anthony 1292
 Muller, Charles 148, 820
 Müller, Christoph 628, 765
 Müller, Frank Henrik 930
 Müller, Hans 535
 Müller, Stefan 925, 927–929
 Munkres, James 880
 Muñoz, Rafael 620
 Murata, Masaki 1259
 Murison-Bowie, Simon 987
 Murphy, Brona 415, 1011
 Murray, Denise E. 301
 Murray, James A. H. 8–9
 Murty, M. Narsimha 891, 894
 Myles, Florence 262
- N**
- Nagao, Makoto 231, 687, 1181
 Nagata, Masaaki 540
 Nakamura, Masami 541
 Naranan, Sundaresan 816
 Nardi, Bonnie A. 366
 Narin, Francis 358
 Narita, Masumi 1283
 Naro, Anthony J. 1132
 Natalis, G. 535
 Nathan, Mordecai (Isaac Nathan ben Kalony-mus) 2
 Nation, Paul 115
 Nattinger, James R. 116
 Naumann, Karin 296, 583, 591, 593, 768
 Navarro, Borja 620
 Needleman, Saul B. 1259
 Neijt, Anneke 908–909, 914
 Nelfelt, Kerstin 217
 Nelson, Gerald 155, 165, 193, 227, 231, 398,
 424, 738–739, 742, 744, 746–748, 750–
 753, 757, 989, 1009, 1011
 Nenkova, Ani 1278–1279
 Nerbonne, John 602, 896
 Nerima, Luka 1215, 1224
 Nero, Shondel 1011
 Nesselhauf, Nadja 112, 117, 261, 263, 271
 Neuhaus, Hans J. 909–910
 Neumann, Gerald 391
 Neumann, Stella 1160–1161, 1167, 1170
 Nevala, Minna 56, 1099
 Nevalainen, Terttu 55–56, 61, 243, 248, 404,
 1097–1100, 1103–1104
 Neves, Luís 1038
 Newman, Eamonn 1280

- Newman, Mark E. J. 330, 333, 339–340, 342–343, 345–346, 348, 355, 361
 Newmeyer, Frederick J. 19, 988
 Newport, Elissa L. 1207
 Ney, Hermann 287, 697, 1164, 1186, 1188–1189
 Ng, Andrew Y. 861, 871
 Ng, Hwee Tou 572, 574
 Niblett, Tim 858
 Nicely, Patricia E. 895
 Nicholls, Diane 266
 Niculescu-Mizil, Alexandru 865
 Nida, Eugene 1160
 Nie, Jian-Yun 319, 324
 Niemann, Heinrich 1038
 Niessen, Sonja 1188
 Nijholt, Anton 213, 215
 Nilsson, Jens 762
 Nioche, Julien 313
 Nirenburg, Sergei 1182
 Nishimoto, Eiji 903
 Nishimoto, Kazushi 304
 Niu, Cheng 532
 Nivre, Joakim 218, 220, 226, 232, 763
 Noël, Dirk 939
 Nolan, Francis 412
 Nölke, Henning 1056
 Nöller, Claudia 667, 680
 Nordberg, Bengt 1036
 Norén, Coco 1056
 Norling-Christensen, Ole 135
 Norrby, Cartrin 1034
 North, Ryan 1244
 Nöth, Elmar 1038
 Nunberg, Geoffrey 846, 1253
 Nunes, Ricardo 1038
 Nurmi, Arja 245, 249–250, 1099–1100, 1103
 Nygaard, Lars 287
- O**
- O'Donoghue, Timothy 505, 507
 O'Donovan, Ruth 955, 957–958
 O'Grady, William 1209
 O'Keeffe, Anne 193, 415, 1011–1013, 1015–1016
 O'Neill, Edward 310, 313
 O'Reilly, Tim 363
 Oakes, Michael 693–694, 801, 803, 888, 891, 894–896, 1071, 1144, 1253–1254
 Oard, Douglas W. 1192
- Och, Franz Josef 287, 697, 1164, 1186, 1188–1192
 Ochs, Elinor 1013, 1045
 Oepen, Stephan 232, 234
 Oflazer, Kemal 231, 555, 561
 Ogden, Charles K. 140
 Ogura, Kanayo 304
 Oh, Sun-Young 1292
 Ohara, Kyoko Hirose 962
 Ohlander, Sölve 322
 Oja, Merja 895
 Oksefjell, Signe 284, 433, 1143
 Okumura, Manabu 1280
 Okurowski, Mary Ellen 618, 621, 762
 Oliva, Karel 922
 Olofsson, Arne 48
 Olohan, Maeve 280, 286, 1141, 1154, 1169–1171
 Oltvai, Zoltán N. 343
 Ondruška, Roman 747
 Ooi, Vincent B.Y. 303, 1011
 Oostdijk, Nelleke 47, 193, 195–196, 202, 644, 678, 843, 846–847, 860
 op den Akker, Rieks 213
 Oppenheim, Alan V. 666
 Orasan, Constantin 583, 585, 591, 593–594, 765, 1273–1274, 1276–1277, 1282, 1284
 Orlov, Jurij K. 331, 373
 Orton, Harold 411, 1128
 Osborne, Randy 602
 Osen, Janet 1250
 Österman, Aune 1101
 Österreicher, Wulf 169
 Ostler, Nicholas 134, 264, 458, 1012
 Östman, Carin 256
 Östman, Jan-Ola 1056, 1058
 Oswald, Victor A. 77
 Otheguy, Ricardo 1017
 Otte, Evelien 334, 338, 360
 Otterbacher, Jahna 1281
 Ovens, Janine 1011
 Owen, Marion 505
- P**
- Paccagnella, Luciano 301
 Padró, Lluís 541
 Page, Ellis B. 77
 Paice, Chris D. 1283
 Pajzs, Júlia 148
 Palander-Collin, Minna 846, 1101–1102

- Palen, Leysia 293
Pallett, David S. 644
Palmer, Chris C. 1096
Palmer, David D. 535, 540, 687, 879
Palmer, Martha 232, 564, 575, 761, 766, 768, 864, 962, 965
Pan, Haihua 1000
Pander Maat, Henk 1050
Panovová, Jarmila 763
Pāṇini 3
Pankow, Christiane 295
Panyametheekul, Siriporn 295
Paolillo, John 295–296
Papageorgiou, Harris 694, 1179
Papineni, Kishore 1256
Papp, Szilvia 121
Park, Han Woo 334, 361
Park, Juyong 345
Parodi, Giovanni 849
Paroubek, Patrick 74
Parsons, Terence 973
Partington, Alan 112, 114, 977, 979–980, 993, 1053–1054
Passonneau, Rebecca 584, 586, 1278–1279
Patocka, Franz 1130
Paul, Hermann 4, 929
Paulussen, Hans 277
Pawley, Andrew 116, 482
Pearce, Darren 560
Pearson, Egon 783
Pearson, Jennifer 117, 122, 137, 280, 1168
Pearson, Karl 783
Pecina, Pavel 1236
Pedersen, Jan O. 566, 948, 1085, 1275
Pedersen, Jette 136
Pedersen, Ted 569, 896, 1235
Peirce, Charles Sanders 209, 220
Pennock, David M. 362
Pereira, Fernando 602, 863, 948
Pereira de Oliveira, Maria José 117
Pérez, A. 681
Pérez Basanta, Carmen 121
Pérez-Paredes, Pascual 271
Perkuhn, Rainer 139, 142
Perrault, Francois 622, 1178
Perrez, Julien 262
Persson, Olle 357
Petch-Tyson, Stephanie 112, 267, 269
Peters, Carol 698
Peters, Pam 42, 102, 116
Petrakis, Euripides G. M. 1255
Pettersson, Per A. 1036
Pfeil, Ulrike 293
Philippsen, Michael 1263
Piao, Scott 701–702, 1252–1253, 1265
Picchi, Eugenio 698
Pickering, Lucy 1017
Picone, Joseph 655
Piepenbrock, Richard 553, 904
Pietsch, Lukas 1131–1132, 1135
Pike, Kenneth 26
Pine, Julian M. 1198
Pinker, Steven 911, 956, 1206
Pinski, Gabriel 358
Piotrowski, Rajmund 820
Piperidis, Stelios 694, 1179
Pito, Richard 923
Plaehn, Oliver 234–235, 765
Plag, Ingo 905, 908, 914–916
Planas, Emmanuel 1179
Pleasant, Nigel 313
Pluymakers, Mark 911
Pöchhacker, Franz 1164
Poesio, Massimo 584, 587, 594, 628
Poggi, Isabella 211, 220
Polanyi, Livia 333
Pollard, Carl 70, 599
Pollatschek, Moshe A. 1075
Pollatsek, Alexander 914
Polomé, Edgar 243
Poos, Deanna 164
Popescu-Belis, Andrei 583
Poplack, Shana 1132, 1134
Porter, Constance Elise 300
Porter, Martin F. 553, 697
Poschenrieder, Thorwald 252
Post, Brechtje 412
Postal, Paul M. 26–29
Pothos, Emmanuel M. 911
Poutsma, Arjen 1191
Poutsma, Hendrik 4, 10, 28
Power, Richard 622
Prandi, Michele 973–974
Pravec, Norma A. 261
Prechelt, Lutz 1263
Preece, Jenny 614
Preiss, Judita 564
Preston, Dennis R. 1136
Prillwitz, Siegmund 217
Prime, Camille 356, 359
Prince, Alan 1027
Prince, Ellen F. 1292
Prinsloo, Daniel J. (Daan) 132, 135, 139–141, 472

- Probst, Julia 1161
 Procter, Paul 572
 Prodromou, Luke 1017
 Prolo, Carlos A. 955, 957
 Pruvost, Jean 135
 Przepiórkowski, Adam 387, 555
 Pullum, Geoffrey K. 29, 1146
 Punyakanok, Vasin 863
 Pusch, Claus D. 97, 1018, 1134
 Pustejovsky, James 139, 574, 762, 765, 973
 Putnam, Hilary 979
 Pyle, Dorian 876–878, 880, 882, 888, 895

Q

- Qiao, Hong Liang 505
 Quaglio, Paulo Marques 845
 Quasthoff, Uwe 349, 351–352
 Quemada, Bernard 35
 Quereda, Luis 112
 Quinlan, J. Ross 535, 858
 Quirk, Chris 1259
 Quirk, Randolph 10–12, 35, 39–40, 48, 148, 209, 749, 1009, 1014

R

- Rábade, Luis Iglesias 1143
 Rabin, Michael O. 71
 Rabiner, Lawrence 666
 Radday, Yehuda T. 1075
 Radev, Dragomir 1272, 1278, 1281–1282
 Radford, Andrew 988
 Ragan, Peter H. 262–263
 Raghavan, Prabhakar 362, 368
 Rahav, Giora 194
 Raible, Wolfgang 330
 Rainer, Franz 794
 Rajman, Martin 877, 880, 888, 896
 Ramshaw, Lance A. 605
 Randall, Beth 747, 923
 Rapoport, Anatol 330–331, 340, 343, 365
 Rapp, Reinhard 557, 1179
 Rapp, Stefan 1039
 Ratnaparkhi, Adwait 535, 539–541, 687, 858, 866
 Raumolin-Brunberg, Helena 55–56, 243, 1097–1100, 1102–1103
 Ravasz, Erzsábet 343, 346–347, 352
 Ravichandra Rao, Inna Kedage 356

- Rayson, Paul 104, 106, 108, 158, 161–162, 256, 421, 571, 620, 726, 729, 1072, 1118–1120
 Reah, Danuta 1251
 Rebeyrolle, Josette 137
 Reddick, Allen 7–8
 Reder, Stephen 263, 1018
 Redington, Michael 1206–1207
 Redner, Sidney 358–359
 Rehm, Georg 296, 363
 Reid, Elizabeth M. 292, 303
 Renouf, Antoinette 41–42, 116, 158, 310–312, 314, 316–317, 322, 709, 901, 905, 909–910, 936, 938, 1122
 Rényi, Alfréd 342
 Reppen, Randi 132, 134, 386, 791, 803, 827, 847–848, 948, 1010, 1012, 1015, 1056, 1092, 1289
 Resnik, Philip 77, 318, 328, 601, 923, 961–962
 Rey, Jennifer M. 845
 Reynar, Jeffrey C. 535, 687, 858
 Ribeiro-Neto, Berthier 329, 707, 1256
 Ricca, Livio 904–905
 Riccardi, Giuseppe 1028
 Richardson, Graham 664, 670
 Richardson, Malcolm 1104
 Rickford, John R. 1123
 Ridings, Daniel 277, 1168
 Rieger, Eliezer 16
 Rieger, Matthias 1261, 1263
 Riezler, Stefan 237
 Rigau, German 569
 Riley, Michael D. 535, 540
 Riloff, Ellen 628
 Ringbom, Håkan 269
 Riordan, Oliver 330, 336, 340–342, 344, 347, 353
 Rissanen, Jorma 859
 Rissanen, Matti 43, 57, 64–65, 244, 246, 248, 402, 859, 1092, 1100–1104
 Ristad, Eric Sven 1259
 Ritchie, Graeme 554
 Ritter, Helge 667, 680
 Rivlin, Ehud 361
 Riza, Hammam 394
 Roberts, Andrew 517
 Robertson, Stephen 886
 Rocha, Marco 583, 589
 Rocha, Paolo Alexandre 440
 Roche, Emmanuel 558
 Rocio, Vitor 227

- Rock, Frances 1054
 Rohde, Douglas 747, 765, 923
 Rohdenburg, Günter 1115, 1117, 1121
 Rohrbach, Jan-Marc 121
 Rohrer, Christian 925
 Roland, Douglas 949, 955
 Romaine, Suzanne 56, 101, 1097, 1100, 1120
 Romary, Laurent 235, 287, 583–584, 587,
 674, 680, 772, 1178
 Römer, Ute 112, 114–116, 120
 Rooth, Mats 955–959, 967
 Röscheisen, Martin 689, 696, 1260
 Roseberry, Robert L. 112, 117
 Rosenbach, Anette 316
 Rosenbaum, Peter 28
 Rosenblatt, Frank 858
 Rosenkrantz, Daniel J. 601
 Rosenzweig, Joseph 766
 Ross, Alan S. C. 103
 Ross, John Robert 29, 927
 Rossini Favretti, Rema 174, 390
 Roth, Dan 863
 Roth, Michal 602
 Rottweiler, Gail Price 156
 Roukos, Salim 79, 858
 Rounds, William C. 70
 Rousseau, Brendan 358
 Rousseau, Ronald 334, 338–339, 357–358,
 360
 Rousseeuw, Peter J. 967
 Roweis, Sam T. 895
 Rubin, Donald B. 79, 869, 1185
 Rubin, Gerald M. 43–44, 90, 503, 541
 Ruiz, Victor 367
 Rumelhart, David E. 858
 Rundell, Mike 116, 310, 313, 322, 1009
 Runkehl, Jens 297, 303
 Russell, Graham 689, 1179, 1260
- S**
- Sacau Fontenla, Elena 435
 Sack, Harald 364
 Sacks, Harvey 618, 623
 Sadoff, Steven J. 653
 Sag, Ivan A. 70, 599, 1213
 Sager, Juan C. 1169
 Saggion, Horacio 1282
 Sahakyan, Marina 1039
 Sahlgren, Magnus 1224
 Sáiz, Marina 849
 Sajavaara, Kari 1142
 Salager-Meyer, Françoise 1282
 Salkie, Raphael 287, 1141
 Salmon-Alt, Susanne 959
 Salton, Gerard 877, 884, 886, 896, 1276
 Sampson, Geoffrey 45, 225, 227–228, 231,
 233, 386, 421–423 614, 616, 644, 760, 804,
 995
 Samuels, Michael 56–58
 Samuelson, Pamela 1250
 Samuelsson, Yvonne 227, 1161
 Sand, Andrea 395
 Sanders, Eric 659
 Sanders, Ted 1050
 Sanderson, Mark 1252
 Sandiford, Helen 116
 Sanfilippo, Antonio 608
 Sang, Erik Tjong Kim 605
 Sankoff, David 896, 1258–1259
 Santamaría, Celina 328
 Santamaria, Carmen 1018
 Santana, Juan 112
 Santini, Marina 1071
 Santorini, Beatrice 84, 386, 425, 503, 508,
 540, 760, 765, 857, 864, 923, 955
 Santos, Diana 287, 440, 1151
 Sarkar, Anoop 959
 Sasaki, Felix 232
 Sato, Hiroaki 962
 Sato, Satoshi 1181
 Saul, Lawrence K. 895
 Savage, T. R. 1275
 Savitch, Walter J. 71
 Savoy, Jacques 697
 Saxena, Anju 762
 Scannell, Kevin P. 461–462, 465
 Scha, Remko 602
 Schabes, Yves 74, 558, 601–602
 Schachter, Paul 26
 Schaden, Stefan 1038–1039
 Schaeder, Burkhard 33, 40
 Schafer, Ronald 666
 Schäfer, Lande 265
 Schäffler, Hildegard 1143
 Schegloff, Emanuel A. 618, 623
 Schenker, Adam 333
 Scherer, Carmen 910
 Scherre, Maria Marta Pereira 1132
 Schervish, Mark J. 780, 783–784, 786–787,
 789, 792, 801, 1235, 1245
 Schiehlen, Michael 1224
 Schiffer, Stephen R. 627

- Schiffrin, Deborah 1287
Schiller, Anne 505, 924
Schleimer, Saul 1263–1264
Schlesinger, Izchak M. 1206
Schlobinski, Peter 297, 303
Schlüter, Julia 1121
Schlüter, Norbert 112, 115
Schmid, Hans Jörg 256
Schmid, Helmut 532, 536, 540–541, 545, 560–561
Schmid, Tanja 1028, 1033
Schmidt, Eric 312
Schmidt, Heinz 1263–1264
Schmidt, Ingrid 492
Schmidt, Jan 366
Schmidt, Ruth 515
Schmidt, Thomas 769
Schmied, Josef 118, 287, 403, 1011, 1142–1143, 1145, 1154
Schmitt, Norbert 1016
Schneider, Peter 405
Schnörch, Ulrich 135
Scholze-Stubenrecht, Werner 138
Schönberger, Klaus 366
Schone, Patrick 1221, 1238
Schonefeld, Oliver 769
Schönefeld, Doris 189, 949
Schonell, F. J. 1009
Schreuder, Robert 916
Schröder, Ingo 601
Schüller, Dietrich 1129
Schulte im Walde, Sabine 142, 601, 608, 948, 958–959, 966–968
Schultink, Henk 907, 916
Schummer, Joachim 359
Schuster, Heinz Georg 348
Schütze, Carson 777, 989, 991
Schütze, Hinrich 77, 351, 506, 537, 543, 545, 566, 570, 574–575, 600–601, 608, 707, 801, 819, 846, 877, 879, 882, 888, 891, 896, 948, 1187, 1213, 1221, 1224, 1245, 1253, 1256, 1260
Schuurman, Ineke 864
Scott, Brad 411
Scott, Dana 71
Scott, Mike 112, 157, 385, 730–731
Searle, John Rogers 614, 616–617, 622, 632
Sebba, Mark 35
Seglen, Per O. 359
Seidenberg, Mark 914
Seidlhofer, Barbara 121, 260, 270–271, 1017
Selinker, Larry 264
Selman, Bart 334
Selting, Margret 202
Semino, Elena 166, 619, 634, 1061
Sengupta, I. N. 364
Seppänen, Aimo 310, 323
Serratrice, Ludovica 1200
Setzler, Kristen 263, 1018
Seuren, Pieter 989
Sgall, Petr 232, 763
Shah, Mehul 334
Shannon, Claude E. 80, 87, 914
Sharoff, Serge 314, 319, 389–390
Sharp, Harriet 1029
Shastri, S. V. 42
Shavlik, Jude W. 857
Shaw, Philip 1027
Shei, Chris 988
Sheldon, Amy 1209
Shen, Libin 539
Shepard, Roger N. 895
Shi, Rushen 1207
Shimonov, Anat Rachel 846, 1072–1073, 1084
Shinya, Yusuke 1261
Shipley, Elizabeth 1198
Shivakumar, Narayanan 1252, 1255, 1257, 1263–1264
Shlesinger, Miriam 1164
Shneiderman, Ben 361
Shobbrook, Katherine 190
Shohamy, Elana 845
Shorrocks, Graham 1128
Short, Charles 208
Short, Mick 166, 619, 634, 1061
Shriberg, Elizabeth 203–204, 1025
Shuy, Roger 1016
Si, Antonio 1264
Salm, Ambros 1245
Sichel, Herbert S. 1073–1074
Siebenhaar, Beat 295
Siegel, Eric V. 965, 967
Siegel, Sidney 1276
Siemund, Rainer 185, 395
Siever, Torsten 297, 303
Siewierska, Anna 1132, 1136–1137
Sigman, Mariano 344, 354–355
Signorini, Allessio 312
Sigogneau, Anne 357, 360
Silverman, Emily 364
Silverman, Kim 654
Simard, Michel 283, 319, 324, 532, 686, 693–695, 1151, 1178, 1260
Šimková, Maria 392

- Simmons, Robert F. 89–90, 541
 Simon, Herbert A. 344, 365, 749, 803, 817
 Simons, Gary 772
 Simon-Vandenbergen, Anne-Marie 277, 248, 285–286
 Simov, Kiril 226, 232, 235
 Simpson, Rita C. 164, 1011
 Sinclair, John McH. 41–42, 48–49, 112, 114, 116, 118–120, 123–124, 138, 157, 165–166, 260, 278, 351, 618, 624, 631, 712, 726, 818, 936–938, 977–980, 982–985, 989, 991, 1012, 1047, 1056–1059, 1213–1215, 1217, 1221, 1223, 1229, 1245
 Singhal, Amit 880
 Sireci, Stephen G. 894
 Siro, Paavo 26
 Skandera, Paul 395
 Skinner, Burrhus Frederic 22
 Skopeteas, Stavros 765
 Skousen, Royal 904
 Skut, Wojciech 605, 607
 Skutnabb-Kangas, Tove 1027
 Slade, Diana 1009
 Sleator, Daniel 603
 Slembrouck, Stef 172
 Slobin, Dan I. 911, 1209
 Sloman, B. 497
 Smadja, Frank 699, 1215, 1234–1235
 Small, Henry 358, 360
 Small, Steven L. 568, 571
 Smith, Carlota 997–1000
 Smith, Jennifer 1133
 Smith, M. W. A. 1072, 1085
 Smith, Marc A. 304, 364, 381
 Smith, Nicholas 396, 1121
 Smith, Noah A. 318, 328
 Smith, Temple F. 1259
 Smith Cairns, Helen 1197
 Smitterberg, Erik 1100
 Smolensky, Paul 1027
 Smyth, Padhraic 342, 361–362, 820
 Sneddon, James 521
 Snell-Hornby, Mary 1160
 Snow, Catherine 468, 1199–1200
 Söderberg, Barbro 1033
 Solé, Ricard V. 347, 351, 352, 353
 Somers, Harold 566, 1178, 1180–1181
 Soricut, Radu 1256
 Sorjonen, Marja-Leena 1055
 Souter, Clive 428, 503, 505, 507, 512
 Spärck Jones, Karen 885–886, 1272
 Spencer, Andrew 554
 Sperber, Dan 614, 617, 626
 Sperberg-McQueen, C. Michael 47, 282, 486
 Spiegel, Murray F. 1028
 Spohr, Dennis 147
 Spöttl, Carol 1016
 Spranger, Kristina 959
 Sproat, Richard 554, 701
 Srinivas, Bangalore 1191
 Sripicharn, Passapong 119
 Stabler, Edward 70
 Stahl, Daniel 1198
 Stamou, Sofia 232
 Stampe, David 1028
 Stark, Mariza 135
 Starkweather, John A. 1008
 Stavestrand, Helga 123
 Stede, Manfred 76, 618, 620, 763
 Steedman, Mark 70, 74, 232, 762
 Stefanowitsch, Anatol 921, 935, 941–942, 944, 948–949, 992, 1000
 Stegbauer, Christian 366
 Stein, Dieter 293, 1100
 Steinbach, Michael 891, 895–896
 Steiner, Erich 1143, 1170
 Steiner, Ilona 747, 923
 Steingrimsson, Steinthor 1039
 Stemmer, Georg 1038
 Stenström, Anna-Brita 33, 39, 48, 97–98, 104–105, 108, 110, 194, 410, 503, 630–631, 1010, 1016–1017, 1053–1055, 1130
 Stern, Clara 1198
 Stern, Karen 412, 1009
 Stern, William 1198
 Stevens, Vance 119–120
 Stevenson, Mark 564–565, 567, 570, 576, 896
 Stevenson, Suzanne 964–968, 1244
 Stewart, Dominic 112, 280, 1168
 Stewart, Miranda 1056
 Steyvers, Mark 344, 348–350, 354–355
 Stibbard, Richard M. 201, 423
 Stiles, William B. 614, 621–622, 632–633, 637
 Stockwell, Robert 1092
 Stokoe, William C. 217
 Stolcke, Andreas 623–624, 637
 Stone, Philip J. 567
 Storjohann, Petra 135
 Stork, David G. 891, 894–895
 Storrer, Angelika 293, 299, 303, 332, 581
 Strogatz, Steven H. 330, 335–336, 339, 341–342, 346

- Strömquist, Siv 256
Strube, Michael 590, 620, 628, 765
Strunk, Jan 687
Strzalkowski, Tomek 896
Stubbe, Maria 1055
Stubbs, Michael 114, 156, 328, 330–331,
 351, 373, 949, 992–993, 996, 1054, 1238,
 1297
Stump, Gregory T. 553
Subirats, Carlos 962
Suderman, Keith 136, 769, 1010
Sumita, Eiichiro 1181, 1188
Sun, Jian 538–539
Sutcliffe, Richard 503
Suzuki, Yoshimi 1281
Svartvik, Jan 11–12, 38–40, 48–49, 105,
 194, 196, 209, 408, 654, 760, 1009, 1014,
 1054
Svensson, Mikael 433
Swales, John M. 1053, 1063, 1274
Sweet, Henry 4
Swerts, Marc 189
Syder, Frances H. 116
Syrdal, Ann K. 654
Szembricsanyi, Benedikt 1132
- T**
- Taavitsainen, Irma 248, 1062, 1103–1104
Tabachnick, Barbara 876, 888, 890–891
Tagliamonte, Sali A. 1132–1136
Taglicht, Josef 48
Tagnin, Stella E. O. 262
Tajima, Keishi 363
Takahashi, Masako 70
Tamburini, Fabio 390, 1144
Tan, Pang-Ning 891, 895–896
Tanaka, Izumi 584–585
Tanaka, Katsumi 363
Tanaka, Kumiko 698
Tanaka, Takaaki 1245
Tang, Rong 360
Tannen, Deborah 1287
Tao, Hongyin 1016
Tapanainen, Pasi 533–534, 536, 540–541
Tapper, Marie 1053
Tarp, Sven 132–133, 136, 146
Tatlock, John S. 3
Tauberer, Joshua 498
Tavakolian, Susan L. 1209
Taylor, Ann 227, 232, 234–235, 254, 426
Taylor, Lolita (Lita) 40, 408, 503, 654, 1009
Taylor, Malcolm J. 895
Teahan, Bill 505
Teich, Elke 677, 1161–1164, 1167–1170
Teleman, Ulf 226, 233, 762, 1034
Temperley, Davy 603
Temple, J. T. 896
Temple, Rosalind 1133
ten Bosch, Louis 1039
Tenenbaum, Joshua B. 344, 348–350, 354–
 355, 381, 895
Tenfjord, Kari 262
Tengi, Randee 575
Terra, Egidio 1220
Tesnière, Lucien 70, 603
Teubert, Wolfgang 118, 156–158, 166, 276,
 281, 979, 1142, 1154, 1168
Teufel, Simone 505, 924, 1275, 1277, 1283
Thal, Donna J. 1198
Thelwall, Mike 328, 339, 360, 365
Thewissen, Jennifer 265
Thibault, Paul J. 160
Thielen, Christine 295, 505, 924
Thomas, Dylan 24
Thomas, Jenny 566, 1053
Thomas, Margaret 264
Thomason, Sarah 1028
Thompson, Cynthia A. 868–869
Thompson, Geoff 1056
Thompson, Henry S. 769
Thompson, Paul 419, 1018
Thompson, Richard A. 600
Thompson, Sandra A. 762–763, 1002, 1009,
 1208, 1281, 1292
Thorndike, Edward L. 10
Thurstun, Jennifer 120
Tichy, Walter F. 1259
Ticrea, Miruna 1280
Tiedemann, Jörg 287, 694, 1259
Tilling, Philip M. 1128
Tillmann, Christoph 1188–1189
Tinsley, Howard E. A. 876, 891
Tirkkonen-Condit, Sonja 1169
Tishby, Naftali 948
Tissari, Heli 1100
Todla, Sunisa 295
Tognini Bonelli, Elena 114, 134, 138, 278,
 707, 711, 993–994, 996, 1012, 1048–1049
Tokunaga, Takenobu 1245
Tomás, Jesús 319
Tomasello, Michael 1198, 1208–1209
Tomita, Masaru 601

- Tomokiyo, Takashi 366–367
 Tono, Yukio 116, 132, 991
 Torres, Armando 365
 Tottie, Gunnell 48
 Toury, Gideon 1160
 Toutanova, Kristina 237
 Towell, Geoffrey 575
 Tran, Nicholas 1259
 Trancoso, Isabel 1038
 Trap-Jensen, Lars 146
 Trask, Larry 482
 Traugott, Elizabeth Closs 58, 60
 Travis, Catherine E. 1132
 Tribble, Chris 112, 119, 121–122, 124, 271,
 730–731
 Tricas, Fernando 367
 Trier, Jost 974
 Trippel, Thorsten 491, 493–494
 Trosterud, Trond 483
 Trotta, Joe 310, 323
 Trudgill, Peter 98, 896, 1112–1113, 1126,
 1129
 Trujillo, Arturo 1182
 Tsang, Vivian 964, 966–967
 Tsou, Benjamin K. 401, 1275
 Tsujii, Jun-ichi 1181
 Tugwell, David 139, 144, 608
 Tuldava, Juhani 351
 Tür, Gökhane 561
 Turnbull, Jill 121
 Turney, Peter D. 322
 Tutin, Agnes 583, 588–589, 1213–1214,
 1245
 Tweedie, Fiona 819, 1071–1072, 1079, 1082
 Tyson, Stephanie 427, 1053
 Tzeras, Kostas 1255

U

- Uchimoto, Kiyotaka 430
 Ueffing, Nicola 1188
 Ueyama, Motoko 319, 322, 812
 Uitdenbogerd, Alexandra L. 1255
 Ule, Tylman 235, 869, 930
 Ullendorf, Edward 482
 Ullman, Jeffrey D. 558
 Upton, Thomas 112, 845, 847
 Ushioda, Akira 955–957
 Utiyama, Masao 702
 Utsuro, Takehito 699, 1283
 Uzar, Rafal 261
 Uzuner, Ozlem 1253–1254

V

- van Bergen, Linda 407
 van de Velde, Hans 653
 van Deemter, Kees 586, 593
 van den Bosch, Antal 864, 870
 van den Heuvel, Henk 643, 659
 van den Heuvel, Theo 46–47, 747
 van der Hulst, Harry 217
 van der Wouden, Ton 227, 1055
 van Dijk, Teun A. 331, 1251
 van Eckert, Edgar 301
 van Eynde, Frank 866
 van Gijn, Ingeborg 217
 van Halteren, Hans 47, 747, 865–866, 922,
 1071–1072
 van Hout, Roeland 653
 van Leeuwen, Theo 221
 van Raan, Anthony F. 343, 349, 357–360
 van Rijssbergen, Cornelis Jost 884, 886, 896
 van Rooy, Bertus 265
 van Son, Rob J. J. H. 672
 van Sterkenburg, Piet 132–133
 van Zaanen, Menno 79
 Vandeghinste, Vincent 864
 Vander Beke, George E. 140
 Vangsnes, Øystein A. 1134
 Vapnik, Vladimir N. 539, 858
 Váradi, Tamás 389, 422, 760
 Varantola, Krista 324
 Varjokallio, Matti 561
 Vasishth, Shravan 802
 Vasko, Anna-Liisa 1128
 Vaughan, Liwen 339
 Vázquez, Gloria 966
 Veblen, Thorstein 24
 Veenstra, Jorn 605, 869
 Vela, Mihaela 1161, 1167, 1169
 Vendler, Zeno 973, 997–998
 Ventola, Eija 331
 Verdejo, Felisa 328
 Verhagen, Arie 1208
 Verkuyl, Henk J. 997
 Verleyen, Michel 882, 887–888
 Verlinde, Serge 146
 Véronis, Jean 283, 564–566, 766, 1143–1144
 Verschueren, Jef 632, 1061
 Viberg, Åke 277, 283, 1148
 Vickrey, David 576
 Viegas, Fernanda B. 293
 Vieira, Renata 584, 587, 594, 628
 Vihman, Marilyn May 1209

Vilkuna, Maria 469
Villada Moirón, Begoña 1245
Vilmi, Ruth 292
Vilnat, Anne 230
Vinay, Jean-Paul 1160
Vine, Bernadette 414, 1013, 1055
Vintar, Špela 1154, 1168
Virtanen, Tuija 1047, 1050, 1052, 1056, 1058
Visser, Frederikus Th. 1113
Viterbi, Andrew J. 79, 537, 543, 605
Voegelin, Charles W. 17
Voghera, Miriam 1018
Volk, Martin 227, 323, 1161, 1164
Voorhees, Ellen M. 575
Voss, Jakob 369–371
Vossen, Piek 572
Voutilainen, Atro 255, 505, 541, 605
Vronay, David 304

W

Wagner, Robert A. 1259
Wagner, Susanne 1128, 1131
Wahlster, Wolfgang 286
Waibel, Alex 1188–1189
Wakelin, Martyn F. 1128
Walinski, Jacek 261
Walker, Terry 56, 60, 66, 250, 406, 1099
Wallace, David L. 77, 1076–1077, 1083
Wallis, Sean A. 155, 165, 193, 234, 424,
738–739, 742, 744, 746–748, 750–753,
757, 1011
Walter, Elizabeth 143
Wang, Jianxin 393
Wang, Kefei 438
Wang, Lixun 121
Wang, Wei 702
Wang, Ye-Yi 1188–1189
Ward, Gregory 620
Warren, Martin 417, 1011, 1014
Wärvik, Brita 1054, 1061
Warwick, Susan 689, 1260
Wasow, Thomas A. 599, 989–990
Wasserman, Stanley 334, 339–340
Watanabe, Taro 1188
Waterhouse, Keith 1251
Waterman, Michael S. 1259
Waterman, Scott A. 948
Wattenberg, Martin 293
Watters, Paul A. 896

Watts, Duncan J. 330, 335–336, 339, 341–
346
Watts, Richard J. 1095
Waugh, Linda R. 1014
Waugh, Sam 1082
Wauschkuhn, Oliver 958–959
Way, Andy 1181, 1191
Weaver, Warren 81, 87–88, 90, 566
Webb, Andrew R. 888, 891, 896
Webber, Bonnie 620
Weeninck, David 197
Weerasinghe, Ruvan 1191
Wehrli, Eric 1215, 1224
Weigelt, Ladeana F. 653
Weinert, Regina 1053, 1061
Weinreich, Max 458
Weinreich, Uriel 56, 1027, 1092–1093
Weir, David 961–962
Weischadel, Ralph 545
Weiss, Benjamin 1038
Weiss, Sholom M. 869
Weisser, Martin 620, 624–626
Wells, John C. 645, 649, 653
Wenker, Georg 1127
Werry, Christopher C. 301, 303
West, Michael 10, 115, 140
Westergaard, Marit R. 1134
Wettler, Manfred 557
Whale, Geoffrey Robert 1249
Whistle, Jeremy 119
White, Howard D. 356–357, 360
White, Margaret N. 902
White, Margie 847–848
Wible, David 269, 271
Wichmann, Anne 112, 189, 200–201, 203–
204, 654, 1008, 1055
Widdows, Dominic 351
Widdowson, Henry G. 270, 988
Widmann, Thomas 137
Widmer, Gerhard 858
Wiebe, Janyce 575
Wiegand, Herbert Ernst 132–133, 138
Wierzbicka, Anna 632
Wikberg, Kay 282
Wikborg, Eleanor 1053
Wilhelm, Stephan 123
Wilkerson, Daniel S. 1263–1264
Wilks, Yorick 565, 568, 570–571, 574, 576,
896, 1250
Willems, Dominique 1145
Willett, Peter 896, 1253, 1256–1257, 1265
Williams, Briony 200, 654, 1009

- Williams, Geoffrey 1213, 1245
 Williams, Raymond 730
 Williams, Ronald J. 858
 Williamson, Nigel 1088
 Willis, Dave 114–115
 Willis, Jane 114
 Wilson, Andrew 34, 49, 106, 132, 158, 161–162, 170, 208, 279, 286, 567, 571, 614, 616, 739, 803, 990–991, 1053, 1257
 Wilson, Christine 1082
 Wilson, Deirdre 614, 617, 626
 Wimmer, Gejza 331, 344, 365
 Winski, Richard 628, 643
 Wirdeñäs, Karolina 1034
 Wise, Michael J. 1255, 1259, 1263, 1265
 Witt, Andreas 769
 Witten, Ian 820
 Wittenburg, Peter 394, 497, 645, 660, 674–675
 Wittig, Thomas 349, 351–352
 Wokurek, Wolfgang 1039
 Wolff, Christian 329
 Wolfram, Dietmar 360
 Woliński, Marcin 555
 Woltring, Hermann J. 666
 Wong, Anita 877
 Wong, Hongsing 1191
 Wong, William 1189
 Woods, Anthony 801, 888, 891, 895, 1078
 Woodward, Guy 1039
 Woolls, David 1178, 1263
 Wörner, Kai 771
 Wouters, Paul 328, 365
 Wright, Jerry 1028
 Wright, Joseph 10, 1127
 Wright, Laura 1092
 Wu, Andi 540
 Wu, Dekai 532, 576, 688–689, 694, 699–701, 1183, 1191, 1260
 Wu, Youzheng 538
 Wulff, Stefanie 802, 935–936
 Wunsch, Christian D. 1259
 Wurm, H. Lee 906
 Wynne, Martin 256, 694

X

- Xiao, Richard (Zhonghua) 110, 116, 132, 287, 396, 991, 993, 997–1001, 1003
 Xin, Xu 701
 Xu, Bo 538
 Xue, Nianwen 231, 445, 539

Y

- Yamada, Kenji 698, 1191
 Yamazaki, Takefumi 699
 Yang, Chung-Shu 877
 Yang, Huizhong 430
 Yang, Suying 1000
 Yang, Xiaofeng 582
 Yang, Yiming 1085
 Yardi, M. R. 1072
 Yarnold, Paul R. 876, 888, 891, 895
 Yarowsky, David 564, 567, 570, 573–574, 869, 896
 Yates, Frank 1237
 Yates, Simeon J. 295, 1056
 Yee, Lo Yuen 698
 Yianilos, Peter N. 1259
 Yip, Virginia 1200
 Youd, Nick 188
 Younger, Daniel H. 72
 Yule, George Udny 636, 1071, 1074–1075

Z

- Zampolli, Antonio 34
 Zandbergen, René 820
 Zanettin, Federico 122, 280, 287, 1143, 1167, 1170
 Zaphiris, Panayiotis 293
 Zaslavsky, Arkady 1263–1264
 Zavrel, Jabub 603, 865–866
 Zé Amvela, Etienne 567
 Zelman, Andrés 365–366
 Zeman, Daniel 959
 Zeng, Y. 655
 Zernik, Uri 896
 Zhan, Weidong 438, 446
 Zhang, Hua-Ping 539
 Zhang, ZhenQiu 210, 1281
 Zhao, Jun 538
 Zhou, Liang 1280
 Zifonun, Gisela 927
 Zinsmeister, Heike 608–609, 766, 1244
 Zipf, George Kingsley 148, 331, 803–804, 813, 816, 820, 861, 1026, 1207
 Zitt, Michel 356, 359
 Zitzen, Michaela 293, 296
 Zlatic, Vinko 350, 370, 372–373
 Zobel, Justin 1255, 1264
 Zorzi, Daniela 119
 Zuraidah, Mohd Don 554
 Zwicky, Arnold M. 554

Index of corpora and repositories

A

- A Representative Corpus of Historical English Registers 59, 64, 242–243, 245, 247–248, 250, 402–403, 1095, 1102
Aboriginal Studies Electronic Data Archive 477
Academia Sinica Balanced Corpus 443
Academia Sinica Tagged Corpus of Early Mandarin Chinese 444
Academic Corpus 419–420
ACE *see* Australian Corpus of English
ACIP – Thousand Books of Wisdom 474
ACT corpus of Old Church Slavonic 245
AGD *see* Archiv für Gesprochenes Deutsch
AILLA *see* Archive of Indigenous Languages of Latin America
Air Traffic Control Corpus 191
Aix-MARSEC 197, 205, 408–409
Alaska Native Language Center Archives 478
AMALGAM multi-tagged corpus 502–504, 512–513, 755
American English Microphone corpus 643
American English Meetings Corpus 192
American National Corpus 100, 136, 385–386, 431, 1010
An Crúbadán corpus 318, 461, 465
ANC *see* American National Corpus
Aozora Bunko 476, 479
AP *see* Associated Press corpus
APU Spanish learner corpus 262
Arabic Gigaword 472, 479
ARCHER *see* A Representative Corpus of Historical English Registers
Archiv für Gesprochenes Deutsch (Institut für Deutsche Sprache in Mannheim) 97, 178, 181 *see also* Institut für Deutsche Sprache Resources
Archive for spoken German *see* Archiv für Gesprochenes Deutsch
Archive of Indigenous Languages of Latin America 461, 478
Archive of natural speech in “rare” languages 466
ASBC *see* Academia Sinica Balanced Corpus
ASJ Continuous Speech Corpus of Japanese 477
Associated Press corpus 582
Association for Computing in the Humanities 34
Association for Literary and Linguistic Computing 34

- ATIS corpus of human-machine dialogues 864
Augustana 250
Australian Corpus of English 99, 102, 395–397
Austrian Phonogram recordings 1129

B

- Babel Chinese-English parallel corpora 438
Banca dati dell’ italiano parlato 1018
Banco de datos del español 97
Bank of English 136, 142–143, 158, 181, 234, 310, 394, 401, 736, 980, 994, 1009
BASE *see* British Academic Spoken English corpus
Basque Spoken Corpus 1018
BAWE *see* British Academic Written English corpus
Bensei Database 476
Bergen Corpus of London Teenage Language 99–100, 108, 159, 190, 194, 197, 409–410, 503, 1010, 1013, 1016, 1130
BFSU Chinese-English Parallel Corpus 438
Birmingham Collection of Texts 40–41, 48
Birmingham Corpus 40
BNC *see* British National Corpus
BNC-Baby *see* British National Corpus Baby
BNC-World *see* British National Corpus
Bochumer Mittelhochdeutsch Korpus 249
BoE *see* Bank of English
BOKR *see* Russian National Corpus
Bonner Frühneuhochdeutsches Korpus 243, 245, 254
British Academic Spoken English corpus 418, 1018
British Academic Written English corpus 419
British National Corpus 11, 98, 101–104, 106, 108–110, 114, 118–119, 134–136, 142–144, 155, 158–159, 162, 170, 172, 177–179, 194, 197, 243, 252, 256, 305, 318, 321–322, 351, 384–387, 389–391, 393, 400, 410, 412, 423–424, 431, 482, 487, 502–506, 509, 511, 513, 515, 517–518, 520, 523, 531, 533, 628–629, 634, 659, 687, 713, 726, 729, 731, 733–734, 736, 738, 751, 808–812, 814–818, 844, 908, 934–940, 955, 962–964, 1009–1010, 1013, 1072–

- 1073, 1117–1121, 1130–1131, 1136, 1144, 1163, 1169, 1214, 1218–1220, 1229–1230
– Baby 385, 713–715, 717, 719, 722, 726, 736
British National World Corpus *see* British National Corpus
Brooklyn-Geneva-Amsterdam-Helsinki Corpus of Old English 59, 255, 426
Brown Corpus 1, 34–48, 90–91, 99, 101–102, 108–109, 148, 171–175, 177, 179, 181–184, 227, 242–243, 305, 318, 383, 395–398, 401, 420, 422, 425–426, 441, 448, 502–504, 506–508, 511, 513–514, 534–535, 575, 687, 760–761, 777, 779, 787, 791, 793, 796–798, 806–819, 864, 871, 955, 1010, 1048, 1114–1116, 1121, 1224–1225, 1227, 1230, 1240
Buddhist Scriptures Information Retrieval 475
BUDSIR *see* Buddhist Scriptures Information Retrieval
BYU Corpus of American English 386
- C**
- C.R.A.T.E.R. *see* ITU/CRATER parallel corpus
Callhome corpus 192, 204, 1010
Cambridge and Nottingham Business English Corpus 1011
Cambridge and Nottingham Corpus of Discourse in English 410–411, 415, 1011, 1013, 1015–1016
Cambridge Learner Corpus 261, 429
Canada's Defense and Civil Institute of Environmental Medicine corpus 195
Canadian Hansards 287, 432, 574, 686, 688, 690, 695, 697, 1141, 1163, 1177–1178, 1184
CANBEC *see* Cambridge and Nottingham Business English Corpus
CANCODE *see* Cambridge and Nottingham Corpus of Discourse in English
Canterbury Tales Project 242
CCG bank *see* Combinatory Categorial Grammar Bank
CDLI *see* Cuneiform Digital Library Initiative
CED *see* Corpus of English Dialogues 1560–1760
CEEC *see* Corpus of Early English Correspondence
CEECE *see* Corpus of Early English Correspondence Extension
CEECS *see* Corpus of Early English Sampler
CEG *see* Cronfa Electroneg o Gymraeg
CELEX 553, 678
CELT *see* Corpus of Electronic Texts
Century of Prose Corpus 242, 244, 246, 251–252
CETEMPúblico corpus 440
CGN *see* Spoken Dutch Corpus
Charlotte Narratives 1010
Chemnitz English-German Translation Corpus 287, 1146–1147, 1149
Child Language Data Exchange System 427, 431, 467–468, 499, 677–678, 681, 1197, 1199–1200, 1202–1203, 1205–1209
CHILDES *see* Child Language Data Exchange System
Chinese Interlanguage Corpus 431
Chinese Learner English Corpus 430
CHRISTINE Corpus 227, 423–424
Clay Sanskrit Library 473
CLEC *see* Chinese Learner English Corpus
CLUVI Parallel Corpus 434
– CONSUMER 435
– FEGA 435
– LEGA 435
– LEGE-BI 435
– LOGALIZA 435
– TECTRA 435
– UNESCO 435
CNC *see* Czech National Corpus
COBUILD Corpus 41, 48, 114, 116, 136, 140, 142, 148, 277, 394, 979, 1009, 1207, 1213, 1217
COLT *see* Bergen Corpus of London Teenage Language
Combinatory Categorial Grammar Bank 70, 232, 762
COMPARA/DISPARA 287
CONCE *see* Corpus of Nineteenth-Century English
COPC *see* Century of Prose Corpus
CORIS *see* Corpus di Italiano Scritto
Corpus de Referencia del Español Actual 136, 400, 1018
Corpus del Español 243, 245, 407
Corpus di Italiano Scritto 174, 177–178, 181, 390–391
Corpus do Português 408
Corpus of Early American English 60

- Corpus of Early English Correspondence 59, 63, 242–246, 248–250, 252, 255, 404–405, 1095, 1097–1100, 1102, 1104
 – Extension 59, 405
 – Sampler 245–246, 404–405
- Corpus of Early English Medical Writing 243, 248, 1095, 1103
- Corpus of Electronic Texts 470
- Corpus of English Dialogues 1560–1760 60, 63, 243–244, 246–248, 250, 406, 1095
- Corpus of Irish English 60, 249
- Corpus of Late Eighteenth-Century Prose 406
- Corpus of Late Modern English Prose 407
- Corpus of Middle English Prose and Verse 60, 255, 407
- Corpus of Nineteenth-Century English 243–245
- Corpus of Professional English 421
- Corpus of Spoken American English *see* Santa Barbara Corpus of Spoken American English
- Corpus of Spoken Hebrew 194
- Corpus of Spoken Professional American-English 100, 420–421, 1011, 1117
- Corpus of Spoken Professional English 159
- Corpus of the Berlin-Brandenburgische Akademie der Wissenschaften 136
- Corpus of Translated Finnish 1163
- Corpus Oral de Referencia del Español Contemporáneo 1018
- Corpus Scriptorum Latinorum 245
- Corpus Search, Management and Analysis System 143, 181, 305, 440 *see also* Institut für Deutsche Sprache Resources
- COSMAS *see* Corpus Search, Management and Analysis System
- CoSy:50-Corpus 295
- CREA *see* Corpus de Referencia del Español Actual
- Croatian National Corpus 394
- CroCo Corpus 1161, 1167
- Cronfa Electroneg o Gymraeg 441
- Cross-Towns Corpus 1038
- CSAE *see* Santa Barbara Corpus of Spoken American English
- CSPA (CSPAЕ) *see* Corpus of Spoken Professional American-English
- CSTBank 1281, 1284
- CTF *see* Corpus of Translated Finnish
- Cuneiform Digital Library Initiative 472
- Czech National Corpus 136, 388, 442, 471
- D**
- Danish Arboretum *see* Danish Dependency Treebank
- Danish Dependency Treebank 231, 234
- Data Novels 476
- Datenbank Gesprochenes Deutsch *see* Archiv für Gesprochenes Deutsch
- Davis-Howes Count of Spoken English 1009
- DCCLT *see* Digital Corpus of Cuneiform Lexical Texts
- DCIEM *see* Canada's Defense and Civil Institute of Environmental Medicine
- DCPSE *see* Diachronic Corpus of Present-Day Spoken English
- Der Digitale Grimm 255
- DeutschDiachronDigital 252
- Deutsches Referenzkorpus (DeReKo) *see* Institut für Deutsche Sprache Resources
- Diachronic Corpus of Present-Day Spoken English 194, 196, 424–425, 1114
- Diccionario de la Lengua Española 97
- Dictionary of Old English Corpus (in Electronic Form) 60, 245–246, 255, 403–404, 426
- Digital Corpus of Cuneiform Lexical Texts 472
- Digital Wenker Atlas 1127
- Digitales Wörterbuch der deutschen Sprache 135–136, 146, 181, 305, 391–392, 1111
- DiWA *see* Digital Wenker Atlas
- DoBeS *see* Dokumentation bedrohter Sprachen
- Document Understanding Conferences Corpora 1279–1283
- DOEC *see* Dictionary of Old English Corpus (in Electronic Form)
- Dokumentation bedrohter Sprachen 194, 464–466
- Dortmund Chat Corpus 296
- DUC *see* Document Understanding Conferences Corpora
- Düsseldorf CMC Corpus 296
- Dutch Polyphone corpus 643
- Dutch Spoken Corpus *see* Spoken Dutch Corpus
- DWDS *see* Digitales Wörterbuch der deutschen Sprache
- E**
- Early English Books Online 246, 404, 1096
- Early Modern English Dictionary Database 255

- Early Modern English Medical Texts 59
ECCO *see* Eighteenth Century Collections Online
ECI/MCI *see* European Corpus Initiative Multilingual Corpus I
EDR Japanese Corpus 476
EEBO *see* Early English Books Online
Eighteenth Century Collections Online 404
ELAR *see* Endangered Language Archive
ELDA *see* Evaluation and Language Resources Distribution Agency
Electronic Text Corpus of Sumerian Literature 460–461, 472
elexiko 135
ELRA *see* European Language Resources Association
ELSNET *see* European Network in Language and Speech
EMEMLT *see* Early Modern English Medical Texts
EMILLE *see* Enabling Minority Language Engineering Corpus
Enabling Minority Language Engineering Corpus 437, 473, 516–517, 701
Endangered Language Archive 466
English Gigaword 1252
English-Chinese Parallel Corpus 287, 1282
English-Italian Translational Corpus 287
English-Norwegian Parallel Corpus 122, 281–285, 287, 432–433, 1163
English-Swedish Parallel Corpus 277, 285, 287, 433, 1163
ENPC *see* English-Norwegian Parallel Corpus
Enron Email Dataset 295
ESPC *see* English-Swedish Parallel Corpus
ET 10/63 parallel corpus 434
ETCSL *see* Electronic Text Corpus of Sumerian Literature
ETL Spoken Dialog Corpus 477
EU Translation Corpus 1150, 1153
Europarl *see* OPUS Corpus, Europarl
EuroParl Corpus 97, 287, 439, 810–812, 1163
European Corpus Initiative Multilingual Corpus I 435, 465, 1163
European Language Resources Association 396, 434, 436–437, 446–447, 461, 464–465, 864, 1163
European Network in Language and Speech 435, 446–447
European Speech Database Project 461, 463, 465, 468, 470–471, 473, 643–644, 646, 648–649, 658
EuroWordNet 571
Eus3LB Corpus of Basque 231
EVA Corpus of Norwegian School English 262
Evaluation and Language Resources Distribution Agency 447, 1010
- F**
- FALKO* *see* Fehlerannotiertes Lernerkorpus Fehlerannotiertes Lernerkorpus 117
FIDA Corpus of Slovenian 442
Finnish-English Contrastive Corpus 287
FLOB *see* Freiburg-LOB Corpus of British English
FrameNet 762, 962, 965, 967
Frantext 243, 247, 394, 959
FRED *see* Freiburg English Dialect corpus
Freiburg English Dialect corpus 1128, 1136
Freiburg-Brown Corpus of American English 99, 101–102, 108–109, 173, 182, 395–397, 401, 1001, 1003, 1114–1116
Freiburg-LOB Corpus of British English 99, 101–102, 108, 173, 182, 395–397, 401, 1001, 1003, 1114–1115
French Linguistic Development Corpus 431
French Progression Corpus 431
French-German Parallel Corpus 281
FROWN *see* Freiburg-Brown Corpus of American English
- G**
- GeFre PaC* *see* French-German Parallel Corpus
GEPCOLT *see* German-English Parallel Corpus of Literary Texts
German-English Parallel Corpus of Literary Texts 280
German-Swedish IRC-Corpus 295
Global English Monitor Corpus 395
Göteborg Spoken Language Corpus 1034
Gothenburg Corpus 760–761
GSLC *see* Göteborg Spoken Language Corpus
Gymnasisters språk- och musikvärldar (GSM) 1034–1036

H

- Hansards *see* Canadian Hansards
 HC *see* Helsinki Corpus of English Texts
 HCOS *see* Helsinki Corpus of Older Scots
 HD *see* Helsinki Corpus of British English Dialects
 Hellenic National Corpus 390
 Hellenic National Treebank of Greek 232
 Helsinki Corpus of British English Dialects 1016, 1128
 Helsinki Corpus of English Texts 42–43, 59–60, 62–63, 242–250, 252–253, 255, 401–402, 407, 426, 1061, 1095–1097, 1101–1102, 1110, 1113
 Helsinki Corpus of Older Scots 60, 248–249, 401, 1095
 HKCSE *see* Hong Kong Corpus of Spoken English
 HKUST *see* Hong Kong University of Science and Technology Learner Corpus
 HNC *see* Hungarian National Corpus
 Hong Kong Corpus of Spoken English 416–417, 1011, 1013
 Hong Kong Parallel Text 438–439
 Hong Kong University of Science and Technology Learner Corpus 262, 429–430
 Hungarian National Corpus 389, 391
 Hypnotic corpus 296

I

- IBM Paris Treebank 230
 ICAME *see* International Computer Archive of Modern and Medieval English
 ICAMET *see* Innsbruck Computer Archive of Middle English Texts
 ICE *see* International Corpus of English
 ICE-East Africa *see* International Corpus of English, East Africa
 ICE-GB *see* International Corpus of English, Great Britain
 ICE-New Zealand *see* International Corpus of English, New Zealand
 ICLE *see* International Corpus of Learner English
 ICSI Meetings Corpus *see* International Computer Science Institute Meetings Corpus
 IDS *see* Institut für Deutsche Sprache Resources

- IJS-ELAN Slovene-English Parallel Corpus 434, 1165
 Indogram 296, 423
 Industrial Parsing of Software Manuals corpus 503
 INL *see* Instituut voor Nederlandse Lexicologie Resources
 Innsbruck Computer Archive of Middle English Texts 243–244, 246, 248, 252, 406
 Institut für Deutsche Sprache Resources 143, 172, 178, 181, 440, 1128–1129 *see also* Archiv für Gesprochenes Deutsch
 Institute for Dutch Lexicology 97, 440
 Institute for Dutch Lexicology Resources *see* Instituut voor Nederlandse Lexicologie Resources
 Instituut voor Nederlandse Lexicologie Resources 277–278
 International Computer Archive of Modern and Medieval English 34, 38–39, 99, 181, 305, 396, 402–406, 408–410, 415, 422, 428, 446–447, 502, 1092, 1128, 1160
 International Computer Science Institute Meetings Corpus 190
 International Corpus of English 42, 99, 101, 155, 161, 170, 175, 193, 197, 398–399, 424, 502–503, 511, 513, 739–741, 747, 752–754, 1009, 1011
 – East Africa 398, 1011
 – Great Britain 155, 165, 192–194, 196–197, 203–204, 227, 231, 398, 421, 424–425, 743, 749–751, 753–754, 843, 847, 943–945, 947, 1114, 1116
 – New Zealand 398, 1011
 International Corpus of Learner English 117, 160–161, 261–262, 265, 427–428, 1056, 1061–1062
 International Organization for Standardization 145, 490, 660
 International Telecommunications Corpus 696
 INTERSECT Corpus 287
 Intonation(al) Variation in English (Corpus) 193–194, 197–198, 203–204, 411
 IPI PAN Corpus of Polish 387
 IPSM *see* Industrial Parsing of Software Manuals corpus
 ISO *see* International Organization for Standardization
 Italian Syntactic-Semantic Treebank 231–232

ITU/CRATER parallel corpus 434, 1144
 IViE Intonation(al) Variation in English
 (Corpus) *see* Intonational Variation in English (Corpus)

J

Janus Pannonius University learner corpus 431
 Japanese EFL Learner corpus 430
 Japanese Map Task corpus 199
 JEFLL *see* Japanese EFL Learner corpus
 JPU *see* Janus Pannonius University learner corpus
 JRC-ACQUIS Multilingual Parallel Corpus 434

K

KEMPE *see* Korpus of Early Modern Playtexts in English
 KIAP (Kulturell Identitet i Akademisk Prosa) 1056
 Kids' Audio Speech Corpus 1018
 KNC *see* Korean National Corpus
 Kolhapur Corpus of Indian English 42, 173, 395–397
 Korean National Corpus 393–394
 Korpus 2000 (for Danish) 135–136, 146–147, 394
 Korpus deutschsprachiger Newsgroups 295
 Korpus of Early Modern Playtexts in English 407
 Kyoto Text Corpus 477

L

L-CIE *see* Limerick Corpus of Irish English
 La Banque de Textes du LASLA 242
 la Repubblica corpus 810–812, 814–816, 818
 LACITO *see* Archive of natural speech in “rare” languages
 Lampeter Corpus of Early Modern English Tracts 60, 243–245, 247–248, 251, 253–254, 403, 1095
 Lancaster 1931 256, 396, 397
 Lancaster Anaphoric Treebank 582

Lancaster Corpus of Mandarin Chinese 173, 396–397, 1001, 1003
 Lancaster *Newsbook Corpus* 43, 252, 1252–1253
 Lancaster Parsed Corpus 173, 229, 422, 760
 Lancaster/IBM Spoken English Corpus 408, 502–503, 654, 955
 Lancaster-Leeds Treebank 421
 Lancaster-Oslo/Bergen corpus 35–36, 38–42, 45–48, 90–91, 99, 101–102, 108–109, 148, 171, 173, 182, 185, 242, 395–397, 401, 420–422, 441, 502–504, 506–511, 513–514, 516, 518, 709, 736, 760, 787, 955, 1048, 1114–1116
 LC *see* Lampeter Corpus of Early Modern English Tracts
 LCMC *see* Lancaster Corpus of Mandarin Chinese
 LDC *see* Linguistic Data Consortium
 Limerick Corpus of Irish English 415–416, 1011, 1013
 LINDSEI *see* Louvain International Database of Spoken English Interlanguage
 LinGO Redwood Treebanks 234
 Linguistic Atlas of Early Middle English 249
 Linguistic Atlas of Late Mediaeval English 1098
 Linguistic Atlas of Older Scots 249
 Linguistic Data Consortium 49, 386, 413–414, 425–426, 432, 435, 438–440, 443, 445–447, 464, 466, 658, 682, 864, 920, 1010, 1252, 1279, 1281–1283
 Linguistic Variation in Chinese Speech Communities Corpus 400–401
 LIVAC Corpus *see* Linguistic Variation in Chinese Speech Communities Corpus
 Lívð Tekstðod 471
 LLC *see* London-Lund Corpus
 LOB *see* Lancaster-Oslo/Bergen corpus
 LOCNESS *see* Louvain Corpus of Native English Essays
 London-Lund Corpus 12, 35, 39–40, 47–48, 99, 104–105, 110, 194, 196, 200, 204, 408, 423, 425, 502–503, 509, 513–514, 589, 654, 659, 760, 831, 1009–1010, 1014, 1016, 1049, 1054, 1114
 Longman British Spoken Corpus 412
 Longman Learners’ Corpus 261, 429
 Longman Spoken American Corpus 412, 1009
 Longman Written American Corpus 400
 Longman/Lancaster Corpus 398, 400

- Louvain Corpus of Native English Essays
160, 427–428
- Louvain International Database of Spoken English Interlanguage 428, 1012
- LPC *see* Lancaster Parsed Corpus
- LUCY Corpus 423–424
- M**
- Machine Readable Spoken English Corpus 408–409
- Macquarie corpus *see* Australian Corpus of English
- Macquarie University Corpus of Australian English 173
- MAELC *see* Multimedia Adult ESL Learner Corpus
- Malay Concordance Project 476
- Map Task Corpus 189, 191, 193, 199–200
- MARSEC *see* Machine Readable Spoken English Corpus
- Maya Hieroglyphic Database 478
- MCLC *see* Modern Chinese Language Corpus
- MED *see* Middle English Dictionary
- Medieval Nordic Text Archive 487
- Medline 582
- Meertens Instituut Resources 97
- MEMEM *see* Michigan Early Modern English Materials
- MEMT *see* Middle English Medical Texts
- Menota *see* Medieval Nordic Text Archive
- Message Understanding Conference corpora 85, 580, 582, 595
- METER corpus 1252–1253, 1265
- METU Treebank of Turkish 231
- MICASE *see* Michigan Corpus of Academic Spoken English
- Michigan Corpus of Academic Spoken English 114, 163–164, 417, 418, 1011–1012, 1018, 1062
- Michigan Corpus of Upper-level Student Papers 417–419
- Michigan Early Modern English Materials 407
- MICUSP *see* Michigan Corpus of Upper-level Student Papers
- Middle English Collection 407
- Middle English Compendium 60, 255, 1096
- Middle English Dictionary 60–62, 255, 1096
- Middle English Medical Texts 59

- MidEng *see* Middle English Collection
- Mittelhochdeutsche Begriffsdatenbank 245, 252, 255
- MLCC *see* Multilingual Corpora for Cooperation
- Modern Chinese Language Corpus 392–393
- Modern Tamil Prose Tagged Corpus 474
- MUC *see* Message Understanding Conference corpora
- MULTEXT *see* Multilingual Tools and Corpora
- Multilingual Corpora for Cooperation 436–437, 1163
- Multilingual Learner Corpus 262, 435–436
- Multilingual Tools and Corpora 435, 489, 1144
- Multimedia Adult ESL Learner Corpus 1018

N

- National Corpus of Irish 394
- NECTE *see* Newcastle Electronic Corpus of Tyneside English
- NEGRA 231, 234, 605, 607, 766
- Netscan Usenet database 295
- Network of Eighteenth-century English Texts 1100
- New York Times Newswire 955, 1279
- Newcastle Electronic Corpus of Tyneside English 1128
- Newdigate Newsletters 243
- NICT Japanese Learner English Corpus 430
- NomBank 761
- Nordic Teenage Language Project 97

O

- OED *see* Oxford English Dictionary
- OGI Corpus *see* Oregon Graduate Institute Multilingual Corpus
- OLAC *see* Open Language Archives Community
- Old English Corpus *see* Dictionary of Old English Corpus (in Electronic Form)
- OMC *see* Oslo Multilingual Corpus
- Open Language Archives Community 461, 464, 491, 494–495, 497, 499, 682–683, 770, 772, 920
- OpenSubtitles *see* OPUS Corpus, OpenSubtitles

- OPUS Corpus 287, 439
 – Europarl 97, 287, 439, 810–812, 1163
 – OpenSubtitles 439
- Oral Vocabulary of the Australian Worker 1009
- ORCHID corpus 476
- Oregon Graduate Institute Multilingual Corpus 466
- Oslo Corpus of Tagged Norwegian Texts 1134, 1144
- Oslo Multilingual Corpus 277, 285, 433, 1143–1144, 1161
- OTA *see* Oxford Text Archive
- OVAW *see* Oral Vocabulary of the Australian Worker
- Oxford English Dictionary 8–9, 35, 43, 60, 135, 148, 180, 255–256, 909–910, 1039, 1102, 1115–1116, 1121
- Oxford Text Archive 35, 180, 402, 405, 407, 426, 431, 446–447, 467
- P**
- Pacific and Regional Archive for Digital Sources in Endangered Cultures 477
- ParaConc 283, 287, 736, 1143, 1167, 1178
- PARADISEC *see* Pacific and Regional Archive for Digital Sources in Endangered Cultures
- PARC 700 Dependency Bank 763
- PAROLE 391, 436, 468, 470, 489, 1034–1037
 – Corpas Náisiúnta na Gaeilge 470
- Partially Parsed Corpus of Medieval Portuguese 227
- PCC *see* Potsdam Commentary Corpus
- PDT *see* Prague Dependency Treebank
- Pedant Corpus 277
- PELCRA Reference Corpus of Polish 387, 431
- Penn Arabic Treebank 231
- Penn Chinese Treebank 231, 444, 446
- Penn Discourse TreeBank 232, 762, 769
- Penn Korean Treebank 231
- Penn Parsed Corpus of Modern British English 59
- Penn Treebank 84, 227, 229–232, 234–235, 237, 425–426, 502–503, 508, 513, 540, 542, 548, 582, 600, 602, 604, 752–753, 760–762, 764–765, 767–770, 857, 864, 871, 956–957, 962, 1281
- Penn-Helsinki Corpus of Middle English 1096
- Penn-Helsinki Parsed Corpus of Early Modern English 59, 255
- Penn-Helsinki Parsed Corpus of Middle English 59, 227, 254–255, 426
- PERSEUS Digital Library 468
- PKU-CCL-Corpus of Modern and Ancient Chinese 446
- Plato’s ‘Republic’ corpus 287
- PMK *see* Prague Spoken Corpus
- Polish Learner English Corpus 431
- Polish National Corpus 385, 387
- Polytechnic of Wales corpus 427, 502–503, 506, 511, 513
- Pongal-2000 Project 474
- Português Falado – Documentos Autênticos 1017
- Post-Observation Trainer-Trainee Interactions 1011
- Potsdam Commentary Corpus 763, 765
- POTTI *see* Post-Observation Trainer-Trainee Interactions
- PoW *see* Polytechnic of Wales corpus
- Prague Czech-English Dependency Treebank 227, 230, 763, 959, 1191
- Prague Dependency Treebank 230–232, 234–235, 442–444, 763, 768, 959, 1281
- Prague Spoken Corpus 388–389
- Project Gutenberg 180, 354
- Project Madurai 474
- PropBank *see* Proposition Bank
- Proposition Bank 232, 761–762, 765, 768–770, 864, 962–963, 965
- PWN Corpus of Polish 387
- R**
- RAT *see* Reading Academic Text Corpus
- Reading Academic Text Corpus 419
- Reading Emotion Corpus 201
- Renaissance Electronic Texts 254
- RET *see* Renaissance Electronic Texts
- Reuters 582, 1252
- Rhetorical Structure Discourse Treebank 762
- RNC *see* Russian National Corpus
- Rosetta-ALL Language Archive 467, 480
- RST Discourse Treebank 232
- Russian National Corpus 385, 389–390
- Russian Reference Corpus *see* Russian National Corpus
- RWC Text Database 476

S

- Saarbrücken Corpus of Spoken English 413
Saarbrücken Lexical Semantics Annotation Project 147
SALSA *see* Saarbrücken Lexical Semantics Annotation Project
Santa Barbara Corpus of Spoken American English 159, 194, 200, 202, 413, 1055, 1117
SBCSAE *see* Santa Barbara Corpus of Spoken American English
SCCD *see* Spoken Chinese Corpus of Situated Discourse
SCoSE *see* Saarbrücken Corpus of Spoken English
SCOTS *see* Scottish Corpus of Text and Speech
Scottish Corpus of Text and Speech 441
SCRIBE *see* Spoken Corpus Recordings In British English
SEC *see* Lancaster/IBM Spoken English Corpus
SEC *see* Spoken English Corpus
SED *see* Survey of English Dialects
SemCor corpus 575
SEMiSUSANNE 423
SEU *see* Survey of English Usage (Corpus)
Sinica Treebank of Chinese 232, 444
Sketch Engine 144–145
Slovak National Corpus 392
SMARTKOM project 670
SNK *see* Slovak National Corpus
SpamAssassin Public Corpus 295
SpeechDat 461, 465, 468, 470, 473, 643–644, 646, 648–649, 658 *see also* European Speech Database Project
Spoken Chinese Corpus of Situated Discourse 445–446
Spoken Corpus of British English *see* Spoken Corpus Recordings In British English
Spoken Corpus of the Survey of English Dialects 411
Spoken Corpus Recordings In British English 192, 502
Spoken Dutch Corpus 193, 227, 654–656, 659, 678–679, 682, 860, 862, 864, 866
Spoken English Corpus 192, 197, 200, 204, 408–409, 502–503, 511, 654, 1009
SSE *see* Survey of Spoken English
Standard Speaking Test Corpus *see* NICT Japanese Learner English Corpus

- STRAND bilingual database 467
Stuttgart-Tübingen-Tagset 924
SUMMAC corpora 1278–1280
SUMMBank 1282
Survey of English Dialects 411, 1128, 1133
Survey of English Usage (Corpus) 1, 10–12, 35, 39–40, 48, 105, 148, 170, 175, 177, 194, 408, 1009, 1049, 1114
Survey of Spoken English 39, 408, 1009
SUSANNE Corpus 227–229, 231, 422–424, 760–761, 955
SWB *see* Switchboard corpus
Swedish Talbanken 413–414, 468, 762, 1034–1036
Switchboard corpus 192, 203, 227, 232, 413–414, 423, 490, 623, 652, 655, 864, 1010
SYN2000 388–389
- T**
- Talbanken corpus *see* Swedish Talbanken
Tamil Digital Corpus 474
TDT *see* Topic Detection and Tracking Corpora
TEC *see* Translation(al) English Corpus
TELC *see* Thai English Learner Corpus
TELEC Student Corpus 271
Telekorp corpus 262
TELRI Research Archive of Computational Tools and Resources 446–447, 1163
Text Retrieval Conference Corpus 85, 1252, 1274, 1278–1279, 1281, 1283
Text Summarisation Challenge Corpora 1282
Textes de Français Ancien 250
Thai Chat Corpus 295
Thai English Learner Corpus 430
Thesaurus Indogermanischer Text- und Sprachmaterialien 250, 469
Thesaurus Linguae Graecae 34, 469
Tibetan & Himalayan Digital Library 474
TiGer dependency bank 763
TiGer treebank 147, 231–232, 235, 609, 674, 763, 768, 924–926, 928–929
TimeBank 762, 765, 769
TIMIT Acoustic Phonetic Continuous Speech Corpus 192, 652
TITUS *see* Thesaurus Indogermanischer Text- und Sprachmaterialien
Topic Detection and Tracking Corpora 1278, 1280–1281

TRACTOR *see* TELRI Research Archive of Computational Tools and Resources
 TRAINS Corpus of Dialogues 191
 Translation(al) English Corpus 280, 1141, 1163, 1165, 1169
 TREC *see* Text Retrieval Conference Corpus
 Treebank II *see* Penn Treebank
 Triptic Corpus 277
 TSC *see* Text Summarisation Challenge Corpora
 TüBa-D/S *see* Tübinger Baumbank des Deutschen/Spontansprache
 TüBa-D/Z *see* Tübinger Baumbank des Deutschen/Zeitungssprache
 Tübingen Partially Parsed Corpus of Written German 930
 Tübingen Treebank of spoken German *see* Tübinger Baumbank des Deutschen/Spontansprache
 Tübingen Treebanks of spoken German, English and Japanese 227
 Tübinger Baumbank des Deutschen/Spontansprache 227, 230
 Tübinger Baumbank des Deutschen/Zeitungssprache 230, 763
 TüPP-D/Z *see* Tübingen Partially Parsed Corpus of Written German
 Turin University Treebank of Italian 231–232
 Tycho Brahe Corpus of Historical Portuguese 252

U

University of Maryland Parallel Corpus 467
 University of Pretoria Research Corpora – Sepedi 472 – Zulu 472
 University of Pretoria Research Corpora 472
 University of Stellenbosch African Speech Technology Project 473, 479
 University of Virginia Chiricahua and Mescalero Apache Texts 478
 University of Virginia (Classical) Japanese Text Initiative 476

UNO *see* Nordic Teenage Language Project
 Uppsala Student English Corpus 430
 USE *see* Uppsala Student English Corpus

V

Voices of Taiga and Tundra (Saint-Petersburg Collections) 471

W

Wall Street Journal Corpus 227, 426, 527, 535, 541–542, 582, 761–762, 864, 871, 923, 955 *see also* Penn Treebank
 War of the Worlds 810–812, 814–815
 WCNZE *see* Wellington Corpus of Written and Spoken New Zealand English
 Wellington Corpus of Written and Spoken New Zealand English 42, 99, 102, 108, 173, 182, 395–397, 414, 420
 Wendekorpus 181
 West Point Arabic Speech Corpus 472, 479
 WK *see* Wendekorpus
 WordNet 85–86, 92, 322, 349, 353–355, 423, 566, 569, 571–575, 591, 766, 960–962, 964, 966–968, 1261
 WSJ *see* Wall Street Journal Corpus

Y

YCOE *see* York-Toronto-Helsinki Parsed Corpus of Old English Prose
 York-Helsinki Parsed Corpus of Old English Poetry 255, 426
 York-Toronto-Helsinki Parsed Corpus of Old English Prose 255, 426, 1096

Z

ZEN *see* Zurich English Newspaper Corpus
 Ziff-Davis Corpus 1275, 1277
 Zurich English Newspaper Corpus 59, 243–244, 246, 248, 251, 405–406, 1095

Subject index

A

- active learning *see* machine learning, active learning
- age *see* social variables, age
- alignment 202, 204, 210, 217–218, 279, 282–283, 645, 651–652, 656, 665, 1143, 1164, 1182
 - sentence 686, 688–695, 1184, 1259–1261
 - tree 699–700
 - word 686, 695–698, 1185
- ALPAC report 77, 88
- alternation *see* diathesis alternation
- anaphora annotation *see* annotation, anaphora
- anaphora resolution *see* annotation, anaphora
- annotation 45, 82–83, 134, 143, 208–209, 218, 252, 303, 294, 296–297, 301–302, 490, 760, 827, 855, 859–870, 922, 955, 995, 1145, 1164–1165
 - anaphora 579–598
 - aspectual information 762
 - automatic 199, 219, 233, 650, 1276
 - chunking 230, 235, 302, 604–605, 609–610
 - consistency 233–234, 547, 593, 763–767, 869 *see also* evaluation
 - constituency 229–231
 - coreference 579–598
 - dependency structure 232, 352, 603, 763, 1191
 - disfluency 212, 232
 - error 266, 1202–1203
 - functional 229–232, 609
 - header 252–253, 394, 495–496, 628, 683, 1165–1166
 - hierarchical 605, 675
 - information structure 763, 767
 - inline 713, 769
 - manual 200, 203, 219, 233, 547, 651, 680–681, 1273–1275
 - meta-data 216, 299 *see also* meta-data
 - morpho-syntactic 228, 490, 1202
 - multi-tier 228, 266, 489, 766–769, 1167, 1205 *see also* annotation, standoff
 - part-of-speech 43, 45, 82, 89–90, 145, 228, 254, 265, 282, 302, 501–502, 504–505, 507, 510, 527, 539–548, 856, 859, 866–870, 1165
 - phonetic 197, 199, 202
 - phonological 197, 199, 642–663, 859
 - phrase structure 231

- pragmatic 491, 613–642
- prosodic 196, 199–200, 202–203, 209, 644, 653–654, 859, 1014
- semantic 229, 232, 255, 564–598, 762, 962
- standoff 487, 489–491, 678–679, 716, 768–769
- syntactic 46, 228–233, 598–613, 760, 859, 922, 1071 *see also* parsing
- annotation graph 674, 767
- annotation guidelines 233, 547, 580, 594, 655, 764–765, 770, 923
- annotation standard 235, 306, 484–501, 510, 514, 520, 767
- annotation tools *see* tools, annotation
- anonymisation 157, 196–197, 300
- ANOVA *see* statistical test, ANOVA
- antonymy 974, 976–978
- aspect 996–1004
 - Mandarin Chinese 999–1001
- ASR *see* speech recognition, automatic association
 - Bayesian 80, 544, 752, 780, 800, 1077
 - statistical 941–942, 1215, 1293–1295
- association measure 143, 949, 1215–1218, 1236, 1243, 1298
 - chi-squared 1085, 1228, 1235, 1237
 - Dice coefficient 693, 694, 1182, 1234–1235, 1237, 1257
 - log-likelihood 697, 1235, 1237
 - mutual information 286, 696–698, 1085, 1218, 1226, 1228–1230, 1237, 1241, 1297–1298
 - odds ratio 793
 - simple-ll 1228, 1230, 1237
 - t-score 1227–1228, 1230, 1237
 - z-score 789, 791, 1078, 1227–1229, 1237, 1242
- audio recording *see* recording, audio
- authorship attribution *see* stylometry
- automatic annotation *see* annotation, automatic
- automatic speech recognition *see* speech recognition, automatic
- automatic summarisation *see* summarisation, automatic

B

- balance *see* corpus design, balance
- Bayesian statistics *see* association, Bayesian
- bilingual dictionary *see* dictionary, bilingual

binomial distribution *see* statistical distribution, binomial
 binomial test *see* statistical test, binomial
 blog corpus *see* corpus, blog
 Brill tagger *see* tagger, Brill

C

CALL *see* computer-assisted language learning
 CCG *see* grammar, combinatory categorial
 central tendency *see* statistical distribution, central tendency
 CES *see* Corpus Encoding Standard
 CFG *see* grammar, context-free
 child language corpus *see* corpus
 chi-squared *see* association measure, chi-squared *see* statistical distribution, chi-squared *see* statistical test, chi-squared
 Chomsky hierarchy 71, 78
 chunking *see* annotation, chunking
 CIA *see* contrastive interlanguage analysis
 citation graph 356–361
 classification *see* machine learning, classification
 CLAWS tagger *see* tagger, CLAWS
 cluster analysis *see* machine learning, clustering
 clustering *see* machine learning, clustering
 CMC *see* corpus, computer-mediated communication
 collexeme 942–947, 1145
 colligation 716, 938, 996, 980, 1057–1058, 1145
 collocation 49, 116, 165, 286, 351, 353, 570, 725, 979–980, 994, 1057–1059, 1066, 1071, 1101, 1145, 1212–1248, 1291, 1295
 collostructional analysis 940
 combinatory categorial grammar 70, 74, 232, 762
 comparable corpus *see* corpus, comparable
 competence vs. performance 25, 27, 189, 777, 988–989, 991
 complementizer 1133, 1292–1295
 compositionality 906, 916, 997–998, 1213–1214
 computational linguistics 68–96, 505, 859, 863, 870–871, 875, 1063, 1213–1214, 1224
 – history 87–96

computational pragmatics *see* pragmatics, computational
 computer-assisted language learning 119, 271
 computer-mediated communication *see* corpus, computer-mediated communication
 concordance 1–3, 33, 119, 121, 123, 138, 143–144, 180, 267–268, 316–317, 706–737, 1143, 1177–1178
 confidence interval *see* statistical test, confidence interval
 constituency annotation *see* annotation, constituency
 constraint-based grammar *see* grammar, constraint-based
 construction grammar 941, 1050
 context 618–619, 628–630, 921, 1044–1045, 1049, 1057
 context-free grammar *see* grammar, context-free
 context-sensitive grammar *see* grammar, context-sensitive
 contingency table 696, 792–795, 800, 949, 1085, 1182, 1231–1234, 1243, 1245
 contrastive analysis 278, 285–286, 1001–1004, 1140–1159, 1170
 contrastive interlanguage analysis 117, 267–269, 1141
 contrastive linguistics 276, 280, 286, 1140–1159, 1160
 conversation analysis 6, 159, 202, 218, 293, 618–619, 1014, 1043, 1055 *see also* discourse analysis
 cooccurrence 351, 696, 725, 824–825, 830, 1214–1215, 1218, 1220–1242, 1295
 copyright *see* corpus design, copyright issues
 coreference annotation *see* annotation, coreference
 coreference resolution *see* annotation, coreference
 corpus
 – blog 295, 297, 313, 328, 366–368
 – child language 1199–1200
 – comparable 122, 137, 276–277, 432, 685, 698, 1141, 1162
 – computer-mediated communication 213, 262–263, 292–309, 335, 365, 1061–1062
 – dialect 1127–1137
 – first-generation 35–40, 101, 171–173
 – historical 53–68, 157, 242–259, 401–408, 426, 401–408, 1091–1109
 – historical, spoken language 60, 247

- learner 117, 120–121, 123, 159, 259–275, 426–432, 992, 1017, 1168
- monitor 42, 157, 174, 394–395
- multilingual 276, 319, 432–439, 685–705, 1141, 1144, 1149, 1151–1154, 1162
- multimodal 160, 207–225, 459, 664–685
- parallel 97, 118, 121, 137, 275–292, 319, 432, 574, 685, 687, 698, 1141, 1143, 1145, 1151, 1154, 1161–1167, 1178, 1184, 1282
- pre-electronic 1–14
- reference 154, 170–172, 174, 178–182, 383–394, 733, 1144
- speech 39, 48, 97, 157–160, 169, 187–207, 227, 408–417, 458–459, 489, 642–664, 1008, 1024–1026, 1029–1036, 1060
- text 48, 135, 157, 169–187, 208, 227, 243, 305, 331, 334, 383–408, 458, 459, 602, 607
- Web 108, 138, 158, 328, 335, 922, 1169, 1252
- corpus compilation
 - ethics 11, 157, 195, 300–301, 463, 1065
- corpus design 11, 36, 134, 136, 154–168, 169, 175–176, 189, 190–195, 213, 226, 242–243, 248, 250, 261–265, 282–284, 297, 1012, 1049, 1065, 1136, 1144, 1161–1164, 1287–1289
- balance 160, 163–166, 214, 246, 306, 311, 318, 463, 798
- best practice 462–464, 499, 642–644, 660, 760–761, 771–772
- copyright issues 38, 157, 180, 245–246, 282, 319, 463, 659, 1250
- representativeness 40, 160–162, 164, 166, 172, 244, 246–247, 798, 991–992, 994, 1011, 1013, 1049, 1095, 1287–1289
- size 11, 40, 135, 160, 162, 164–166, 227, 245, 261, 1012, 1136, 1288–1289
- whole text vs. sample 12, 165, 178, 246, 1289
- Corpus Encoding Standard 386, 389, 485, 488–489, 660, 675, 1167
- corpus size *see* corpus design, size
- corpus-based approach 143, 992–996, 999–1001, 1005, 1046–1047, 1051, 1066, 1283
- corpus-driven approach 114, 138–139, 143, 711–712, 973, 980, 990, 993–996, 999, 1005, 1046–1048, 1051, 1058–1059, 1062, 1066, 1283
- cuneiform 472
- cusum charts 1078–1080

D

- data sparseness 818–819, 880–888, 910, 964
- data-driven learning 118–123
- DCMI *see* Dublin Core Metadata Initiative
- DDL *see* data-driven learning
- decision tree *see* machine learning, decision tree
- dependency grammar 70, 230, 443, 603–604, 739, 741–743
- dependency structure annotation *see* annotation, dependency structure
- descriptive statistics *see* statistics, descriptive
- diachronic corpus *see* corpus, historical
- diachrony 56, 909–910, 1133, 1134 *see also* language change
- dialect 10, 57, 100, 249, 1016, 1126–1140
- dialect corpus *see* corpus, dialect
- dialectology 896, 1126–1140
- dialogue *see* interaction, dialogue
- diathesis alternation 954, 963–964
- Dice coefficient *see* association measure, Dice coefficient
- dictionary 6–10, 41, 43, 48, 60–62, 131–153, 255–256, 270, 276, 280, 482, 503, 565, 569–570, 909, 958–959, 1102, 1121, 1127 *see also* lexicography
- bilingual/multilingual 137, 255, 276, 280, 283, 698, 1145
- electronic 135, 180, 317, 569
- learner 116, 135, 140, 146, 270, 1217
- machine-readable 43, 133, 145, 180, 569, 571, 572, 1213
- specialized 133, 136–137, 146
- dictionary as corpus 255–256, 403
- digitisation of manuscripts *see* manuscript digitisation
- dimensionality *see* vector space
- dimensions of variation *see* variation, dimensions of
- Dirichlet distribution *see* statistical distribution, Dirichlet
- discourse 1017, 1043–1070, 1102–1103, 1149–1151
- discourse analysis 301, 618, 1043–1070
- disfluency 159, 196, 212, 635, 649, 644–646, 648, 1025, 1119
- dispersion *see* statistical distribution, dispersion
- Document Type Definition *see* XML, Document Type Definition
- DTD *see* XML, Document Type Definition

- Dublin Core Metadata Initiative 493–495
 duplicate detection *see* text re-use
- E**
- EA *see* error analysis
 EAGLES *see* Expert Advisory Group on Language Engineering Standards
 effect size *see* statistical test, effect size
 EFL *see* English as a Foreign Language
 electronic dictionary *see* dictionary, electronic
 Electronic Metastructure for Endangered Language Data *see* E-MELD
 ELF *see* English as a Lingua Franca
 elicitation 12, 15, 17, 29–30, 189–190, 459, 631, 669–670, 845, 1016, 1137, 1145
 EM algorithm *see* Expectation Maximisation
 embedded annotation *see* annotation, inline E-MELD 463, 499
 emotional speech *see* speech, emotion in endangered language 194, 458, 462, 464–466, 482, 499
 English as a Foreign Language 115–116, 121, 271 *see also* corpus, learner
 English as a Lingua Franca 260, 1017, 1027
 error analysis 121, 268
 error annotation *see* annotation, error
 error detection in annotation *see* annotation, consistency
 ethics *see* corpus compilation, ethics
 evaluation 75, 84–86, 237, 573–576, 594–595, 608–610, 658, 695, 743, 921–922, 958, 967–968, 1239, 1275–1280
 – gold standard 75, 84–85, 237, 548, 573, 575, 608, 610, 637–638, 687, 764, 922–923, 967–968, 1273, 1275, 1283
 – precision 84, 695, 922, 968
 – recall 84, 695, 922, 968
 Expectation Maximisation 79, 869, 1185–1186
 expected frequency *see* frequency, expected
 experiment design 1288–1291
 Expert Advisory Group on Language Engineering Standards 145, 485, 488, 497, 499, 510, 514–515, 628, 637, 643, 660, 680
 eXtensible Markup Language *see* XML
- F**
- factor analysis 827–834 *see also* multi-dimensional approach
- field-work 17, 20–21, 30, 463, 1128
 film 121, 160, 208, 221, 683, 1027
 finite-state automaton 71–72, 558–559
 finite-state language 19–20
 finite-state techniques 71–72, 558–560, 604–605, 957, 1191
 finite-state transducer 71, 561
 first generation corpus *see* corpus, first generation
 first language acquisition 22, 25, 427, 1008, 1197–1211
 Fisher’s exact test *see* statistical test, Fisher’s exact
 Fisher-Yates exact test *see* statistical test, Fisher-Yates exact
 foreign language acquisition *see* second language acquisition
 FrameNet 762–763, 962
 frequency
 – expected 780–781, 943, 1225–1226, 1231–1234, 1240–1241, 1297
 – observed 780–781, 792, 795, 798, 943, 1225–1226, 1231–1234, 1240–1241, 1297
 frequency normalisation 162, 164, 1299
 frequency spectrum 806–813, 816–817
 FSA *see* finite-state automaton
 functional annotation *see* annotation, functional
 functional linguistics 1130, 1287, 1291
- G**
- Gaussian distribution *see* statistical distribution, normal
 gender *see* social variables, gender
 General Ontology for Linguistic Description 499, 680, 770
 Generalized Phrase Structure Grammar 507
 generative linguistics 14–32, 900, 921, 988, 990–991, 1027, 1134–1135
 genre *see* text type
 gesture 208–212, 215, 217, 220–222, 664, 666, 669–670
 global English 103, 395, 1027
 GOLD *see* General Ontology for Linguistic Description
 gold standard *see* evaluation, gold standard
 GPSG *see* Generalized Phrase Structure Grammar
 grammar 3–4, 15, 17, 19–20, 70, 72–75, 189, 281, 599–601, 933–952, 1014

- constraint-based 46, 601, 739
 - context-free 19, 71–73, 78–79, 230, 599–600, 957, 959
 - context-sensitive 70–72
 - prescriptive 3, 5, 16, 715–716, 1122
 - probabilistic 78, 79, 237, 600–601, 1131–1132
 - probabilistic context-free 600–601
 - symbolic 599–600, 602
 - grammar development *see* grammar writing
 - grammar induction 602–603
 - grammar writing 21, 23, 601–603
 - grammaticalisation 58–59, 61–62, 194, 1010, 1099–1101, 1120, 1133–1134, 1142–2243, 1146–1148
 - grammaticality judgments *see* native speaker intuition
 - graph theory 333, 335–348
 - guidelines for annotation *see* annotation guidelines
- H**
- hapax legomenon 41, 76, 482, 818–819, 902, 905–906, 912, 992, 1026, 1073
 - Head-driven Phrase Structure Grammar 70–71, 73–74, 232, 599, 605, 958
 - header information *see* annotation, header hesitation *see* disfluency
 - Hidden Markov Model 78, 81–82, 532, 539, 541–546, 605, 961, 1188, 1277
 - hierarchical annotation *see* annotation, hierarchical
 - historical corpus *see* corpus, historical
 - historical pragmatics *see* pragmatics, historical
 - homonymy 16, 139, 393
 - HPSG *see* head-driven phrase structure grammar
 - human annotation *see* annotation, manual
 - hyperlinks 298, 330–331, 334, 361–373
 - hyponymy/hypernymy 960, 974, 977–978
 - hypothesis testing *see* statistical test
- I**
- IAA *see* inter-annotator agreement
 - ideographic script *see* tokenisation, ideographic script
 - IMDI *see* ISLE Meta-Data Initiative
- inferential statistics *see* statistics, inferential
 - inflectional morphology *see* morphology, inflection
 - Information Retrieval 85, 315, 317, 324, 329, 552, 566, 580, 884–885, 896, 1072, 1254, 1256
 - information structure annotation *see* annotation, information structure
 - inline annotation *see* annotation, inline interaction
 - dialogue 190–191, 621–624
 - multi-party 161, 192, 628, 630, 645
 - inter-annotator agreement 202, 235, 581, 593–595, 764, 766, 1273, 1276 *see also* evaluation *see also* annotation, consistency
 - inter-coder reliability *see* inter-annotator agreement
 - International Organisation for Standardisation *see* ISO
 - International Phonetic Alphabet 217, 656, 1203
 - International Standards for Language Engineering *see* ISLE
 - internet 138, 158, 292, 294, 296, 298, 312, 314, 334, 361–373, 1027 *see also* corpus, Web *see also* corpus, computer-mediated communication
 - interval scale *see* statistical variable, interval
 - intuition *see* native speaker intuition
 - inverse document frequency *see* TF.IDF
 - IPA *see* international phonetic alphabet
 - ISLE 497–498, 660, 680, 682, 770
 - ISLE Meta-Data Initiative 497–498, 660, 678, 682, 683, 770
 - ISO 145, 235, 385, 485–487, 490–491, 499–500, 660, 674, 676–677, 680, 683, 770
- K**
- KWIC *see* concordance
- L**
- LAF *see* Linguistic Annotation Framework
 - language change 54–58, 66, 77, 109, 243, 245, 306, 910, 1091–1109, 1109–1125, 1161 *see also* corpus, historical *see also* variationism

- recent and ongoing 107, 306, 909–910, 1094, 1109–1125
 - language contact 1021–1043
 - Language for Special Purposes 117, 120, 280, 1168
 - language model *see* machine learning, language model
 - language teaching 112–131, 236, 262–263, 270–271, 276, 288
 - Large Number of Rare Events *see* statistical distribution, LNRE
 - learner corpus *see* corpus, learner
 - learner dictionary *see* dictionary, learner
 - lemmatisation 265, 302, 552–564, 859, 866–870 *see also* stemming *see also* morphological analysis
 - Lexical Functional Grammar 70, 73–74, 230, 232, 507, 599, 605, 762, 957
 - lexical priming 353, 982–985
 - lexicography 6–8, 131–153, 280, 353–354, 505, 565, 567, 607, 686, 818, 1027, 1145, 1154, 1227 *see also* dictionary
 - LFG *see* Lexical Functional Grammar
 - Linguistic Annotation Framework 490, 491, 499, 676, 772
 - LNRE distribution *see* statistical distribution, LNRE
 - log-likelihood *see* association measure, log-likelihood
 - long-term storage *see* sustainability
 - LSP *see* Language for Special Purposes
-
- ## M
- machine learning 81, 237, 332, 507, 537, 570, 602, 855–874, 874–899
 - active learning 868–870
 - classification 295, 535–536, 884, 964–967, 1253
 - clustering 143, 332, 339–342, 846, 890–894, 967, 1074, 1076, 1082
 - decision tree 535, 548, 585, 863
 - supervised 82–84, 856, 542, 546, 967
 - support vector machine 539, 585, 858, 863
 - training 79, 82–84, 506, 863–864
 - unsupervised 82–83, 536, 542, 546, 1261
 - machine translation 77, 81, 85, 87, 276, 324, 566, 580, 687, 698, 756, 1175–1196, 1178, 1217, 1256, 1260
 - example-based 279, 283, 1176–1177, 1181–1182
 - language model 1186–1187
 - statistical 279, 283, 695, 1176–1177, 1183, 1188–1196
 - machine-readable dictionary *see* dictionary, machine-readable
 - magnetic tape *see* recording, magnetic tape
 - manifold *see* topology, manifold
 - manual annotation *see* annotation, manual
 - manuscript digitisation 66, 158, 244–245, 250–254, 487, 1105
 - markedness 141, 857, 906, 974, 1028, 1131
 - MATE 219, 489–490, 587–588, 677, 682 *see also* NITE
 - maximum entropy 535, 539, 1178
 - mean *see* statistical distribution, mean
 - median *see* statistical distribution, median
 - mental lexicon 565, 911–916
 - meta-data 138, 144, 492, 299–300, 493, 495, 726–728, 770 *see also* annotation, meta-data
 - meta-data standard 492–498
 - MI *see* association measure, mutual information
 - Minimalism 70, 72
 - mode *see* statistical distribution, mode
 - monitor corpus *see* corpus, monitor
 - morphological analysis 552–564 *see also* lemmatisation *see also* stemming
 - morphology 19, 552–564, 895, 899–919, 1099–1100
 - inflection 552–564, 1100
 - productivity 899–919, 1026
 - word formation 107, 553–554, 557–559, 561–562, 899–919, 1033–1034, 1036, 1039
 - morpho-syntactic annotation *see* annotation, morpho-syntactic
 - multi-dimensional approach 172, 822–855, 949, 1080
 - multi-document summarisation
 - multi-level annotation *see* annotation, multi-tier
 - Multilevel Annotation, Tools Engineering *see* MATE
 - multilingual corpus *see* corpus, multilingual
 - multimodal corpus *see* corpus, multimodal
 - multi-party interaction *see* interaction, multi-party
 - multi-tier annotation *see* annotation, multi-tier
 - multivariate analysis 874–899, 1080–1082

multivariate test *see* statistical test, multivariate
 multiword expression 513, 516, 519, 522, 529–530, 548, 1214, 1239, 1242 *see also* collocation
 multiword extraction 734–735, 1238–1239
 mutual information *see* association measure, mutual information

N

named-entity recognition 655, 699
 national corpus *see* corpus, reference
 native speaker intuition 6, 15, 18–19, 21–30, 166, 211, 712, 905, 921, 988–1000, 1004–1005, 1015, 1152
 Natural Interactivity Tools Engineering *see* NITE
 natural language processing 77–78, 132–133, 145, 237–238, 324, 505, 552, 565, 580, 764, 858, 859–864, 952, 964–965, 1279 *see also* computational linguistics *see also* machine learning *see also* machine translation
 neologism 140, 322, 394, 905–906, 1121
 neural networks 334, 535, 863, 1082–1083
 newswire text 1250, 1252–1253, 1264–1265
 n-gram 81–82, 536, 543–544, 1187, 1256–1257, 1264–1265
 NITE 219, 489–490, 499, 674, 677, 682
 NLP *see* natural language processing
 nominal scale *see* statistical variable, nominal
 non-parametric test *see* statistical test, non-parametric
 non-randomness 796–799, 1224, 1244
 normal distribution *see* statistical distribution, normal
 normalisation, spelling 65–66, 102, 250, 252

O

observed frequency *see* frequency, observed
 OCR *see* optical character recognition
 odds ratio *see* association measure, odds ratio
 OLAC *see* Open Language Archives Community
 ongoing change *see* language change, recent and ongoing

Open Language Archives Community 491, 494–495, 682–683, 770, 772, 920
 optical character recognition 77, 251, 694
 orality vs. literacy *see* spoken vs. written language
 orthographic transcription *see* transcription, orthographic

P

parallel corpus *see* corpus, parallel
 parameter estimation 79, 542–543
 parametric test *see* statistical test, parametric
 PARSEVAL 237, 608
 parsing 73, 88, 235, 237, 254, 323, 530, 598–613, 1191 *see also* annotation, syntactic
 partial parsing *see* annotation, chunking
 part-of-speech tagging *see* annotation, part-of-speech
 Pattern Grammar 937–938, 940, 951, 994
 PCA *see* principal components analysis
 PCFG *see* grammar, probabilistic context-free
 pedagogy 112–131, 236, 270–271, 324, 1017 *see also* language teaching
 – applications 113–114, 117–118, 120–123
 performance *see* competence vs. performance
 period disambiguation 529–536
 phonetic annotation *see* annotation, phonetic
 phonetic transcription *see* transcription, phonetic
 phonological annotation *see* annotation, phonological
 phonology 21, 71, 101, 200, 895, 1028–1034, 1097, 1127, 1207
 phrase structure annotation *see* annotation, phrase structure
 phraseological unit *see* collocation
 plagiarism 1250, 1253, 1262–1263
 polysemy 139–140, 142, 147, 284, 354–355, 569, 967, 973, 975
 population *see* statistics, population
 pos tagging *see* annotation, part-of-speech
 pragmatic annotation *see* annotation, pragmatic
 pragmatics 54, 204, 613–642, 1016, 1043, 1050, 1055
 – computational 619–621
 – corpus-based 613–642, 632–637
 – historical 56, 1061, 1097–1098

precision *see* evaluation, precision
 pre-electronic corpus *see* corpus, pre-electronic
 prescriptive grammar *see* grammar, prescriptive
 principal components analysis 888, 895, 908–909, 1076, 1081–1082
 probabilistic context-free grammar *see* grammar, probabilistic context-free
 probabilistic grammar *see* grammar, probabilistic
 probability *see* statistics, probability
 productivity *see* morphology, productivity
 prosodic annotation *see* annotation, prosodic
 prosody 12, 39–40, 47–48, 188, 193, 196–205, 209, 211, 653–655, 660, 664, 859–860, 1014, 1016
 – semantic 712, 980–982, 1058
 psycholinguistics 22, 30, 72, 620, 1050, 1063, 1132, 1160
 punctuation 303, 305, 529–536, 647, 687, 804–805, 1166

Q

query *see* search

R

random sampling *see* sample, random
 random variation 778–779, 784, 791, 1225, 1242
 rank frequency profile 806–813
 RDF *see* Resource Description Framework
 recall *see* evaluation, recall
 recent change *see* language change, recent and ongoing
 recording
 – audio 209–210, 216, 650–670
 – magnetic tape 39
 – video 209–210, 215, 665–670
 reference corpus *see* corpus, reference
 register 156, 162, 244, 248, 511, 514, 520, 823–855, 896, 908, 1052, 1060, 1094, 1100, 1102–1103, 1288, 1293, 1295, 1301
 regular expressions 71, 145, 530, 710
 regularisation *see* normalisation, spelling
 representativeness *see* corpus design, representativity

representativity *see* corpus design, representativity
 Resource Description Framework 498, 677
 reusability 759–776
 Rhetorical Structure Theory 762–763, 1281
 RST *see* Rhetorical Structure Theory

S

SAMPA 217, 645, 649, 653
 sample
 – proportional 1288
 – random 778–779, 783, 794, 796–797
 – stratified 1288
 Schema *see* XML, Schema
 search 315–316, 552, 706–737, 1167
 search engine *see* tools, search
 search tools *see* tools, search
 second language acquisition 259–275, 1017–1018 *see also* language teaching
 semantic prosody *see* prosody, semantic
 semantic roles 139, 232, 575, 761–763, 954, 962–963 *see also* verb frame
 semantic tagging *see* annotation, semantic
 semantics 16, 353, 745, 896, 960–964, 972–987, 1100–1102, 1206
 sense *see* word meaning
 sense tagging *see* annotation, semantic
 SENSEVAL 575, 963
 sentence alignment *see* alignment, sentence
 sentence boundary detection 529–535
 SGML 47, 253–254, 282, 294, 298, 486, 533, 1165, 1167
 significance *see* statistical test, significance
 similarity measure 1255–1271
 – edit distance 1179, 1258–1259
 simple-II *see* association measure, simple-II
 single document summarisation *see* summarisation, single-document
 size *see* corpus design, size
 social class *see* social variables, social class
 social variables 11, 97, 100, 109, 155, 159, 161, 190, 248–249, 264, 300, 635, 907, 1016, 1031, 1093, 1097–1098, 1100
 – age 107, 194
 – gender 104, 248, 293, 845, 1072, 1102–1103
 – social class 97, 103
 sociolinguistics 96–111, 629, 896, 1016, 1050, 1093, 1097, 1099, 1103, 1127, 1129, 1132, 1134

- historical 249 *see also* pragmatics, historical
- sociopragmatics 629–630, 634–635
- sparseness of data *see* data sparseness
- sparsity of data *see* data sparseness
- specialized dictionary *see* dictionary, specialized
- speech
 - emotion in 194–195, 201, 204, 211, 213
 - speech act 190, 204, 616–617, 620–626, 634, 1016–1017, 1202
- speech assessment methods phonetic alphabet *see* SAMPA
- speech corpus *see* corpus, speech
- speech recognition
 - automatic 77, 188, 191, 195, 650, 658, 1029, 1038
 - speech synthesis 77, 188, 192, 201, 567, 650, 654–655, 818, 1038–1039
- spelling variant 66, 102, 254, 320–321, 1098 *see also* normalisation, spelling
- spoken language 101, 110, 159, 187–207, 458, 845, 1008–1024
- spoken language corpus *see* corpus, speech
- spoken vs. written language 11, 106, 110, 169, 247, 262, 305, 827, 845–847, 1014–1015, 1025–1026, 1052, 1058, 1060–1061, 1093–1094, 1103, 1259
- standard deviation *see* statistical distribution, standard deviation
- Standard Generalized Markup Language *see* SGML
- standoff annotation *see* annotation, standoff
- statistical distribution
 - binomial 787–789, 797–798
 - chi-squared 791–792
 - Dirichlet 861
 - dispersion 949
 - LNRE 800, 882 *see also* statistical distribution, Zipfian
 - mean 811
 - median 811
 - mode 811
 - normal 787–789
 - standard deviation 787, 789–790, 799, 832, 1078, 1085, 1290, 1300–1301
 - Zipfian 330–331, 343, 348, 358, 537, 813–818, 861, 870, 882, 1026, 1244 *see also* Zipf's law *see also* statistical distribution, LNRE
- statistical test 779–784, 786, 790, 794, 798–799, 949, 1072, 1076–1078, 1227, 1235, 1245, 1301, 1303
- ANOVA 799–800, 832, 834, 1300, 1303
- binomial 782–787, 789–790, 798, 949
- chi-squared 782, 792–793, 795, 1073, 1076, 1085
- confidence interval 785–787, 789, 793, 800
- effect size 784–787, 793, 799, 942, 1227–1238, 1244–1245
- Fisher's exact 792–793, 1235, 1237
- Fisher-Yates exact 942–946
- multivariate 825, 1080–1082 *see also* multivariate analysis
- non-parametric 800, 1303
- parametric 1303
- significance 144, 781–782, 784, 1228, 1235, 1237
- t-test 798–799, 1072, 1300
- two-sample 790–794
- two-sided 782–783, 789, 1227–1228
- statistical variable
 - interval 798–799, 1290, 1300
 - nominal 798, 1290, 1293, 1298, 1300, 1306
- statistics
 - descriptive 1286–1304
 - inferential 778, 800, 1303 *see also* statistical test
 - population 779–783, 794
 - probability 780
- stemming 553–554, 697, 883 *see also* lemmatisation
- structuralism 5, 15, 17–19, 21, 446, 707, 907, 1142
- style 105–107, 1061, 1070, 1149–1151, 1160
- stylometry 77, 846, 896, 907–908, 1070–1090, 1253–1254
- subcategorisation *see* verb, subcategorisation
- subjacency 925–927
- summarisation 581, 1260–1261, 1271–1286
 - automatic 85, 1271–1286
 - multi-document 1253, 1261–1262, 1277–1278
 - multilingual 1272, 1282
 - single-document 1272–1274
- supervised learning *see* machine learning, supervised
- support vector machine *see* machine learning, support vector machine
- sustainability 484, 493, 759–776
- SVD *see* principal components analysis
- synonymy 92, 142, 322, 353–354, 960, 974, 977–980
- syntactically annotated corpus *see* treebank

syntax 4, 14–32, 46–47, 895, 920–933, 934–952, 1208–1209 *see also* treebank
see also parsing *see also* annotation, syntax

syntax annotation *see* annotation, syntactic
 syntax-semantics interface 953, 965

Systemic-functional Grammar 428, 506, 512

T

tagger 82, 88–91, 256, 303, 502–527, 540–551

– Brill 510

– CLAWS 45, 90–91, 254, 510, 516–517, 634

– TAGGIT 43, 45, 90–91

tagging *see* annotation

tagging guidelines *see* annotation guidelines

TAGGIT *see* tagger, TAGGIT

tagset

– anaphora 583–589

– part-of-speech 504–507, 512–513

– pragmatic 620–632

TEI *see* Text Encoding Initiative

terminology 136, 283, 490, 686, 696

test *see* statistical test

text category *see* text type

text corpus *see* corpus, text

Text Encoding Initiative 47, 306, 485–488, 491, 495–497, 499, 1165

– guidelines 254, 306, 460, 485–488, 496

text re-use 1249–1271

text summarisation *see* summarisation

text type 105, 171–172, 190, 247–248, 293, 297, 306, 332–333, 511, 514, 520, 752, 846, 1052, 1060–1062, 1072, 1298–1299

text-to-speech systems *see* speech-synthesis

TF.IDF 886, 1256

theory-driven approach 992–993

thesaurus 350, 353–356, 370, 567, 570–573

ToBi *see* Tones and Break Indices

tokenisation 508, 513, 516, 519, 522, 527–540, 687

– ideographic script 532, 536–540

Tones and Break Indices 200, 202, 654

tools

– annotation 218, 235, 589–592, 658, 680, 765

– search 143, 157, 315, 317, 712, 716–717, 720, 722, 724, 735–736, 747, 923–924, 1143, 1160, 1167

topic/focus annotation *see* annotation, information structure

topology 342, 351, 361–362, 371, 739–743, 752–757

– manifold 880–881, 886–892

training *see* machine learning, training

transcription 12, 39, 159, 208–210, 216, 217, 218, 642–663, 1012–1013, 1045, 1200–1203

– orthographic 189, 195–197, 645–660, 1014, 1135–1136, 1201

– phonetic 645, 647, 649–653, 656, 658–659, 1030, 1135, 1203

– phonological 650, 1207

translation 276, 278–283, 286, 320, 686, 698, 1001–1004, 1063–1064, 1141–1142, 1146–1148, 1151, 1159–1175, 1175–1196

translation corpus *see* corpus, parallel

translation correspondence 278, 283–284, 286

translation effect *see* translationese

translation memory 1178–1181

translation practice 1168–1169

translation unit 1166

translation universal 1169–1170

translationese 278, 1088, 1142, 1154, 1162

Tree Adjoining Grammar 70, 601

tree alignment *see* alignment, tree

treebank 46, 225–241, 421–426, 598–613, 738–759, 922, 924, 1191

– historical 227, 426

– parallel 228

– spoken language 227, 232, 423

t-score *see* association measure, t-score

t-test *see* statistical test, t-test

TTS *see* speech-synthesis

Turing machine 71, 73

two-level morphology 71, 74, 559–560

two-sample test *see* statistical test, two-sample

two-sided test *see* statistical test, two-sided

type-token statistics 879, 902, 1073, 1169, 1204

typology 1130, 1143, 1151

U

UML *see* universal modelling language

Unicode 253, 460, 485, 532, 679

universal modelling language 677

unsupervised learning *see* machine learning, unsupervised

V

- valency *see* verb, subcategorisation
variable selection *see* experiment design
variation 110, 154, 192, 331, 824–825,
1091–1094, 1097, 1130, 1161, 1291–1295
– dimensions of 830–834
– textual 846, 1059–1062
variationism 54–60, 66, 98, 100, 109, 825,
1093, 1103, 1110, 1129–1133 *see also* language change
vector space 329, 877, 878, 880–882, 884,
1218, 1255–1256
verb
– features 965–966, 998, 1100
– subcategorisation 133, 142, 952–972
verb class 952–972, 997–998
verb frame 552, 952–972
VGC *see* vocabulary growth
video recording *see* recording, video
Viterbi algorithm 79, 537–538, 543–546,
605
vocabulary growth 373, 819, 902–903,
1026–1027 *see also* word frequency distribution
vocabulary richness 1071, 1073–1075 *see also* word frequency distribution

W

- Web corpus *see* corpus, Web
Web linguistics 310–311 *see also* corpus,
Web *see also* internet
wiki 293, 298, 328, 338, 350, 368–373

- word alignment *see* alignment, word
word formation *see* morphology, word formation
word frequency distribution 318, 803–804,
806, 813, 817–820, 902, 907, 1026, 1073,
1244
word meaning 972–987
word sense *see* word meaning
word sense disambiguation 87, 142, 323,
565–576, 896, 963, 965, 973, 1175
written corpus *see* corpus, text

X

- XCES *see* Corpus Encoding Standard
xenophones 649, 1029–1034
XML 253, 266, 282, 294, 298, 485–487,
498–499, 533, 677–679, 713, 763, 768–
770, 772, 1165, 1167
– Document Type Definition 485, 764, 487,
677
– Schema 485, 487

Z

- Zipfian distribution *see* statistical distribution, Zipfian
Zipf-Mandelbrot's law 365, 804, 813–818,
907
Zipf's law 331, 365, 813–818, 920, 923, 930,
1026, 1215, 1244 *see also* statistical distribution, Zipfian
z-score *see* association measure, z-score