Keyness in Texts

Studies in Corpus Linguistics (SCL)

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

Volume 41

Keyness in Texts
Edited by Marina Bondi and Mike Scott

Keyness in Texts

Edited by

Marina Bondi
University of Modena & Reggio Emilia

Mike Scott
Aston University

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

Keyness in texts / edited by Marina Bondi and Mike Scott.
p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 41)
Includes bibliographical references and index.
1. Semantics. 2. Discourse analysis. 3. Corpora (Linguistics) 4. Phraseology. I. Bondi, Marina. II. Scott, Mike, 1946-
P302.K48 2010
401'.43--dc22 2010029119
ISBN 978 90 272 2317 3 (Hb ; alk. paper)
ISBN 978 90 272 8766 3 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Perspectives on keywords and keyness

An introduction

Marina Bondi
University of Modena and Reggio Emilia, Italy

"All words are equal, but some are more equal than others"
(adapted from Orwell's *Animal Farm*)

Lexical items enjoy equal status in the lexicon of a given language, but their importance varies from the point of view of text. Each individual word form contributes to the construction of meaning in text, but only some words are key-words, i.e. words that play a role in identifying important elements of the text. Similarly, any given language is constituted by all the lexical elements that become part of it, but only some lexical elements are taken to characterize its cultural specificity.

Starting from the different interpretations of the expression "keywords" – as searching tools, in text mining and classification, but also as analytic tools in text interpretation and discourse analysis – this introduction focuses on the relationship between words and text, looking at the co-text of the word, but also at the cultural context that informs the text, where culture is taken to mean the repertoires of meanings shared within a community (e.g. national, or local, but also disciplinary). Keywords are often taken to be markers of the "aboutness" and the style of a text (Scott & Tribble 2006: 59–60): what we want to investigate here is what structures of textuality keywords point to and how far they are also influenced by the position of the writer, in the context of text production.

1. Keywords and keyness in language studies

The notion of keyword has no well-defined meaning in language studies. The definition of a "word" as such may be seen as problematic in modern linguistics; de Saussure's search for the basis of a scientific study of language as system led him to units different from the word at various levels of analysis – phonetics and phonology, syntax, morphology, semantics.

Lexical analysis has long been concerned with the ways in which language, and lexis in particular, instantiates culture. Already in the nineteen-thirties, Firth's lexical semantics proposed the study of "sociologically important words, what one might call focal or pivotal words" and advocated an analysis of the distribution of words whose meanings characterize a community by occurring in specific contexts, with specific associations and values (1935: 40–41). On the basis of anthopological notions of context, and referring in particular to Malinowski, his colleague at the London School of Economics, Firth showed how the study of words in context can illuminate meanings that characterize a culture and a community, referring for example to the development of the meanings of *clerk* in Middle English from medieval clerics.

Similarly, Cultural Studies – Williams (1976) in particular – made an attempt to produce an analysis of contemporary culture through the study of a number of "cultural keywords", i.e. the 'dictionary' of a culture and a social group. The meanings of words like *alienation, capitalism, family, fiction, hegemony, literature, media, tradition* etc. were taken to represent the most distinctive features of contemporary western culture, by integrating synchronic and diachronic perspectives in a full appreciation of meaning. Williams thus made the link between keywords and discourse communities even more explicit, but he clearly oriented the analysis to historical and social macro-contextual factors only, not paying much attention to text and genre and leaving methodological tools for the analysis of meaning completely undiscussed.

A similar focus, but with a different perspective – oriented to the distinction between semantic universals and cultural underpinnings of a language – is provided by Anna Wierzbicka (1999, 2006). Wierzbicka looks at lexical semantics through her Natural Semantic Metalanguage (NSM) as a key to the history, culture and society that produced it, considering the impact of values on interaction and its strategies. Her approach aims at counteracting both the tendency to mistake Anglo English for the human norm and widespread attempts to deny the existence and continued relevance of the cultural baggage of English in international communication. She looks for example at typical features of "anglo" culture such as the ideal of accuracy, the practice of understatement, recourse to "facts" and emphasis on rationality as against emotions. The importance of the meanings associated with a word like *reasonable* shows that "reasonableness" may prove to be the most effective persuasive strategy in an anglophone cultural context, which leaves little room to asymmetrical relations and denies persuasive power to both pleading and authority claims. Her study of the historically shaped cultural meanings of words like *right, wrong, reasonable, fair* aims at revealing covert meanings making the heritage of a common spirit perceivable. Her framework combines cognitive and interactional perspectives, attention to thinking, speaking and

doing, with an interesting emphasis on the impact of values on interactive strategies, while still relying on an intuitive process of keyword and data selection.

Keywords are not necessarily a key to culture, however: they may facilitate understanding of the main point of a text, constituting chains of repetition in text. Whether referring to words that are key to the intepretation of a text or key to the interpretation of a culture, the study of keywords has become central in corpus linguistics, especially through the development of techniques for the analysis of the meaning of words in context. In a quantitative perspective, keywords are those whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus (Scott 1997; Baker 2004; Scott & Tribble 2006).

Identifying elements that are repeated to a statistically significant extent does not in itself constitute an analysis or an interpretation of the text or corpus. It does however point to elements that may be profitably studied and need to be explained. It certainly does point to fundamental elements in describing specialised discourse or in placing a text in a specific domain. The problem for the researcher, of course, lies both in the design of appropriate and adequately representative corpora and in the delicacy of the analysis, with its capacity to isolate specific questions and avoid overgeneralization.

In a corpus perspective, keywords are studied through their typical co-occurrence with other lexico-semantic units. Michael Stubbs (1996, 2001), for example, has shown the importance of concordance analysis in this field: the cultural and ideological implications of a lexical element can be illuminated by an analysis of its collocation and semantic preference – the tendency of the word to co-occur with other words and with words belonging to a specific semantic category or field (see also Sinclair 1996).

The notion of quantitative keyness applies equally to word forms, lemmas and word sequences[1]. The definition thus easily adapts to more complex units than the word, pointing towards a perspective that is gaining ground in present-day descriptive and theoretical language studies: phraseology. Keywords, in fact, are not necessarily single words: we can look at key-clusters (repeated strings of words)[2] or even key-phrases, when extended units of meaning (Sinclair 1996)

1. In a semantic perspective the notion has also been recently extended to semantic elements (Rayson 2008). These of course can only be based on previous semantic analysis and tagging of the corpus, on the basis of given semantic descriptors.

2. In the field of *natural language processing*, computational linguistics and corpus linguistics, research on 'n-grams', also called 'word clusters', 'lexical clusters' or 'bundles' (cf. Biber, Conrad & Cortes 2004; Carter & McCarthy 2006) studies contiguous word forms building up to create repeated word sequences in the corpus.

are considered, i.e. words in combination originating a unit of meaning that can be different from the sum of the constituent lexical units. In the words of John Sinclair (2005), a corpus perspective looks at words in combination and finds in phraseology the ideal starting point for the exploration of the systematic relation between text and form.

Emphasis on phraseology has been increasing in corpus research (e.g. Hunston & Francis 2000; Moon 2002; Hunston 2004). The revived interest finds its origin in Sinclair's notion of collocation (e.g. Sinclair 1991) and in his "idiom principle", highlighting that in the linearity of text each choice narrows down the range of possible choices in the elements that follow and that "a language user has available to him or her a large number of semi-preconstructed phrases that constute single choices" (1991: 110).

Phraseological studies have shown a tendency to shift their attention from fixed, opaque multiword units to a much wider range of units. The focus of interest can thus be extended to discontinuous or inverse relations ("concgrams", Cheng, Greaves & Warren 2006) and patterns (Hunston & Francis 2000).

The key lexical elements of a text create a dense network of intercollocation, including both continuous and discontinuous phraseological patterns. Following Phillips (1989), for example, we can look at a collocation like that between *electric* and *charge*, but also at the patterns created in text between their collocates (e.g. for *charge*: *distribution, density, point, uniform*; for *electric*: *dipole*). The network of lexical relations of this kind would contribute to an identificaton of the "aboutness" of a text.

When lexical analysis combines with semantic analysis, looking at the extended unit of meaning with its corollary of semantic preference and semantic prosody (Sinclair 1996), attention to the co-text means identifying both the potential semantic associations between otherwise different forms and the association of the unit with further textual-pragmatic meanings. A recent development along these lines is the corpus study of semantic sequences, i.e. "recurring sequences of words and phrases that may be very diverse in form […] more usefully characterised as sequences of meaning elements rather than as formal sequences (Hunston 2008: 271).

2. The keyness metaphor in knowledge management: "Aboutness" as subject matter

The meaning of keywords is often explored through the metaphor on which the expression is based. A key is a tool that gives you access to something. The metaphor refers to the power of opening (and closing), revealing (or hiding) what is unknown or unclear. A keyword gives access to features of a text or corpus that

are not immediately obvious: but what are these features? What textual doors are opened by keywords?

The first meaning of *keyword* is perhaps the most obvious in knowledge management, where keywords are those that help identify a text in structured data-bases, such as for example library resources. Textual data-bases can be searched making use of keywords to be automatically retrieved in pre-defined fields: title, author, abstract, subject descriptors, or the text itself. A range of tools is needed because titles are not always the best indicators of the subject matter of a text. This is quite obvious in literary writing: no-one thinks of *Hamlet* or *Othello* as indicators of subject matter or theme. It is less obvious but equally true of professional communication. A text entitled *The Danger Model: A Renewed Sense of Self* cannot be automatically attributed to a subject or a discipline. When we see it is a viewpoint article published in *Science* in 2002, we can probably exclude some of the expectations created by the title, but we need at least descriptors to understand that the field is immunology. The abstract reveals that the text discusses a change in paradigm in immunological studies, a shift from a vision of immunology as based on the distinction between self and non-self to a vision of the immunological system as worried about danger rather than foreignness. Titles can be seen as a key to texts, though not always the most direct key to their subject matter.

With the proliferation of scientific publications and the ever increasing use of textual data-bases, keyword searches have become central to knowledge management. Subject classification, however, is mostly realized from a perspective that is external to the text itself, making use of bibliographical classifications of knowledge such as the Dewey system. Author-produced keywords have also been used, though with unstable results. A priori categorizations are intersubjectively valid but they lack flexibility. Author-produced keywords are more flexible but they lack intersubjective comparability. Knowledge representation has become a key issue: from general and domain ontologies, to semantic networks and "frames". The attention often shifts from lexical units characterizing the surface of text to the possibility of recovering meaning structures beyond lexical forms.

The development of information science and of web-based knowledge, however, has shifted attention from information retrieval to information extraction. The availability of enormous quantities of unstructured data on the web poses the question of information "extraction": text mining requires tools that can move from lexical forms to meanings and their structures, thus finding keywords through the text rather than outside the text. In text mining, just like in current linguistic research, phraseological units are gaining importance. Tools for the identification of keywords are being developed on the basis of frequency data that do not simply look at individual word forms, but rather at relations between words that frequently co-occur.

This brings us back to keywords as words whose frequency (or infrequency) in a text or corpus is statistically significant. The vast majority of the keywords that can be determined by automatic analysis of a text will be key to its subject matter.

3. The keyness metaphor and text interpretation: Subject matter and organization

The notion of text has been one of the most influential in theoretical and methodological developments in linguistics. The past forty years have shown growing interest in meaning making processes beyond the basic syntactic unit of the clause or clause complex, starting for example with work on textuality and text cohesion (e.g. Halliday & Hasan 1976; Beaugrande & Dressler 1981; Conte, Petöfi & Sözer eds. 1989) and leading up to recent interest in patterns of lexis in text (Hoey 1991) as well as meaning units in the linearity of text (Sinclair 2004; Sinclair & Mauranen 2006). Lexical elements can be shown to play a key role in the cohesion of text (signalling and establishing relations between lexical units) and in textual coherence (the conceptual and functional unity of a text). In such a textual perspective, words can become key to the conceptual structure of the text – very much in the same way as in librarianship they define its subject matter – but also to the organizational structure of text – in ways that may also be illustrative of its communicative purpose.

Cognitive and pedagogic approaches to text have often shown that for the act of reading the words that organize text may be more important than those that identify its "content", because they guide the reader towards the elements of content. Signals of organizational structure will thus be key (or "pivotal") in reading because they facilitate access to the information required. If exploring a data-base requires use of keywords that constitute a map of existing knowledge, exploring a text requires use of organizational keywords that act as a textual map. Keywords signalling textual organization act as signposts and help readers identify generic patterns and locate information. When looked at from this perspective, keyness also links to a vast literature on meta-discourse and its role in reading (e.g., Vande Kopple 1985; Crismore 1989; Hyland 2005; Ädel 2006). Let us take the basic metadiscursive structure of two abstracts like the following as an example:

(1a) In this paper we investigate the implications of… In the received theory of …In our model… Hence, we conclude that …

(1b) Recent studies highlight increasing recognition of… It is understood now that …… This article overviews … and outlines…, including …

Irrespective of whether we are talking about nanotechnologies or market structures, the elements reported highlight the basic communicative structure of the text they represent: in the first case the text presents a new model contrasting it with more consolidated theories, while the second introduces a critical review of an issue on the background of recent developments in the field. If subject matter is essential in retrieval, communicative purpose and genre (research article vs review article) may be equally important in reading and metadiscursive elements act as signposts to actual content.

Key-words, key-clusters and key-phrases are not always elements of the conceptual structure of a text. There may be elements of grammatical structure or elements of self-reference. These become useful pointers to the most frequent textual structures of a text as well as its most frequent metadiscursive phaseology.

We may thus think of two kinds of keywords, much in the same way as Sinclair and Mauranen's "Linear Unit Grammar" distinguishes two kinds of unit in the linearity of text – "message-oriented elements", contributing to the topical continuation of discourse, and "organization-oriented elements", that contribute to managing discourse (2006: 59–60). On the one hand, there are keywords that point at the conceptual structure of a text, its "aboutness", what the text is about. On the other, there are keywords that point at issues that may prove to be useful indicators of the communicative purpose and micro- or macro-structure of the text, what the text does and how.

4. Keyness in text and discourse: A sample analysis

As will be apparent from the rest of the volume, work on keyness in text easily leads to work on discourse, linking language use beyond the sentence to the study of social practices and ideological assumptions associated with language (thus involving the different definitions of discourse listed by Schiffrin, Tannen & Hamilton 2001: 1). Words and phrases that are key in a text or in a corpus may be shown to be indicative of the writer's position and identity, as well as of the discourse community,with its values and beliefs about the subject matter and the genres that characterize it (e.g. Baker 2006; Biber, Conrad & Cortes 2007; Ädel & Reppen eds. 2008).

In studies of academic discourse, for example, the acquisition of academic literacy has often been seen as a process of enculturation of students into disciplinary communities through a process of informal learning, of apprenticeship into the ways of speaking of the community (Berkenkotter & Huckin 1995: 7). Academic discourse communities are seen by John Swales as social groupings identified by "a broadly agreed set of common public goals", participatory mechanisms of intercommunication, specific genres and lexis, and "a threshold level of members

with a suitable degree of relevant content and discoursal expertise" (1990: 24–27). Research perspectives have become more and more interested in cross-disciplinary analysis, focusing on the role played by "disciplinary culture in defining what is conventionally seen as acceptable argument or textual organization" (Hyland 2000; Hyland & Bondi eds. 2006). Cross-disciplinary research has recently extended the attention traditionally paid to domain terminology to include interest in general lexis, particularly in the "general academic lexis" that is used across a wide span of domains. It has been shown that different disciplines tend to use it in slightly different ways, on the basis of their methodological tenets. A word like *case*, for example, is frequently used both in economics and in business studies, but it is used in contexts that are fundamentally different and representative of different argumentative frameworks. The word occurs most frequently in collocations like (*the*) *case of* or (*BE*) *the case* in economics, thus signalling the setting up of hypotheses and scenarios, whereas in business studies it is more often found in expressions like *case study* or *case in point,* signalling an exemplification or an illustration (Bondi 2006).

The words and expressions that recurrently identify the conceptual structures and the organizational structures of a text or corpus can be studied to illuminate features of the discourse that produces the text or corpus. The keywords that point to the aboutness of a text or corpus will be key to the ontology of the discourse. The keywords that point to textual organization will be key to the epistemology.

We can explore these preliminary statements through a case study of a landmark text: the *General Theory* by John Maynard Keynes.[3] In the full title – *The General Theory of Employment, Interest and Money* – we can see that *Employment, Interest and Money* have been chosen as key to the subject matter, whereas the choice of *General Theory* is meant to provide a form of self-representation that highlights the main communicative purpose of the writer, as well as his theoretical position. In the first one-paragraph chapter of the volume, Keynes presents his position against classical economic theories, anticipating a criticism of their fundamental postulates as based on a special case, rather than a more general vision:

(2) Chapter 1 – THE GENERAL THEORY

> I have called this book *the General Theory of Employment, Interest and Money*, placing the emphasis on the prefix *general*. The object of such a title is to contrast the character of my arguments and conclusions with those of

3. John Maynard Keynes, *The General Theory of Employment, Interest and Money*, New York, Harcourt and Brace 1936; e-text available from The University of Adelaide Library Electronic Texts Collection (http://ebooks.adelaide.edu.au/).

> the classical theory of the subject, upon which I was brought up and which dominates the economic thought, both practical and theoretical, of the governing and academic classes of this generation, as it has for a hundred years past. I shall argue that the postulates of the classical theory are applicable to a special case only and not to the general case, the situation which it assumes being a limiting point of the possible positions of equilibrium. Moreover, the characteristics of the special case assumed by the classical theory happen not to be those of the economic society in which we actually live, with the result that its teaching is misleading and disastrous if we attempt to apply it to the facts of experience.

The contrast between *special* and *general* provides the starting point for the whole book. The book itself is commonly called "The General Theory", thus giving prominence to what Keynes presented as the element of novelty of his book.

An analysis of the keywords of the text will clearly show that the words in the title also recur as keywords. Using *Wordsmith 5* (Scott 2008), we have calculated keywords with reference to different corpora: a previous book by the same author (*The economic consequences of the peace*[4]), a reference corpus of current economic articles (HEM-Economics[5]) and a general reference corpus (BNC-written component).

The relative positions of *employment*, *interest* and *money* vary slightly but they remain among the top five keywords in all three cases. Although not too much weight can be placed on the order of KWs, as argued by Scott (this volume), nevertheless where the terms are of similar frequency the first positions in keyword lists are often indicative of subject matter and they are relatively consistent across corpora.

The other words included in the top 5 are also worth considering. Both the general written language and Keynes' previous book highlight other content words: *rate* and *investment*. These point at important conceptual elements of the *General Theory* that distinguish it from previous work. The word *rate* is typically used in the cluster *the rate of interest* (348/737 occurrences), which is one of the foci of

---

4. John Maynard Keynes, *The Economic Consequences of the Peace*, New York, Harcourt, Brace and Howe, 1920; E-text available from The Project Gutenberg Online (http://www.gutenberg.org/files/15776/15776.txt).

5. The corpus comprises 436 articles published in 2000–2001 from the following journals: *European Economic Review (EER), European Journal of Political Economy (EJOPE), International Journal of Industrial Organization (IJOIO), International Review of Economics and Finance (IREF), Journal of Corporate Finance (JOCF), Journal of Development Economic (JODE), Journal of Socio-Economics (JOSE), The North American Journal of Economics and Finance (NAJEF).*

the book, essential to a theory of investment. *Investment* also marks an important shift in Keynes' work, moving from the international economic framework of the first book – basically an example of economic historical analysis – to an emphasis on the micro-economic foundations of macro-economics in the *Theory*.

The corpus of current economics articles, on the other hand, highlights grammatical words: *of* and *which*. *Which* is mainly to be attributed to frequent use of relative defining and non-defining clauses; a look at collocates will show that the nouns specified by the relative include all the important conceptual elements of the text. Here are the top twenty nouns in the position immediately preceding the relative, in order of decreasing frequency: *interest, factors, investment, money, employment, equipment, level, amount, income, consumption, factor, rate, capital, cash, sum, production, cost, theory, demand, value*. The list includes the three keywords we started from, as well as many other words referring to related concepts.

The presence of *of* reflects a preference for nominal postmodification against noun + noun constructions. If we look at the clusters it is found in, we see a vast dominance of important phraseological structures: *the marginal efficiency of capital, the rate of interest, the quantity of money* are the most frequent, variously related to a theory of investment.

The grammatical words we have found, then, do not point directly at the subject matter of the text, but rather at typical constructions used: the complex nominals used to characterize complex notions and the need to define these in terms of the processes they are characterized by. Both *which* and *of* can be seen as matters of individual style. In this case, however, they are more likely to reflect differences in register due to genre and diachronic change. The reference corpus in fact is representative of a much denser form of writing and also of a much later stage in the development of the discipline, a stage in which terminology based on nominal contructions has been developed to a great extent.

Moving from simple keywords to key-clusters we are more likely to find the complex notions expressed in phraseological terms and also to find other pointers to the typical structure of discourse. Looking at 3–5-word clusters with reference to the *Economic Consequences of the Peace* and to the current economics articles shows very similar clusters, with 4/5-word content keyphrases like *the rate of interest*, (*the) marginal efficiency of capital* and *the quantity of money* among the top five. 3-word lists also show a number of organizational key-phrases like: *in terms of, is equal to, as a whole, it is the, in the sense, it follows that, as a rule* etc. While hardly signals of subject matter, all these expressions act as signals of frequent communicative acts: defining (*in terms of, in the sense*), identifying in mathematical terms (*is equal to*), highlighting (*it is the*), deducing (*it follows that*) etc. These can again be considered matters of individual style, but they also point at important features of the genre and of the writer's authorial identity, highlighting that

we are dealing with scientific argument that favours logical structures. The fact that signals of logical deduction are particularly distinctive in comparison with Keynes' earlier work and lose keyness in comparison with current economics articles may be taken as a sign of authorial development towards forms of more formal reasoning that will become characteristic of the discipline at later stages.

Keywords can also be calculated for each chapter with reference to the whole book. Chapter 18, for example, stands out as characterized by relatively little specific language and rather an insistence on general academic lexis, with keywords like *factors, variables, condition, psychological, we*. The most likely explanation of this peculiarity seems to me to lie in the summative nature of the chapter, which is entitled "The General Theory of Employment Re-stated" and begins by stating: "We have now reached a point where we can gather together the threads of our argument". At other points, negative keywords – words that stand out as particularly infrequent – will play an equally important role: *money*, for example becomes a negative keyword at regular points in the book – Chapter 6, 8, 22 and 24, where Keynes tries to correct monetary views of income, saving and investment, analyses the propensity to consume, explains the trade cycle and sums up his social philosophy.

Similarly, if we look at the well known 1936 Preface of the book, we get a very simple picture, with only three keywords.

N	Keyword	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	I	25	2.39	457	0.40	46.83	0.0000000000
2	MY	13	1.24	120	0.11	39.21	0.0000000000
3	BOOK	7	0.67	50	0.04	24.20	0.0000008665

Figure 1. Keywords of the Preface of the General Theory

The Preface is, entirely predictably, about the writer and his book. It is interesting, however, to check the concordances and see that all the three keywords are in fact self-reference items in the Preface. The choice between personal and non-personal elements of self-reference can be better studied in the co-text of concordance lines and a tendency can be noticed to use non-personal reference to introduce the most important statements about the nature of the book that follows, including some among the best known quotes from the text. The occurrences mark the main steps in the rhetorical organization of the preface:

a. Definition of the intended audience: ***This book*** *is chiefly addressed to my fellow economists. I hope that it will be intelligible to others. But its main purpose is to deal with difficult questions of theory, and only in the second place with the applications of this theory to practice.*

b. Relation to previous work on monetary policy: *The relation between **this book** and my Treatise on Money [JMK vols. v and vi], which I published five years ago, is probably clearer to myself than it will be to others; and what in my own mind is a natural evolution in a line of thought which I have been pursuing for several years, may sometimes strike the reader as a confusing change of view.[…] **This book**, on the other hand, has evolved into what is primarily a study of the forces which determine changes in the scale of output and employment as a whole; and, whilst it is found that money enters into the economic scheme in an essential and peculiar manner, technical monetary detail falls into the background.*

c. Acknowledgement of colleagues' support: *The writer of **a book such as this**, treading along unfamiliar paths, is extremely dependent on criticism and conversation if he is to avoid an undue proportion of mistakes. […] In **this book**, even more perhaps than in writing my Treatise on Money, I have depended on the constant advice and constructive criticism of Mr R. F. Kahn. There is a great deal in **this book** which would not have taken the shape it has except at his suggestion.*

d. Definition of the innovative nature of the book: *The composition of **this book** has been for the author a long struggle of escape, and so must the reading of it be for most readers if the author's assault upon them is to be successful, – a struggle of escape from habitual modes of thought and expression. The ideas which are here expressed so laboriously are extremely simple and should be obvious. The difficulty lies, not in the new ideas, but in escaping from the old ones, which ramify, for those brought up as most of us have been, into every corner of our minds.*

To conclude, we can focus on the word *theory*, which we saw so prominent in the title and in Chapter 1. *Theory* is again an important keyword with all the three reference corpora (with a keyness score of 230.05 when measured against Keynes's previous work, 232.59 against current economics articles, 881.6 against general writing). The word *general*, on the other hand, is not key in comparison with Keynes' previous work and its keyness score in comparison with current economic writing is relatively low (31.32). It is true however that if we look at the collocation *general theory*, then this becomes highly distinctive: the expression is absent from earlier work and it becomes virtually exclusive of Keynes in current economic writing (49 of the 50 occurrences in the corpus of economics articles are references to Keynes' book).

The contrast between the classical theory and Keynes' own General Theory is emphasized by the fact that the latter is mostly presented as what *we* assume (as against what the classical theory assumes), with an intensive use of the first person plural pronoun *we*, which is key against all three corpora. The pronoun is mostly used inclusively, to accompany the development of the argument. The

G*eneral Theory* as such is mostly present in sections and chapters that are characterized by important medadiscursive nodes – introductory sections and conclusion – whereas in the development of the argument the controversy is between "us" and the classical theorists.

The word *classical*, on the other hand, is key everywhere, though most prominent in comparison with general writing (keyness score 514.3), followed by current articles (240.64) and Keynes' previous work (114.61). Keynes admittedly devoted much of his book to a refutation of classical theory as a basis for his own theory (see quote above), and keyword data certainly confirm this. *Classical theory/doctrine/school* are frequent collocations, typically contrasted with the writer's own discourse and associated with refutation of their postulates. The words *school* and *doctrine* are almost exclusively used to represent classical theory and they are often accompanied by words like *accepted*, *dominant* and *orthodox*, only to emphasize the writer's *divergence* from it.

If we look at the concordance of *theory* throughout the book (246 occurrences), it is easy to see that *classical theory* is by far the most frequent collocation (53 occurrences, almost 20%), even more frequent than *the theory* (41). Postmodification with *of* is also confirmed to be very frequent (109), showing that *theory of* is followed by *employment* and *money* 13 times each, whereas *interest* is only present 3 times, but the emphasis lies rather on a theory of the *rate of interest* (19 occurrences) and falls equally on the analysis of a theory of *unemployment* (13). It is also interesting to notice that apart from *classical* (53), *general* (13) and *economic* (10), most other adjectives preceding *theory* are explicitly evaluative, and mostly of the kind that Hunston and Thompson (2000) would call evaluation in terms of "good" and "bad". Here is the full list:

> bad, correct, faulty, foolish, nonsense, central complete, different (2), fundamental (3), foregoing (2), formal, independent, logical, ordinary, peculiar, preceding (2), prevailing, pure, scientific, separate, traditional (3).

Keynes' criticism of the classical theory is very explicit and centres on its oversimplification and faulty premisses; at the same time the presentation of his own model appeals to logical reasoning through formal refutation of the postulates of classical theory but also through his own forms of simplification. An adjective like *special*, for example, (key when compared with both present-day corpora) is repeatedly associated with *case* in criticism of the classical theory, but the second most frequent collocation is with *sense*, where it typically refers to Keynes' own *special* definitions. Special senses are acceptable, but assumptions based on special cases are criticizable.

Keyword analysis can thus also be related to recent interest in the evaluative features of discourse (cf. Hunston & Thompson 2000; Martin & White 2005; Dossena &

Jucker 2007), typically meant to organize discourse, to establish and maintain relations between writer/speaker and reader/listener, or to manifest the value system of the speaker and the discourse community. In the context of argumentative discourse, keywords can be associated with culturally shared assumptions and values that constitute the implicit premises of argument within a socially situated argumentative practice (cf. also Rigotti & Rocci 2004). Both conceptual and organizational keywords may be a guide to the writer's evaluative position, and through this to the writer's position in disciplinary debates.

How much of this analysis depends on a close knowledge of the text is of course debatable. How much sense can one make of a keyword list without having a good familiarity with the text? Keywords, like most frequency dara, point at elements that need to be explained, but part of the explanation is likely to be found in the co-text of the items, and ultimately in the text.

5. Overview of the chapters

The first section of the book explores the notion of keyness from different points of view. Michael Stubbs outlines the field from the point of view of language studies, discussing three loosely related uses of the term "keyword", as cultural keywords, as statistically meaningful repetition and as phraseological patterns involving extended units of meaning. His main theoretical focus lies on the critical link between words, texts and culture, while he argues for the need to relate words and texts to the social institutions which are characterized by texts and text-types.

The more specific problems and challenges of quantitative approaches, largely dominant in corpus linguistics, are presented by Mike Scott. The chapter maps out the problems of defining keyness, discussing statistical issues and the choice of a reference corpus, as well as illustrating issues of corpus stylistics.

One of the problems highlighted by Scott – the role of closed-class keywords – is picked up by Nick Groom and explored fully in a discourse perspective. The chapter presents the case for a specific focus on closed-class keywords as objects of corpus-driven discourse analysis. Their potential lies in the coverage they offer of phraseological data and in their capacity to reflect the constellations of meanings and values of a discourse community.

Jukka Tyrkkö draws a distinction between key words and keywords. Through the examples of hyperlinks in hypertexts, he claims that words may possess a degree of keyness due to their inherent markedness and their functional properties, rather than to statistical significance. Hyperlinks are paradoxically shown not to be reliable indicators of topic, but still indicative of discursively important topics in the process of text production and reception.

The section closes with a chapter by François Rastier, who offers interesting reflexions on the background against which current research on keywords could be set by looking at the Web. He contrasts traditional programmes of knowledge representation with a corpus-linguistics web semantics – situating knowledge with texts rather than outside them – and advocates a re-thinking of the relationship between data and metadata.

Section II looks at keyness in specialised discourse. Martin Warren's text opens the section and links it to the first, by offering a new perspective on "aboutness". He looks at concgramming, identifying the most frequently co-occurring pairs of words, irrespective of constituency and/or positional variation. Analysis of the lexical concgrams looks at meaningful association to draw up a list of aboutgrams identifying the aboutness of a text. An examination of the text's phraseology and phraseological variation is shown to have great potential in defining the aboutness of a text.

The methodology is further discussed in Denise Milizia's analysis of the speeches of Tony Blair and George W. Bush. The focus of the study is first on the word *climate* and on the co-occurrence of *climate* and *change*. The analysis shows the importance of looking at phraseological units rather than individual words in looking for the aboutness of text.

Andrea Gerbig offers an interesting example of how different approaches can be combined in her study of a a corpus of travel writing, from Early modern English literature to contemporary 'blooks'. Starting from statistically determined keywords, she studies key-keywords and associates, before moving on to contextual analysis of some words as extended lexical units and concluding with an analysis of keyphrases and phrase frames, thus including both repeated strings of words and repeated patterns.

Looking at phraseological combinations around selected keywords, Donatella Malavasi and Davide Mazzi study how different disciplines represent their own research activity, focusing in particular on subjects and objects of the activity, as well as on research procedures. By highlighting differences in the general lexis of self-representation in history and marketing, the study confirms the centrality of keywords in characterizing disciplines, as well as a considerable degree of intercollocability between selected keywords.

Gill Philip looks at the problem of metaphorical keyness in a corpus of speeches by Italian female politicians. Starting from an identification of statistically generated keywords as mostly associated to a text's content, Philip looks for tools for the analysis of the relationship between keywords and the message of the text (covert keyness) focusing on evaluative language and metaphors. She sets out a method for semi-automatic identification of metaphors and demonstrates systematic interaction with keywords.

The third and final section looks at critical and educational perspectives. Erica Bassi studies how the Kyoto Protocol has been represented in two national newspapers: the Italian *La Repubblica* and the American *The New York Times*. Keywords are grouped into semantic fields to study the meanings associated to the protocol and closer analysis of words denoting 'disaster' and 'alarm' is carried out, emphasising the different strategies used by the two newspapers.

The study by Soon Hee Fraysse-Kim identifies keywords that trigger national consciousness of Koreans through an analysis of school textbooks used in elementary schools in four Korean communities: in South Korea, North Korea, Japan and China. The sense of homogeneity suggested across the politico-social borders is taken to reflect prevailing ideology, internalized and reproduced by school education.

Along similar lines, but moving towards pedagogical implications for literacy, Paola Leone uses keyness to identify the basic lexical patterns of school textbooks and matches them to the language young learners might be exposed to out of school. Results show discoursal, lexical, semantic, and morphological features which may be unfamiliar to the learner and should therefore deserve special attention in syllabus design.

The investigations presented in this book – originally presented at a conference held in Pontignano, Italy, under the title of the present volume – are quite narrowly focused on keyness in a corpus perspective, mostly involving attention to text and discourse. They are, however, illustrative of different topics, approaches, methods and theoretical assumptions. We are grateful to the contributors for this. Most of the contributions, on the other hand, have largely benefited from John Sinclair's ideas. We would like to add this volume to the long list of books dedicated to his memory, with gratitude.

References

Ädel, A. 2006. *Metadiscourse in L1 and L2 English* [Studies in Corpus Linguistics 24]. Amsterdam: John Benjamins.

Ädel, A. & Reppen, R. (eds.). 2008. *Corpora and Discourse. The Challenges of Different Settings* [Studies in Corpus Linguistics 31]. Amsterdam: John Benjamins.

Baker, P. 2004. Querying keywords. Questions of difference, frequency and sense in keyword analysis. *Journal of English Linguistics* 32(4): 346–359.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Beaugrande, R. de & Dressler, W. 1981. *Introduction to Text Linguistics*. London: Longman.

Berkenkotter, C. & Huckin, T. 1995. *Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power*. Hillsdale NJ: Lawrence Erlbaum Associates.

Biber, D., Conrad, S. & Cortes, V. 2004. *If you look at*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.
Bondi, M. 2006. *A case in point*: Signals of narrative development in business and economics. In *Academic Discourse Across Disciplines*, K. Hyland & M. Bondi (eds), 47–72. Bern: Peter Lang.
Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: CUP.
Cheng, W., Greaves, C. & Warren, M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4): 411–433.
Conte, M. E., Petöfi, J. & Sözer, E. 1989. *Text and Discourse Connectedness* [Studies in Language Companion Series 16]. Amsterdam: John Benjamims.
Crismore, A. 1989. *Talking with Readers. Metadiscourse as Rhetorical Act*. New York NY: Peter Lang.
Dossena, M. & Jucker, H. (eds). 2007. *(R)evolutions in Evaluation*. Special issue of *Textus* 20(1).
Firth, J. R. 1935. Technique of semantics. *Transactions of the Philological Society,* 36–72.
Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.
Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: OUP.
Hunston, S. 2004. The corpus, language patterns, and lexicography. *Lexicographica* 20: 100–113.
Hunston, S. & Francis, G. 2000. *Pattern Grammar* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins.
Hunston, S. & Thompson G. (eds). 2000. *Evaluation in Text*. Oxford: OUP.
Hunston, S. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3): 271–295.
Hyland, K. 2000. *Disciplinary Discourses*. Harlow: Longman.
Hyland, K. 2005. *Metadiscourse*. London: Continuum.
Hyland, K. & Bondi, M. (eds) 2006. *Academic Discourse Across Disciplines*. Bern: Peter Lang.
Martin J. R. & White, P. P. R. 2005. *The Language of Evaluation: Appraisal in English.* Basingstoke: Palgrave Macmillan.
Moon, R. 2002. *Fixed Expressions and Idioms in English*. Oxford: OUP.
Phillips, M. 1989. *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham.
Rayson P. 2008. From key-words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549.
Rigotti, E. & Rocci, A. 2004. From argument analysis to cultural keywords (and back again). In *The Practice of Argumentation* [Controversies 2], F. H. van Eemeren & P. Houtlosser (eds), 903–908. Amsterdam: John Benjamins.
Schiffrin, D., Tannen, D. & Hamilton, H. (eds.). 2001. *The Handbook of Discourse Analysis*. Oxford: Blackwell.
Scott, M. 1997. PC analysis of key words – and key key words. *System* 25(1): 1–13.
Scott, M. 2008. *WordSmith Tools*. Version 5. Liverpool: Lexical Analysis Software.
Scott, M. & Tribble C. 2006. *Textual Patterns. Keywords and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.
Sinclair J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
Sinclair J. McH. 1996. The search for units of meaning. *Textus* 9(1): 75–106.
Sinclair, J. McH. 2004. *Trust the Text*. London: Routledge.
Sinclair, J. 2005. What's in a phrase. Lecture given at the University of Modena and Reggio Emilia, 15 November 2005.
Sinclair, J. McH. & Mauranen, A. 2006. *Linear Unit Grammar* [Studies in Corpus Linguistics]. Amsterdam: John Benjamins.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Swales, J. 1990. *Genre Analysis: English for Academic and Research Settings*. Cambridge: CUP.
Vande Kopple, W. J. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication* 26: 82–93.
Wierzbicka, A. 1999. *Emotions Across Languages and Cultures: Diversity and Universals.* Cambridge: CUP.
Wierzbicka, A. 2006. *English: Meaning and Culture*. Oxford: OUP.
Williams, R. 1976/83. *Keywords: A Vocabulary of Culture and Society.* London: Fontana Press.

SECTION I

Exploring keyness

Three concepts of keywords

Michael Stubbs
University of Trier, Germany

The term "keywords" is widely used to refer to words which are important in some way, either in individual texts or in a given culture. The general idea is perhaps clear enough, but there are two problems. First, there are several different concepts of "keywords". This paper discusses three loosely related uses of the term, which derive from quite different academic traditions, and which are therefore only marginally compatible. Second, there is a very large gap between individual words, texts and culture. The paper argues that this gap can only be bridged by a theory which relates words, phrases and texts to the social institutions which are characterized by texts and text-types.

1. Introduction

Keywords are words which are claimed to have a special status, either because they express important evaluative social meanings, or because they play a special role in a text or text-type. From a linguistic point of view, they contribute to the long "search for units of meaning" (Sinclair 1996). From a sociological point of view, they are part of "a vocabulary of culture and society" (Williams 1976/1983). In work on keywords, semantic and social analysis are inseparable. However, the term "keyword" is used in several different senses which are only loosely related, and there is a large gap between individual words and the social world. Words occur in phrases and speech acts, which in turn occur in speech events in different social institutions. So, this paper is about how words relate to the world.

It is also about the state of the art in corpus linguistics. Many corpus studies since the 1990s analyse the words and phrases used in many different text-types, but these studies do not yet amount to a unified body of social theory. Corpus linguists like nothing better than empirical findings supported by levels of statistical significance. But outside this narrow circle, people want to know how it all hangs together, and how all the empirical information contributes to solving

the great intellectual puzzles of language in society. How should all this work be evaluated? How does empirical linguistics contribute to "wider issues", and how can it be used "as a foundation for a broad range of intellectual exploration"? (Sinclair 2007: 1). One attempt to explain how it does all hang together is made by Searle. In the opening sentence of his 1969 book on *Speech Acts*, he asks "how words relate to the world" (Searle 1969: 3). And in his later 1995 book on *The Construction of Social Reality* (Searle 1995), he shows that social reality depends on language, since there is a logical relation between speech acts, the inter-subjective world of social facts, and the structure of social institutions.

But these two approaches to understanding words and the world are very different in style. Corpus linguistics provides a powerful model of communicative acts (Sinclair 1996, 1998, 2005), which is firmly based on empirical facts and statistics, but it is often weak on social theory. Speech act theory also provides a powerful model of communicative acts, and this in turn provides the basis for a powerful social theory (Searle 1995), but it is weak on empirical linguistic data (it often uses only invented examples).

The paper uses the following presentation conventions. Italics are used for *word-forms*. SMALL CAPS are used for LEMMAS. Single quotes are used for 'meanings'. Double quotes are used for "quotes from other authors". Unless otherwise stated, the examples are from the British National Corpus (BNC), a corpus of one hundred million running words of written and spoken English.

2. Part 1. Three concepts of keywords

I will discuss three very different senses of the term "keywords". Sense 1 derives from cultural studies. It is well known from Raymond Williams' work (Williams 1976/1983), though there is also much earlier French- and German-language work. Sense 2 derives from comparative quantitative corpus analysis, which identifies words which are statistically prominent in particular texts and text collections. It is best known from Mike Scott's *WordSmith Tools* (Scott 1998). Sense 3 derives from work on lexico-grammar. In a 1993 article, Gill Francis proposes an ambitious project to discover the phrasal units which express taken-for-granted cultural meanings, and therefore to "compile a grammar of the typical meanings that human communication encodes" (Francis 1993: 155). These three concepts are very different, and possibly not even compatible, but they do at least share the notion of discourse as recurrent and conventional ways of talking which circulate in the social world and which contribute to ways of thinking about the social world. As J. R. Firth (1957: 29) expresses it in one of his more cryptic utterances: "We are in the world and the world is in us". The general idea

of "keywords" is fairly clear, although the metaphor is rather vague. Keywords are the tips of icebergs: pointers to complex lexical objects which represent the shared beliefs and values of a culture.

2.1 Keywords sense 1 (Williams 1976/1983): Words and culture

Sense 1 is explicitly cultural. As Wierzbicka (1997: 156) phrases it, keywords are a "focal point around which entire cultural domains are organized" (for German, she gives examples such as *Heimat* and *Vaterland*). For English-language scholars, the most famous example of sense 1 is Raymond Williams' book (1976/1983) *Keywords: A Vocabulary of Culture and Society*. However, the concept goes back much further. Forty years before Williams, J. R. Firth (1935: 40, 51) talked of "sociologically important words, which one might call focal or pivotal words" (e.g. *work, labour, trade, leisure*).

In addition, thirty years before Firth, there was the beginning of a long tradition of German-language work on the use of *Schlüsselwörter* (= keywords). Dictionaries of words which are important in social and intellectual history were produced from the early 1900s up to the fall of the Berlin Wall. This tradition is variously called *Schlagwortforschung* (= 'research on catch phrases') and *Begriffsgeschichte* (= 'the history of concepts'). Early 20th century examples include a historical dictionary (Ladendorf 1906) and a dictionary of keywords at the time of the Reformation (Lepp 1908). Late 20th century examples include a dictionary of *Brisante Wörter* (= 'controversial words') *von Agitation bis Zeitgeist* (Strauß, Hass-Zumkehr & Harras 1989), a dictionary of controversial concepts in public discourse (Stötzel & Wengeler 1995), and a dictionary of keywords from the *Wende*, the fall of the Berlin Wall (Herberg, Steffens & Tellenbach 1997). Other work includes an article on *politische Vexierwörter* (= 'ambiguous political words') (Teubert 1989). There is also a long tradition of French-language work. In the 1950s Georges Matoré (1953) discussed *mots clés* (= 'key words') and argued that lexicography is a sociological discipline. Also from the 1950s is an article by Émile Benveniste (1954) who discusses the word *civilisation*, which is used rather differently from the word *civilisation* in English. His work is based on still earlier work by Lucien Febvre from 1930. This tradition was continued by Michel Foucault, who had his favourite keywords (e.g. *labour, madness, prison*) (Hacking 1986: 27).

Williams proposes a rather small set of around 120 words which are important in the culture: though quite how the culture might be defined, I am not sure. Four characteristics of his keywords are as follows. (1) First, Williams identifies words intuitively, on the basis of his extensive scholarship. He then uses the

attested citations in the 12-volume *Oxford English Dictionary* as empirical evidence that his keywords have undergone historical shifts in meaning which have led to complex layers of meanings in contemporary English. They are "difficult words", as he puts it. (2) Second, only some of his keywords are in widespread use (e.g. *country, expert, family, genius*), whereas many are from an intellectual discourse and most native speakers of English would not have the slightest idea what they mean (e.g. *alienation, dialectic, hegemony, utilitarian*). But Williams has no explicit theory of the organization of the vocabulary (e.g. core versus specialized) or of text-types or discourse communities which could explain this distinction. (3) Third, Williams assumes that keywords do not just label, but help create, conceptual categories. He talks of "significant, indicative words in certain forms of thought" (Williams 1983: 15). Work on keywords necessarily implies a constructivist Whorfian perspective. (4) Fourth, Williams' particular interest is a marxist-socialist analysis of the social order. In his article on the discourse on the miners' strike in 1984–85, he discusses four "slippery" keywords / phrases: *management*, *economic, law and order* and *community* (Williams 1985). He discusses "the key issue in the whole modern organization of work": whether workers can control their own production or whether *management* simply means 'employer' (even in nationalized industries).

Example: The semantic field 'work and leisure'

A key semantic and cultural domain identified by both Firth and Williams is that of 'work versus leisure', and Williams (1976 / 1983) shows that related words (such as *work*, *career, job, labour, (un)employment*) have distinctly different uses and connotations. He shows that the word WORK has developed historically from general meanings of 'doing something', to more restricted meanings concerning the social relationship of paid employment.

One limitation in Williams' work is that he does not distinguish between different forms of a lemma, which often have very different collocates and uses. In a small corpus study (Stubbs 1996: 177–78), I showed that the word-forms *work*, *working* and *worker* occur in quite different compounds and fixed phrases:

- workaholic, workforce, workload, workplace, workroom, worksheet, workshop, workstation, worktop, workwear
- working class, working conditions, working mother
- aid worker, factory worker, office worker, social worker

Some of these constructions are highly productive: nouns immediately preceding *worker* include, amongst many others

– airport, bakery, bank, brewery, building, care, charity, coalface, community, construction, defence, farm, forestry, government, health, hospital, hotel, housing, kitchen, maintenance, morgue, sex, steel, welfare, youth

And some of these phrases are replacing other terms, for example

– bank clerk > bank worker, miner > coalface worker, agricultural labourer > farm worker

This is a clear example of how empirical linguistic data can contribute to a social analysis. As Teubert (2007: 59) points out: "without a word for *work* there would not be the social construct 'work'." In addition, this semantic field has undergone rapid historical change, is still strikingly productive, and this is an indicator of social change.

Summary of Williams' approach

Williams' work was within a cultural studies tradition, and was not intended to be a linguistic analysis, so you may think that the following points are unfair. However, from the point of view of linguistic analysis, Williams' list is not a good basis for a general theory of meaning. The list covers a very small set of words, many of which are very specialized. He has no way of providing comprehensive coverage of the vocabulary of a language, he has no theory of how the vocabulary of a language is organized, and he has no way of relating words to texts and text-types.

2.2 Keywords sense 2 (Scott & Tribble 2006): Words and texts

Sense 2 is statistical: keywords are words which are significantly more frequent in a sample of text than would be expected, given their frequency in a large general reference corpus. This concept is extensively discussed by Scott & Tribble (2006), who are very explicit about the merits and limitations of Scott's keywords software (part of the package *WordSmith Tools*: Scott 1998).

(1) First, "keyness is a textual matter" (p. 65). Certain words characterise individual texts (such as *Romeo and Juliet*), as well as text-types and intellectual areas (such as medicine and natural science, p. 29). (2) Second, the software turns texts into word-lists or lists of n-grams, and then compares the lists from different text collections. By filtering and sorting the lists, vast quantities of text are reduced to much simpler patterns (p. viii, 5, 40), which are invisible to the naked eye. (3) Third, content words directly indicate the propositional content of texts. However, although "keyness is a textual matter", since the texts have been ripped apart into lists of individual words and/or n-grams, the patterns ignore

text segmentation. They are a feature of global textual cohesion, but not textual structure: unless the technique is adapted in various (minor) ways, which I will mention below.

I can illustrate some characteristics of this approach with two small case studies.

Example 1: Textual collocates, semantic fields and phraseology

The transcripts of the Hutton Inquiry are available in the world-wide-web[1]. This was an inquiry, chaired by Lord Hutton, into the death in 2003 of Dr David Kelly, a British government weapons inspector during the run-up to the war in Iraq. I compared the transcripts (around 930,000 running words) with the BNC as a reference corpus. As always, amongst the content words in the top 50 keywords were several proper names, plus words which clearly indicate some main topics of the transcripts. If you have forgotten the case, these keywords will remind you of major themes:

- Kelly, Dr, Hutton, Gilligan, Lord, Kelly's, BBC (and other proper names)
- dossier, intelligence, MOD, think, source, FAC, press, JIC, ISC, July, meeting, statement, evidence, committee, draft, September, letter, name, inquiry, document[2]

However, one gains a much better impression of both the content and of the formal nature of the discourse, via longer recurrent expressions. Here are some of the most frequent 4-grams, which all occur 125 times or more, in descending frequency[3].

- the Ministry of Defence; the Foreign Affairs Committee; weapons of mass destruction; the 45 minutes claim; the Joint Intelligence Committee; Intelligence and Security Committee
- I do not think; can I take you; thank you very much; I do not know; I think it is; I think it was; I am not sure; in relation to the

---

1. The transcripts may be reproduced free of charge providing that Crown Copyright is acknowledged. They are available at http://www.the-hutton-inquiry.org.uk/index.htm (accessed July 2007). For help in constructing the corpus I am grateful to Simone Dausner.

2. *MOD* is the Ministry of Defence, *JIC* is the Joint Intelligence Committee, *FAC* is the Foreign Affairs Committee, *ISC* is the Intelligence and Security Committee.

3. The n-grams were extracted with Bill Fletcher's software kfNgrams, http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (accessed July 2007).

In order to capture characteristics of the discourse, we need this phraseology. The most frequent 5-gram (*can I take you to*) is part of the polite, formal, cautious, public usage of Lord Hutton himself. It occurs in utterances such as:

– can I take you to MoD/1/44 page 19. It is a memo ...

Example 2: Textual collocates, semantic fields and point of view
Friedrich (2007) assembled a 30-million-word corpus of newspaper articles, published between 1996 and 2006, in British periodicals such as *The Spectator*, quality papers such as *The Times*, and tabloid papers such as the *Sun*. He selected all articles which contained the word-forms *islam*, *muslim/s* and *middle east*. Using the BNC as a reference corpus, keywords in the top 50 in all three groups of newspapers included:

– *all sub-groups*: Iraq / Iraqi, war, Israel, Saddam, Palestinian, terrorism/t

Some individual papers were clearly distinguished by keywords in the top 50. For example, within their sub-groups, keywords in either *The Business* or *The Morning Star*, but not both, were:

– *Business*: oil, global, prices, investors, companies, economy, opec, markets, growth
– *Morning Star*: killed, troops, attacks, military, occupation, forces, soldiers

The technique can provide empirical evidence of how the same topic can be represented from different points of view. It is not surprising that the business periodical sees the topic from an abstract financial point of view, and that the tabloid sees the topic from a more concrete personal point of view of the soldiers who are involved, but introspection could not accurately predict the lexis used to create these representations.

Usually collocates are extracted from an immediate span of a few words to left or right of a node word, but a keywords analysis extends the span from phrases to texts. We are dealing here with what Mason & Platt (2006) call "textual collocates". This concept of keywords recalls classic structuralist semantics. The software can automatically extract sets of words, which fall into intuitively identifiable semantic fields, and provide an indication of how homogeneous the vocabulary is across a text. The meaning of a word derives, not directly from the relation between words and their denotation in the world, but from internal relations to other words. Again, it is clear that work on keywords implies a constructivist position.

Summary of Scott's approach

A major merit of Scott's approach to keywords is that it provides an empirical discovery method, based on frequency and distribution. Lists of keywords and phrases obviously differ in different text collections. Semantically related keywords are a good indicator of propositional content. Salient words and n-grams are identified, but not more abstract phraseology. Keywords are a mechanism of global textual cohesion, but for information about text structure, we have to adapt the technique. For example, Starcke (2007) identifies keywords in a Jane Austen novel, then looks at their uneven distribution in the text, and uses this as evidence of text segmentation.

2.3 Keywords sense 3 (Francis 1993): Phrases and schemas

Although she does not use the term "keywords", Francis (1993: 155) proposes a third potentially more radical and ambitious approach to culturally significant units of meaning. She aims to identify what people regularly talk about: their conventional ways of expressing their shared values, such as "how difficult or easy life is made for us, how predictable things are, and how well we understand what is going on" (p. 141). One example of the conventional phraseology which speakers use to express their lack of understanding is illustrated in the following concordance lines:

cos I haven't the	faintest idea who they are.
you haven't the	faintest idea how to spell the word
I hadn't the	faintest idea of his intention
but they had no	faintest notion of man's perverse habit
If you have the	faintest notion you might like it
does not have the	slightest idea that the size, shape and ...
we have not the	slightest notion which parts are effective
I have only the	foggiest idea of what is entailed.
none of those would have the	foggiest notion about it
some of you may not have the	foggiest what a fanzine is
"Don't know, can't say, haven't the	foggiest," said Joe.
I've no idea! Not the	foggiest!

The meaning is expressed, not by individual words, but by a variable lexicogrammatical pattern, in which no single word-form is essential. The lemma HAVE is followed by one of a small number of superlative adjectives (*faintest, foggiest, slightest*), plus a word for 'idea', usually *idea* or *notion*, or one of a few other nouns (*she hadn't the faintest doubt, no more than the faintest inkling*). With *foggiest*, even the noun is optional. The whole unit is either negative or hypothetical or contains a word such as *only*.

Phrasal units and cultural schemas

There are other phrasal units which express the familiar experience of 'incomprehension' or 'exasperation'. For example, the 'CANNOT FOR THE LIFE OF ME ' construction expresses irritation at not being able to understand something, often other people's unreasonable behaviour[4]. (The construction can have other pronouns and other psychological verbs.)

I can't	for the life of me understand what it is you see in it
I can't	for the life of me see what motive any of them can have
I can't	for the life of me see what that's got to do with you
I cannot	for the life of me see why children have to take so long
I cannot	for the life of me see why they're so resistant to it
I couldn't	for the life of me see what the old git was moaning about

Similarly, the 'WHAT'S X DOING Y?' construction encodes 'unexpectedness' or 'incongruence' (Kay and Fillmore 1999: 4).

what's that	doing in there? get it out!
what's he	doing here at this time of night?
what's he	doing phoning this time of day?
what's she	doing with a young man like that?
what's Ellie	doing all dressed up in Mother's clothes?

A major challenge for corpus lexicography is to discover such abstract extended phrasal units with their variants, and many other analyses have now followed the same approach, working bottom-up in order to discover constructions which express strong evaluative meanings: little schemas of cultural knowledge. These meanings are not evident to introspection, and are therefore not discussed in traditional speech act theory. They have to be discovered by empirical corpus analysis. Sinclair (1998) analyses the 'WOULDN'T BUDGE' construction, which signals frustration at trying to get something or someone to move, and failing. Channell (2000: 47–50) analyses the 'PAR FOR THE COURSE' construction, which signals that things have gone wrong, yet again, in just the way that you would have predicted. Stubbs (2007) analyses the 'NOT THE END OF THE WORLD' construction, which signals reassurance and sympathy for someone who has suffered some disappointment. Here are some attested examples:

– I tried to persuade him out of it but he wouldn't budge
– tough working conditions are often par for the course in catering
– it's disappointing but it's not the end of the world

4. I am grateful to Katrin Ungeheuer for pointing out this construction to me.

Cameron & Deignan (2006) discuss the emergence of the phrase *emotional baggage*, which has stabilized recently as a preferred way of expressing a culturally shared schema. The phrase expresses "a negative view of past emotions and memories" (2006: 679). The following examples are from the BNC.

> She couldn't help but sound cynical. "... It won't help, you know. You'll still carry that emotional baggage with you wherever you go, wondering what you said or maybe didn't say that frightened him off."
>
> No, he isn't the cause of my nightmares. My past is. It's the emotional baggage I'm hauling around that's causing all the trouble.

This use extends the tendency, over hundreds of years, for *baggage* to be used metaphorically in critical and pejorative ways:

> she was afraid the old baggage was going to start asking awkward questions (BNC)
>
> corpus linguistics offers a fresh start without the baggage that has accumulated over the years (Sinclair 2004a: 185)

Cameron & Deignan (2006: 678) take an explicitly constructivist view and regard this as a case of "an expression becoming fixed and a concept becoming delineated". This conventional way of expressing a critical opinion provides a label which does not refer to an objective thing in the external physical world, but to a subjective perception of the social world. The world is full of entities which exist only because speakers think they exist.

Work on speech acts is best known to linguists in Searle's (1969) version of the theory. However, the kinds of acts discussed in Searle's puritan and analytic world are invented, and often rather trivial: speakers ask each other to pass the salt, open the window and take out the garbage. By working bottom-up, a major discovery of corpus study has been extended lexical items which express much more complex and subtle evaluative speech acts. Unfortunately we are all familiar with the experience of irritation or incredulity when things go wrong, and we have conventional ways of expressing our feelings, but semantic categories such as 'incomprehension', 'exasperation', 'indignation' 'frustration' or 'world-weary despair' do not generally appear in descriptions of English.

Example: Phrasal units around verbs of perception

Two aspects of this work require development. First, no-one quite knows how to talk about such linguistic and conceptual units, and various metaphors are used. Units are said to crystallize, emerge, reify and stabilize. Second, a list of isolated examples does not amount to a theory. We have the problem again of

how to provide comprehensive coverage of the vocabulary of the language. An intermediate step would be to study a well-defined semantic set. For example, the five verbs of sensation and perception, FEEL, HEAR, SEE, SMELL, TASTE, have distinct grammatical and phraseological characteristics. They are not normally used in continuous *-ing* forms, and they are often preceded by a modal (*can* or *could*) which adds little or nothing to the meaning of the main verb (F. Palmer 1974: 117).

In addition, their literal meanings are "marginal to the way in which they are used" (Sinclair 2004b: 278). FEEL is frequently used in conventional expressions of disagreement (*I don't feel like it*) and other non-literal expressions (*get a feel for it*). HEAR is frequently used in conventional expressions of politeness in closing a letter (*I look forward to hearing from you*), or regret (*I'm sorry to hear*), or of rejecting an offer of help (*I wouldn't hear of it*), or to mean 'have recently learned' (*I hear he's got a new job*). SEE is most frequently used in the meaning 'understand' (*if you see what I mean*), and in conventional expressions of tentative agreement (*I don't see why not*) and surprise-cum-complaint (*I've never seen anything like it*). SMELL has many non-literal uses (e.g. *to smell a rat, smelling of roses*). TASTE is certainly used literally, but often in one text-type: recipes. Otherwise, it frequently means 'the ability or lack of ability to judge what is appropriate' (*good taste, bad taste, in poor taste, in the worst possible taste*), or in the extended sense of 'experience' (*a taste of things to come*, *his first taste of freedom*), or in idioms (*a taste of their own medicine*).

Several uses of FEEL provide further conventional ways of expressing uncertainty or tentativeness or surprise and/or lack of understanding of something:

But	I can't help	feeling that there must be more to it than that
But	I can't help	feeling the situation has intensified somewhat
but	I can't help	feeling the future looks black
But	I can't help	feeling that this is strictly a US product
But	I can't help	feeling we've missed all the really vital ones
yet	I can't help	feeling just a bit sorry for the Italians

	I got the	feeling something was going on. It was too quiet.
	a strange	feeling ... that something was going to happen
an	overwhelming	feeling that we must do something to stop it
	There was a	feeling in his bones that something was ...
I suppose it was the		feeling that we were doing something secret

That is, we have a set of high frequency words, whose primary use does not correspond to their literal meaning. Native speakers cannot generate comprehensive lists of such uses from introspection, but they immediately recognize them, in

concordance lines, as preferred and conventional ways of expressing evaluative pragmatic meanings. So, we need a more systematic method of discovering these longer expressions, their forms and their non-compositional meanings. (Stubbs 2007 provides a case study).

Summary of Francis' approach

In summary: This third approach is corpus-driven[5]. It holds the promise of being able to discover speech acts and cultural schemas which are entirely missed by the introspective data used in speech act theory. The examples from Francis and others illustrate one of the major theoretical proposals to come out of corpus studies: what Sinclair (1996, 1998, 2005) has called extended lexical units. In contrast to Williams' examples, which often come from intellectual discourse, Francis' examples are from everyday sociolinguistic acts. These examples now have to be approached from both ends: What are the non-compositional meanings of expressions which contain frequent words (such as FEEL)? And what are the conventional ways of conveying frequent speech acts (such as the expression of incomprehension or uncertainty)?

3. Part 2: The dualism of agency and structure

So far, I have discussed three approaches to keywords, with their strengths and weaknesses. The three approaches share the general idea that certain words and phrases convey meanings which are socially important, but they come out of very different traditions (cultural studies, quantitative corpus analysis and lexico-grammar), they are only loosely conceptually related, and perhaps only marginally compatible. Williams relates individual keywords and cultural concepts, Scott extracts sets of keywords from texts, and Francis shows how conventionally phrased speech acts express widely shared everyday values. What is missing is an explicit model of the relation between phraseology, speech acts, texts and text-types, and social institutions.

Institutions are abstract structures, which change historically, but which typically exist over long periods of time. They nevertheless depend on their constituent speech events, which in turn depend on speech acts which last for only seconds. The problem is to relate things of very different scales in time and space, and the

5. As far as I can determine, Francis (1993: 137, 139) is the first reference in print to a "corpus-driven" and "data-driven" approach to linguistic analysis. Francis distinguishes this approach from the use of a corpus merely to find examples for a theory which has been independently formulated.

main ontological nightmare involves the dualism of agency and structure. Trying to work out the ontology of social institutions has provided "problems in the foundations of the social sciences" (Searle 1995: xii) for around 150 years. Searle (1995: 3–4) admits that he can hardly bear the metaphysics involved in ordering a beer in a French café (well, that's his problem, but you can see his theoretical point). You sit down at a table, and utter a conventional French sentence, which expresses a predictable speech act; a waiter brings a beer, you drink it, utter another conventional speech act or two, pay the bill and leave. It hardly occurs to you that you must know about speech acts (such as question and request), and also about money, property, ownership (the waiter doesn't own the beer but he sells it to you), and so on. These aspects of social reality are not physical facts, but seem just as robust and objective (Searle 1995: xi and passim; Collin 1997; Miller 2007).

Although we tend to think of institutions such as Parisian cafés, churches, universities and all the rest as places, they are better regarded as placeholders for patterns of activities (Searle 1995: 57). This is what gives us the opening we need to relate language use and social institutions. Social institutions and speech events are the same thing looked at from different points of view[6]. But since Searle does not use any empirical language data, he does not discuss texts and text-types. Corpus studies can fill this gap, by providing empirical data on the lexical patterns which make up phrases, texts and text-types. And text-types make up the activities which are referred to in a shorthand way as universities, churches and all the rest. Corpus studies can document how such an emergent model works.

3.1 Texts in society: Some very banal observations

The following remarks are hardly original, but they are, I hope, obvious. That is, I hope you agree that they are simply obviously true and even banal, because I will just state them, then take them for granted, and use them to introduce some points which are certainly not obvious, in so far as they have provided problems in the foundations of the social sciences for the last hundred and fifty years or so, and which remain unsolved. I will then propose that corpus methods can provide a new way of looking at these old puzzles.

There is an inherent and logical relation between social institutions, the professionals who work in them and their clients, and the language which is used there. The rough idea can be illustrated as follows.

---

6. This way of thinking about things has been proposed by Halliday in his analogy of the relation between weather and climate (Halliday 1991).

Scientists	write research papers	for their peers	in specialist journals.
Professors	give lectures	to students	in universities.
Preachers	give sermons	to congregations	in churches.
Doctors	give consultations	to patients	in doctors' surgeries.
Employers	ask questions	of potential employees	in job interviews.
MPs	give speeches	to other MPs	in parliament.
Journalists	write editorials	for readers	in newpapers.
Judges	pass sentences	on the accused	in courtrooms.

The social roles are interdependent: a preacher cannot give a sermon unless there is (at least) a (potential) congregation; a doctor can only give a consultation to a patient; there are no judges without defendants. And just to make this stunningly banal point even more obvious, you don't find other combinations of speakers, speech events and social institutions. You don't find scientists giving sermons to patients in job interviews. We have what seem like natural rules which determine which combinations are possible or not. These rules are not regulative: it is not that scientists go around giving sermons and other people tell them not to. They are constitutive rules which define how society works.

It would be a major undertaking to make such a list anything like comprehensive. However, the list could not be continued indefinitely, because the social world is structured around such speech events. Indeed, if we could list and describe all such speech events, we would have defined the culture. A Hallidayan formulation would be that social reality is an edifice of meanings; the network of meaning potential is what we call the culture. There is a range of social institutions, which are staffed by professionals. Part of their professional communicative competence requires them to engage in particular speech events, with their peers and with their clients. These speech events can be described in terms of their conventionalized speech acts, which are realized by words, phrases, discourse structure and so on.

The model seems most obviously applicable to public and professional spheres, but similar statements can also be made about private spheres: members of the family, friends, acquaintances and neighbours chat, argue and gossip with each other, and express their frustration and incomprehension (so, of course, do doctors and professors). One main difference to the public sphere is that anyone can be a neighbour, but not anyone can be a doctor. Another difference is that professionals may be authorized and/or required to carry out certain speech events (e.g. give lectures every Monday morning). It may be advisable to wish your neighbours "good day" when you see them, but you do not have to, and you do not require special authorization to do so.

When I put these points to a group of students recently, one student objected that the model doesn't work with, for example, novelists. But it seems to me that the student had in mind a rather dubious notion of the free and creative artist, who works outside social conventions. Novelists write novels for readers who have bought their work, and this is only possible within a set of conventions and institutions. The whole institution of literature, in the sense of fiction, is historically quite recent: for example, *Robinson Crusoe* is often discussed as the first novel in English. And literature relies on other institutions, including publishers, not to mention university courses which define the conventional canon.

A more detailed discussion would have to distinguish between at least three senses of the concept of social institution. For example, with respect to medicine, it can refer to a specific place (e.g. hospital, clinic, surgery, etc), or to a whole social system (e.g. the National Health Service) or to a body of knowledge or an academic subject. (This third sense is the one assumed by Scott & Tribble 2006: 82–3.) The same applies to other institutions such as the law and theology, in which people receive a professional training. We therefore find educated groups who have the communicative and textual competence to act in these areas, including command of a technical vocabulary and its phraseology and speech acts. It is the established members of an occupational group or discourse community who create and maintain the genres (Swales 1990). Distinctions are also necessary between different professionals and lay persons (e.g. junior doctor, consultant, nurse, surgeon, patient, etc.), and between different speech events (e.g. consultation, diagnosis, case history, prescription, etc). We might also have to identify specific texts (e.g. the Hippocratic Oath; and an oath is, of course, a speech act). But in principle, we could construct a list such as the following.

In religious institutions, including (Christian) churches (Catholic, Church of England, etc), priests give sermons and baptize and marry people, and together with members of the congregation, they sing hymns, say prayers, and so on, often using specific texts (such as the Bible, the Book of Common Prayer, the marriage ceremony, etc), to which explicit reference is often made in conventional phraseology (*Our reading today is taken from ...*). A church service (speech event) consists of a sequence of hymns, sermon, prayer, announcements (and other speech acts).

In a legal institution, the courtroom, witnesses and defendants swear to tell the truth, they are cross-examined by barristers, and the judge gives a summing up. It is at this level of speaker roles, speech events and speech acts that people understand such social events. Here, for example, is a short extract from the UK Government site on the Criminal Justice System in England and Wales, which is designed to explain to potential jurors their responsibility and what happens in

court[7]. Almost the whole extract consists of explaining the relation between the participants, the speech events and the speech acts, which I have underlined.

> "Once all 12 jurors are in the jury box ... the court clerk will call out each name and each member of the jury will be sworn in. They must either take an oath on a holy book of their choice or they must affirm. This is similar to swearing in, but without the holy book. ... All criminal trials follow similar procedures. A defendant or number of defendants will have been accused of a crime. The prosecution advocate opens the case by explaining the accusations and setting out the facts they will seek to prove during the trial. Witnesses for the prosecution will be called. They take an oath, or affirm, to tell the truth and are then questioned and cross-examined. Next, witnesses for the defence may be called. If they are, they too will take the oath, or affirm, and be questioned and cross-examined. ... When all the evidence has been given to the court, the prosecution and defence advocates may make their closing speeches. They will talk directly to the jury as they argue their cases. The judge will explain the law and summarise the facts of the case. Then he or she will clarify the duties of the jury before they go to the jury room to consider their verdict."

Much of the phraseology in churches and courtrooms is largely fixed, for traditional and/or legal reasons. This is less so in many other institutions, but the same points hold in general. Media institutions (e.g. radio, television, newspapers) have presenters, news readers, journalists and so on, who take part in and/or write news broadcasts, documentaries, editorials and so on. Scientific and technical institutions (e.g. universities, research laboratories) have scientists and researchers who write lab reports, publish articles and so on.

There are few studies which use empirical corpus data to show the micro-macro relations across this range between the linguistic features of text-types and social institutions. However, Atkinson (1999) provides an exemplary case study of the development of a scientific journal and its research articles (text-type) within the Royal Society of London (social institution). He used a sample of texts from seven 50-year intervals between the late 1600s and the late 1900s, in order to study the development of an area of institutionalized knowledge along with its changing norms of language use. The discourse community of The Royal Society is its scientists, who were initially gentlemen amateurs, and only later the professional

7. The website is at http://www.cjsonline.gov.uk/juror/the_trial/index.html (accessed 15 December 2007). A related web-page lists other participants and the speech acts they perform, including Clerk, Crown Prosecution Service Representative, Expert Witness, Probation Representative, Usher, Defence Representative, Dock Officer.

scientists which we know today. Their journal was *The Philosophical Transactions of the Royal Society of London*. The content of the journal evolved from the form of polite letters, often narratives of observations, which relied on the trust and authority of their gentlemen authors, into experimental reports, with a highly conventionalized structure of theory-methods-discussion, with explicit sub-headings, which is so familiar in modern scientific articles. As Atkinson (1999: xvi) puts it, "the evolution of these forms of meaning" from person-centred to object-centred discourse, was "an integral part of the changing form of scientific life".

Atkinson studies things both top-down and bottom-up (p. 56). He studies how a particular text-type developed over time within the institution: this gives the macro perspective. And he studies the rhetorical features of individual texts: this gives the micro perspective. In addition, he uses both qualitative and quantitative methods: traditional rhetorical analysis of the discourse organization of the articles, and also multi-dimensional analysis of the significant co-occurrence of conventionalized linguistic features.[8]

3.2 Puzzles of social theory

We could not operate in the social world if we did not take such things for granted. A social theory of language has to deal with the following relations:

(1) Objective behaviour and subjective meaning. For example, sitting down and drinking a beer versus entering a café, ordering the beer and paying for it. For most of the time, human beings cannot be wrong about what they are doing. If you go into a café and issue a speech act, such as ordering a beer, you must understand what you are doing, you must be doing it intentionally, and if you do not do it in conventional ways you risk causing, at least, confusion.

(2) Scales of time and place. We need a model which explains the relations between large macro structures which exist over long periods of time (social institutions) and small micro events which last a few moments (speech acts and their conventional idiomatic expressions).

8. Miller (2007) points out that the Alan Sokal case shows how formal features of a text-type may fool some of the people some of the time, but cannot fool all of the people all of the time. Sokal, a professor of physics, wrote an article which had all the stylistic features of an academic discussion of post-modernism, and submitted it to a peer-reviewed journal, where it was published (Sokal 1996). These aspects of the world (academic journal, peer-reviewing, etc.) are institutional facts. However, as Sokal then admitted, the content of the article was gibberish and he had written it as a hoax. Style had triumphed over content, but only in the short term.

(3) Structure and agency. The model must also explain how structure and agency interact. They seem to be different kinds of things: one is not reducible to the other, but both are real.

(4) Cause and effect. You can become a doctor only if the required social institutions exist. These institutions create expectations, but it is the agency of the participants, including their speech acts (e.g. giving advice) and speech events (e.g. a medical consultation) which creates and maintains the institutions. The conventions and the institutions are independent of any individual person's activities, but dependent on the cumulative behaviour of the speech community. Corpus methods are good at describing recurrent behaviour across groups of speakers.

The overall model combines structure, knowledge and agency. The social institutions provide the structure. The speakers have the knowledge (communicative competence). The speech events and speech acts are the intentional behaviour of the agents. The conventional phraseology is part of the linguistic system. The evidence for the speech acts and their phraseology is the recurrent textual traces which can be studied in large corpora.

The social institutions are abstract structures, which depend on agency. A university is something which happens. It exists because teachers and students engage in particular kinds of language behaviour which creates conventional social relations between them. In addition, part of being a university, a church, a court of law and so on is being thought to be one, and these social institutions have had their statuses assigned to them by constitutive rules (Searle 1995: 34, 50).

The professionals in such institutions are people with the communicative competence to utter the appropriate speech acts in the conventional way in the required speech events. In addition, they are social groups of a particular kind (Sealey & Carter 2004: 111). You can only belong to a category such as doctor or priest intentionally, and you have to be authorized to have such a status[9]. Having such a status confers rights, obligations and powers, including the power to issue certain speech acts: professors can fail students, priests can marry people, employers can employ you.

The text-types often have everyday names (e.g. church sermon, news broadcast), and these things are also intentional events. You cannot give a church

9. This is quite different from other categories which sociologists use, such as "unemployed men over 50". It is also quite different from a category such as "patient" or "member of the congregation" (which anyone can be), and from a social category such as "adolescent" (which is not voluntary and has no necessarily associated conventions, even if we have a stereotype of how adolescents behave). It is because these statuses have to be authorized, that people can pretend to be doctors or professors when they are not so authorized (Searle 1995: 48).

sermon without intending to. But there is no systematic and comprehensive classification of such things. Various classifications of text-types have been proposed, but they seem only to work as general dimensions or ideal types, based around communicative functions such as informative or persuasive, or around structural types such as descriptive, argumentative or narrative. Most real texts are mixed.

Having proceeded top-down, we now arrive back at texts, which are the only thing which corpus linguists can observe: the traces which are left by these activities. Technology has profoundly changed the traces which people leave and how these traces can be analysed.

I assume that Searle's view of these things is basically correct. He proposes that language plays a special role in institutional reality, and that the social world has a hierarchic structure which explains the relation between speech acts and social institutions. A certain kind of speech act is a promise. A certain kind of promise is a contract. A certain kind of contract is a marriage vow. Only certain kinds of people are authorized to perform marriage ceremonies, because they themselves have entered into other kinds of contracts and have been authorized by other people who are authorized to authorize them. Though, oddly enough, I'm not sure whether Searle says explicitly that different kinds of promises and contract formation are the most basic trait which pervades social behaviour (Wilson 1998: 189).

Although social institutions are complex, they are the result of simple rules which are applied recursively. This is Searle's (1969) proposal: speech acts depend on constitutive rules and social reality is constructed recursively with speech acts at the bottom level (Searle 1995). However, since Searle does not use any empirical language data, he does not discuss texts and text-types. Corpus studies can fill this gap by providing empirical data on the lexical patterns which make up phrases and texts. The predictable co-occurrences of patterns realize text-types, and text-types constitute the activities which are referred to in a shorthand way as universities, churches and all the rest. Corpus studies have shown how such an emergent model can work, and corpus methods provide the possibility of describing complex systems by tracing causation across many levels (Wilson 1998: 207), from phraseology and speech acts to social institutions.

4. Concluding comments

This article is in two distinct parts: part 1 was about keywords and argued bottom-up; part 2 was about social institutions and argued top-down. This is the problem with the concept of keywords: the large gap between individual words and the social world. The term is used in different senses, which are related in only a loose way. It may be a productive concept, but it cannot stand on its own. It assumes

other concepts, such as cognitive schemata, textual collocates and semantic fields, text and text-type. Concepts of text and text-type in turn imply the concept of social institutions.

So how do we get back to keywords? Quantitative corpus data provide evidence of semantic units (extended phrasal units) and thereby extend the empirical basis of speech act theory. This gives us the basis of a theory of language as social action, of the relations between language use and language system, and of the relations between phraseology, texts, text-types and social institutions. There is a series of questions in linguistics which are all logically the same question. How does something arise from nothing? How do extended phrasal units of meaning arise from recurrent collocations? How do social institutions arise from recurrent speech events? How do structures arise from agency? How does the macro arise from the micro? How do the properties of whole systems emerge? The answer is: by recursive application of constitutive rules. Social institutions and text-types imply each other: they are different ways of thinking about the same thing.

Speech act theory asks the right questions, but does not have the data or methods to answer them. It tries to do ordinary language philosophy without attested data on ordinary language. Corpus linguistics has the data and the methods, but has not yet co-ordinated studies in a way which can answer cognitive and social questions. It has not yet moved from description to explanation. If this line of argument can be worked out successfully, it will show how corpus data and methods can help to solve puzzles in the foundations of the social sciences.

Acknowledgements

For valuable comments on a much earlier version of this paper, I am grateful to Kieran O'Halloran and to lecture audiences in Italy in June 2007, at the Catholic University of Milan and at the Certosa di Pontignano, University of Siena.

References

Atkinson, D. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Mahwah NJ: Lawrence Erlbaum Associates.

Benveniste, É. 1954. Civilisation: Contribution d'un mot. Reprinted in *Problèmes de linguistique générale*, 336–45. Paris: Gallimard, 1966.

Cameron, L. & Deignan, A. 2006. The emergence of metaphor in discourse. *Applied Linguistics* 27(4): 671–90.

Channell, J. 2000. Corpus-based analysis of evaluative lexis. In *Evaluation in Text*, S. Hunston & G. Thompson (eds), 38–55, Oxford: OUP.

Collin, F. 1997. *Social Reality*. London: Routledge.
Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society*, 36–72.
Firth, J. R. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis,* 1–32. Special Vol., Philological Society. Oxford: Blackwell.
Francis, G. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In *Text and Technology,* M. Baker, G. Francis, & E. Tognini-Bonelli (eds), 137–56. Amsterdam: John Benjamins.
Friedrich, F. 2007. A Corpus-based Study of British Newspapers. Staatsexamensarbeit, Universität Trier.
Hacking, I. 1986. The archaeology of Foucault. In *Foucault: A Critical Reader,* D. C. Hoy (ed.), 27–40. Oxford: Blackwell.
Halliday, M. A. K. 1991. Corpus studies and probabilistic grammar. In *English Corpus Linguistics*, K. Aijmer & B. Altenberg (eds), 30–43. London: Longman.
Herberg, D., Steffens, D. & Tellenbach, E. 1997. *Schlüsselwörter der Wendezeit*. Berlin: de Gruyter.
Kay, P. & Fillmore, C. J. 1999. Grammatical constructions and linguistic generalizations: The what's x doing y? construction. *Language* 75(1): 1–33.
Ladendorf, O. 1906. *Historisches Schlagwörterbuch*. Berlin: K.J. Trübner.
Lepp, F. 1908. *Schlagwörter der Reformationszeit*. Leipzig.
Mason, O. & Platt, R. 2006. Embracing a new creed: Lexical patterning and the encoding of ideology. *College Literature* 33(2): 155–70.
Matoré, G. 1953. *La méthode en lexicologie*. Paris: Didier.
Miller, S. 2007. Social institutions. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/social-institutions> (10 December 2007).
Palmer, F. 1974. *The English Verb*. London: Longman.
Scott, M. 1998. *WordSmithTools*. Version 3. Oxford: OUP.
Scott, M. & Tribble, C. 2006. *Textual Patterns* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.
Sealey, A. & Carter, B. 2004. *Applied Linguistics and Social Science*. London: Continuum.
Searle, J. R. 1969. *Speech Acts*. Oxford: OUP.
Searle, J. R. 1995. *The Construction of Social Reality*. London: Allen Lane.
Sinclair, J. 1996. The search for units of meaning. *Textus* 9(1): 75–106. Also in Sinclair 2004a, 24–48.
Sinclair, J. 1998. The lexical item. In *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 271], E. Weigand (ed.), 1–24. Amsterdam: John Benjamins. Also in Sinclair 2004a, 131–48.
Sinclair, J. 2004a. *Trust the Text*. London: Routledge.
Sinclair, J. 2004b. New evidence, new priorities, new attitudes. In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.), 271–99. Amsterdam: John Benjamins.
Sinclair, J. 2005. The phrase, the whole phrase and nothing but the phrase. Plenary lecture, *Phraseology 2005*, Louvain-la-Neuve, October 2005.
Sinclair, J. 2007. Introduction. In *Text, Discourse and Corpora*, M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds), 1–5. London: Continuum.
Sokal, A. 1996. Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Social Text* 46–47 spring/summer 1996: 217–52.
Starcke, B. 2007. Korpusstilistik: Korpuslinguistische Analysen literarischer Werke am Beispiel Jane Austens. PhD dissertation, Universität Trier.

Stötzel, G. & Wengeler, M. 1995. *Kontroverse Begriffe: Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik*. Berlin: de Gruyter.

Strauß, G., Hass-Zumkehr, U. & Harras, G. 1989. *Brisante Wörter von Agitation bis Zeitgeist*. Berlin: de Gruyter.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. 2007. Quantitative data on multi-word sequences in English: The case of the word *world*. In *Text, Discourse and Corpora*, M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds), 163–89. London: Continuum.

Swales, J. M. 1990. *Genre Analysis*. Cambridge: CUP.

Teubert, W. 1989. Politische Vexierwörter. In *Politische Semantik,* J. Klein (ed.), 51–68. Opladen: Westdeutscher Verlag.

Teubert, W. 2007. *Parole*-linguistics and the diachronic dimension of the discourse. In *Text, Discourse and Corpora,* M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds), 57–87. London: Continuum.

Wierzbicka, A. 1997. *Understanding Cultures through their Key-words*. Oxford: OUP.

Williams, R. 1976/1983. *Keywords. A Vocabulary of Culture and Society*. London: Fontana.

Williams, R. 1985. Mining the meaning: Key words in the miners' strike. *New Socialist*, March 1985.

Wilson, E. O. 1998. *Consilience: The Unity of Knowledge*. New York: Knopf.

Problems in investigating keyness, or clearing the undergrowth and marking out trails…

Mike Scott
Aston University, UK

The article explores what might be meant by keyness in Corpus Linguistics. Keyness is a textual quality which is beginning to arouse interest, but as yet it is little understood and much exploratory work is needed. To this end, the chapter focuses in on a number of specific issues: the amount of text to take as a unit when computing keyness, statistical problems in making claims of different kinds about they keyness of words and phrases, the choice of an appropriate reference corpus, and the types of repetition which characterize key words. Many of these are illustrated by reference to Shakespeare plays. However, the overall aim is less to characterise the Shakespeare plays than to delineate the research area, to pin down the nature of keyness.

1. Introduction

The purpose of this article is to explore the notion of keyness and consider its implications in corpus-based study. Keyness, as a new territory, looks promising and has attracted colonists and prospectors. It generally appears to give robust indications of the text's aboutness together with indicators of style. Nevertheless, there are problems in identifying exactly what the keyness procedure comes up with and determining the boundaries of the research endeavour, and it is with these problems that we shall be chiefly concerned.

Keyness is a quality possessed by words, word-clusters, phrases etc., a quality which is not language-dependent but text-dependent. That is, in this view of keyness, words are not generally or simply key in a given language, but they may be key in a given text, or in a given set of texts, or in a given culture as in Raymond Williams' classic (1976) study which looks at words which characterise the intellectual culture of late 20th Century Britain, or more generally, Stubbs, (1996: 169): "The study of recurrent wordings is [..] of central importance in the study of language and ideology, and can provide empirical evidence of how the culture is

expressed in lexical patterns". As such, keyness theory suggests these wordings are prominent in some way in that text. Their prominence, their outstandingness (both metaphors suggesting being raised above the general level) may be useful in certain ways. It may lead us to perceive the aboutness of the whole text or of certain parts of it, or it may help us to perceive something about the style of the text which is different from styles of other texts.

Another way of thinking about it is via the metaphor "key" itself. A key grants access to a place which is otherwise restricted, private, sealed off. In this sense a word or phrase is a key which enables one to see something, it is an enabling device.

The nature of these prominences, these access routes, still needs exploration. We are beginning to see uses to which they may be put, too, but this aspect concerning applicability needs development as well, and that is the purpose here. Though the aim is to mark out some trails and chop away some tangled undergrowth, as your guide I can offer no guarantee of riches. The only certainty is that our exploration will bring us some scratches and bruises. We may get to see some interesting views, however, when we get near the top of our path.

2. Colonists and prospectors

Early work in this area has already been carried out. My own first paper of any substance was Scott (1997) in which I investigated the patterns of recurrence of key words (KWs) in a series of feature articles. Berber Sardinha (1999) was quick to follow as were Ku and Yang (1999) studying tourism. Kemppanen (2004) studies key words as indicators in history text; Seale and fellow researchers (e.g. Seale, Charteris-Black & Ziebland 2006) have done extensive work on medical key words. McEnery (e.g. 2009) and Baker (e.g. 2009) have worked a lot on keyness in relation to moral panics and similar social aspects of struggle.

Culpeper (2002) and Culpeper (2009) explore keyness in relation to *Romeo and Juliet*, and this literary purpose is also found in Chapters 4 and 5 of Scott & Tribble (2006). Chapter 7, written by Chris Tribble, looks at keyness and business communications and his Chapter 9 looks at journalism. Michael Toolan also (2004, 2006) has used key word analysis for literary purposes. Many of the sources of such work can be found at http://tinyurl.com/cve22m, which lists texts using *WordSmith* at the author's site; these are snapshots of some of the colonists and prospectors setting foot in the new keyness territory.

3. Issues

In this text, I propose to tackle certain specific issues and then to discuss the applicability of the KW procedure to the study of Shakespeare. The issues are

1. choosing to work with the text section or whole text or the corpus or sub-corpus
2. statistical questions: what exactly can be claimed?
3. how to choose a reference corpus
4. the handling of related forms such as antonyms
5. what is the status of the "key words" one may identify and what is to be done with them?

But first, let us think about the nature of the item we are concerned with, as unlike Raymond Williams we are detecting keyness using machine methods.

3.1 Machine and human KWS

Rigotti and Rocci (2002) warn that machine identification of key words omits all interpretation of the writer's intentions, cannot get at cultural implications and does not spot the congruity of the meanings of each section with the next.

> In our view, a natural language text, slippery and vague as it may be, is not a stone soup where words float free, tied only to their multiple associations within a Foucaultian discourse (Rigotti and Rocci 2002)

Despite the profusion of metaphors here, the general drift is clear. It is perfectly true that automatic analysis works differently from human identification in the case of keyness as it is in the case of current account banking, finding a date, tracing one's ancestors or indeed locating metaphors themselves (Deignan 2009). The computer does things quite differently and that means its results are different in quality and quantity. Of course it doesn't actually understand … or know what is "correct" and of course it can only look at what is found in text or context.

This is precisely the computer's strength. The human being cannot switch off his or her intelligence, we cannot see a constellation of stars merely as blobs of light but will end up seeing patterns with darker or lighter patches, noticing lines of stars, etc.; we cannot see text without seeing meaning, and this is why it is so very hard to spot typos. The computer's very blindness (Scott, forthcoming) can be seen like this:

> Human beings are very good at noticing visual patterns, and this is why WordSmith's lists and plot displays are useful. In effect, the chief purpose of the software is to take a pre-existing shape, the text, then mix it all up and sort it all out, showing it in a quite different order. The computer does not see any patterns, but the human user does – and then gets some sort of insight. (Scott 2008: 104)

A keywords detection procedure which produced exactly the same results as human readers would be problematic in a number of ways. To list the obvious seems unlikely to be of much advantage, at least in the case of texts which the human reader had plenty of time to read. But more fundamentally, the chief impossibility is that human readers do not consistently agree on the key words of a given text. There may be a tendency for certain words to be flagged up by most readers as key, but there are always plenty of others which some readers but not others feel to be key. Abraham Lincoln expressed a similar sentiment about pleasing the people[1].

Keyness is therefore somewhat subjective, anyway. But this need not be seen as a problem; it should be turned around and perceived as a positive advantage. A key word detection routine is of interest precisely because it will typically throw up some items which are unlikely for humans to notice.

4. Text section v. text v. corpus v. sub-corpus

One way of thinking about the differences between computing the key elements of a section of a text, versus a whole text, versus a sub-corpus or a whole corpus, is via this diagram:

SCOPE 1: a few words to left and right
SCOPE 2: the whole sentence
SCOPE 3: the paragraph
SCOPE 4: the story so far (up to *back*)
SCOPE 5: the section or chapter
SCOPE 6: the whole text
SCOPE 7: the colony of texts to which this one belongs
SCOPE 8: other related texts
SCOPE 9: the context of culture
EXTRA-LINGUISTIC SCOPE: where you are when you meet the text

Figure 1. Contextual scope (from Scott & Tribble 2006: 9)

1. Abraham Lincoln: "You can please some of the people all of the time, you can please all of the people some of the time, but you can't please all of the people all of the time". This quote is often attributed to the monk and poet John Lydgate (1370–1451) but I cannot trace the connection and find it most implausible as Lincoln's language is very far from 15th Century in form.

The text section covers scopes 1–5, the whole text level is level 6, and the corpus scopes 7 and 8. In principle a software procedure like that of WordSmith's KeyWords tool can determine the KWs of a text section such as a chunk at scope 3, 4 or 5, or a whole text, or a whole set of somehow-related texts.

But even these apparently simple scope levels are problematic. Is this "text" at level 6 with or without mark-up, images, sounds (as in the case of most corpora at the time of writing)? Also, what do we mean by section, chapter and other non linguistically defined categories – can we define them? And more crucially, is text itself mutating? Consider Figure 2 below. It shows not ordinary text of the kind that scholars have investigated for millennia, but instead small fragmentary notices.

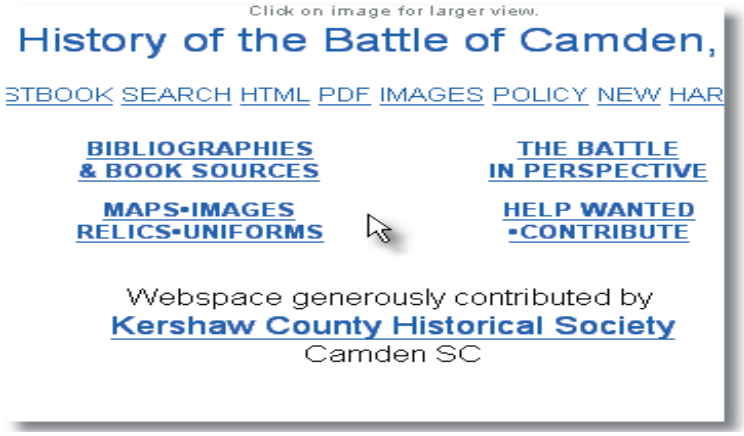


Figure 2. Internet text fragment

Part of the meaning of this new kind of text is the colour blue for hypertext links and the use of underlining, instead of the more prosaic linkage markers such as «click on image for larger view» as visible at the top of the figure.

Nowadays we are so familiar with web text that we no longer regard it as unusual, but it is worth reflecting on the fact that until such text sprang into widespread existence in the 1990s its predecessors, such as notices on bulletin boards and small ads in the window of local shops, were not typically studied as samples of text by students of language or of literature. On the contrary, text meant prose and prose meant paragraphs in pages. That is the underlying assumption behind figure 1, of course, that text scopes can go from small to large, still however disregarding this new and strange mix, a smörgåsbord of offers, instructions, exhortation (*help wanted*), alternative format choices, acknowledgements and hints.

Almost any page of the well-known site Wikipedia will exemplify similar issues. There is a mix of discourse acts in view, several pieces of prose interspersed

with numerous bullet-pointed lists and the separately boxed-off sections forming something like Hoey's classical 1986 'colony text'. At the same time behind the texts is a set of tabs allowing access to a wealth of related matter.

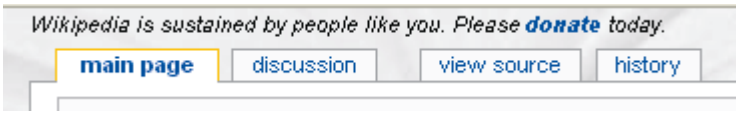


Figure 3. Wikipedia tabs

Log in as a contributor and the view changes:

Figure 4. Wikipedia tabs after logging in

which implies that text now, for the first time in history, varies its form automatically depending on the status of the reader. Text is indeed mutating!

A conclusion for this first section of the investigation would be that the KWs of text depend on text and that as text is a shifting sand, the set of KWs generated on it is also varied.

5. Statistical issues

The next issue to concern us has to do with what can be reasonably claimed when one has generated a set of KWs. The well-established *p value* has been a useful statistical standard for more than a century. It relies on comparing two numbers: the result actually obtained in one's study, and the result which would be obtained if all the results were shared out as equally as possible, that is to say a purely chance outcome. These two are compared and the result is an estimate of the risk of claiming that one's actual obtained result is significant.

In the case of a key word procedure such as that used in WordSmith, this calculation is repeated for every single type in the text we are interested in. For example, the frequency of THE in the text is compared with the frequency of THE in the reference corpus, and the p value is then computed of any difference. If the text has 9% of THE and the reference corpus has only 5% of THE, say, we might get a p value suggesting that we can believe, with little risk of being wrong, that in our text THE is prominent. This process is repeated with the frequencies of WAS, the frequencies of IS, and so on until all word-forms have been examined.

However, if you run lots of comparisons some will spuriously (by chance) appear significant. In the same way, if you throw a normal coin thousands of times, there can sometimes be sequences which you would think very unlikely, such as it landing heads up fifteen times consecutively.

Secondly, text itself doesn't consist of randomly ordered words; words crop up with strange frequencies. There are a small number of very high frequency words like THE, OF, WAS, and an enormous number of words cropping up only once or twice. The representation of this obeys a power law first suggested by George Zipf in 1949. See Scott & Tribble (2006: 28) for a plot showing that the words in a large word-list like that of the BNC obey a power law. A well-known analogy is the blockbuster versus long tail in sales of, for example, movies. There are a few blockbusters which dominate the listings, then a huge long tail of other movies which one cannot get to see in a regular cinema because they attract too few people, but which would still attract a few people, as one can find out from DVD sales from online stores.

The estimation as to what is likely, which our statistical process requires, is itself quite problematic. How likely is it for a text in English with at least 200 words to contain no cases of THE, for example? How can we assess its likelihood? (I have in fact encountered such a text, and was surprised to do so. It turned out to be a newspaper account of election results.) Although THE is far and away the most frequent word in most English, it might not be the most frequent word at all in texts like the ones in Figures 2–4 consisting mainly of hypertext links to other texts. There is a serious difficulty in trying to think about how likely any difference is.

Think also of much less frequent words. How likely is it that a word like *Zanzibar* will occur in text? There are 67 in the British National Corpus, but they come from only 24 texts of the many thousands in the BNC. Suppose a text we are processing contains two references to *Zanzibar*, that will be many times more frequent than our BNC reference corpus suggests is normal, and the machine-computed calculation of prominence will suggest the word is a KW. But what was the likelihood of the BNC itself containing any references to *Zanzibar*? If those 24 texts had not been included in the BNC there would have been none, of course. When we get down to items which occur only say twice or three times in the (one hundred million word) BNC, such as CARJACKED (3), or CARDIOGRAPHER (2), what was their likelihood of appearing at all in the BNC? CARDIOGRAPHER came in only one text, presumably a medical text. It is quite problematic trying to reach a human assessment of likelihood when one is dealing with very low-frequency items.

There are two more problems to be aware of. As explained above the KW procedure computes a likelihood for each word-type in the text we are studying.

Each one gets a different p value. The procedure does not compute the likelihood of getting the *set* of KWs it finds, however, so one cannot make any statistically-based claims about the status of the whole set.

5.1 The farmer and his crops

Finally, there is a problem in thinking about the order of the items in the set. Comparing p values is problematic. The most straightforward way of explaining that is via an armchair example. Suppose we are computing the KWs of a text about a farmer who is growing three crops, namely wheat, barley and oats. He suffers from three problems, namely hail, drought, and gales. The text deals with how he manages and what he does to cope and dedicates roughly the same amount of attention to each crop and each problem. They each get 4 mentions in the text. The frequencies of these in the BNC are as shown below.

Table 1. BNC and text frequencies of 3 crops and 3 blights

	BNC frequency	Text frequency
wheat	1002	4
barley	584	4
oats	288	4
hail	329	4
drought	683	4
gales	194	4

Ignore for present purposes the possibility that HAIL might not only be ice falling from the sky but also a Roman greeting, and GALES in the BNC might sometimes be of laughter. The computer program will ignore this too….

We are assuming for this armchair experiment that in the text these are equally key, i.e. equally important, and that all these items get mentioned an equal number of times. However, a computational procedure based on their frequencies in a large reference corpus will place GALES as the top KW, followed by OATS, HAIL, BARLEY, DROUGHT and WHEAT at the bottom of the list. The order of KWs is not intrinsically trustworthy, because it depends not only on the frequency in the text we are studying (4 times) but also on their frequencies in the reference corpus.

These difficulties do not mean that the KW procedure is not useful. They do mean though that one must take the output of this procedure with discernment and discretion.

Conclusions for this second issue are thus that there is no statistical defence of the whole set of KWs, but only of each one, though the more there are the higher

the chance that some of the comparisons came up by chance, and that it is not certain that the order of the items in the set itself reflects their importance. The implication of those conclusions is that KWs are pointers which suggest to the prospector areas which are worth mining but they are not themselves nuggets of gold.

6. Choosing a reference corpus

The choice of reference corpus is likely to be important for the KW procedure. However, the use of quite stringent p values (such as .00001 which means a risk amounting to one part in a hundred thousand) has led me to conclude after investigating with various different kinds of reference corpus (Scott 2006 Chapter 5, and Scott 2009), that using a mixed bag RC, the larger the RC the better but that a moderate sized RC may suffice. The keyword procedure is fairly robust, and KWs identified even by an obviously absurd RC can be plausible indicators of aboutness, which reinforces the conclusion that keyword analysis is robust. That is to say, there is a set of common KWs identified both by a plausible and by an implausible RC; the implausible one will also throw up some additional (and probably implausible) KWs.

Genre-specific RCs identify rather different KWs. Scott and Tribble (2006: 80–86) found that speech and writing gave rise to noticeably different KKWs ("key key words", that is KWs found in numerous texts) and that differences between academic and non-academic kinds of text from the same domain gave rise to different kinds of "associates" (KWs found in the same texts as a given KW).

For this section we can conclude that the aboutness of a text clearly need not be one single aboutness, it is numerous different aboutnesses and of course the same applies even more to collections of texts grouped in genres and sub-genres. This brings us back to the issue of the processing of single texts versus groups of texts discussed earlier.

7. Related forms

The issue of lemmas and other related forms is complex. Human readers can easily see that *grow* and *grew* are forms of the same verb, that *she* refers to a woman previously mentioned in the text or obvious from the context, and that *huge, massive, tiny* and *big* are related words in that they all describe extremes of size. Software such as *WordSmith* may be able to treat members of the same lemma as related if it is taught to, for example with a look-up list which enables automatic

transformation of a word into its base form. Similarly, software can be made to handle clusters (Biber & Conrad's 1999 *lexical bundles*) but otherwise ignores relations such as synonymy and antonymy.

For the next few years, key words procedures such as those discussed here are not likely to get round these restrictions. Processing word- and letter-forms is not at all the same as understanding, and distinguishing between *May* as a month and as a person's name, or *may* as a modal auxiliary versus *may* the flower is not a problem which is going to be reliably solved soon.

A conclusion for this section is that the KW procedure cannot yet handle related forms, but that it would be desirable for it to do so.

8. Status of the KW

The status of KWs, then, from what we have seen above, has certain restrictions and limitations.

a. Keyness, unlike say antonymy, is not intrinsic to the word or cluster itself but is context-bound.
b. The size of context is a matter of choice. More investigation is needed of KWs from different sized contexts.
c. KWs act as pointers to specific textual aboutnesses and/or styles. The reference corpus affects this in ways which are still not fully understood.
d. They are statistically arrived at but are not fully and completely established. Definitive status for a whole set of KWs cannot be claimed.
e. The procedure is far from being able to handle related forms which readers can easily distinguish.

To illustrate some of these conclusions, let us now examine some examples of key words, using Hamlet as our source.

9. Shakespeare's KWs

9.1 Hamlet

In this section, we will first examine the KWs of *Hamlet* as identified using default settings in *WordSmith* and using all 37 plays (OUP edition of 1916) as the reference corpus. The KWs identified fit into a number of groups as can be seen in Figure 5 below.

Characters:

FORTINBRAS, GERTRUDE, GUILDENSTERN, HAMLET, HAMLET'S, HORATIO, LAERTES, OPHELIA, PYRRHUS, ROSENCRANTZ

Places:

DENMARK, NORWAY

Themes, events:

MADNESS, PLAY, PLAYERS

Pronouns:

I, IT, T, THEE, THOU

Other ("unexpected"):

E'EN, LORD, MOST, MOTHER, PHRASE, VERY

Figure 5. KWs of Hamlet, by group

Most of the first three sets are obvious and probably uninteresting…. if you know the play you already know it concerns Hamlet and some other characters, that it's set in Denmark, that there is a play within the play and that Ophelia goes mad. However, some of the item-sets are puzzling. Why are IT, LORD and MOST positively key in Hamlet? They are negatively key in some of the other plays; what is special about Hamlet that makes them so prominent, so over-represented in this play?

Another question one might ask is, which characters are they most key of? And where are they found, how are these KWs dispersed throughout the play?

Let us first take the case of the word IT. I am assured that this short unobtrusive pronoun has never really been noticed by the myriad scholars who have worked on the play. In the whole set of 37 plays by Shakespeare *it* or *'t* occur about once every 100 words, i.e. 1% of all running words are *it* or *'t*. In Hamlet, however, they amount to over 1.5%. Here, this little word is cropping up 50% more often! In the case of the character Hamlet himself, it is 1.48%, roughly typical of the whole play, but in the case of Horatio, 2.33%: nearly 250% of the average in this one character's speeches.

In Hamlet's speeches, IT is distributed evenly as is shown in Figure 6:

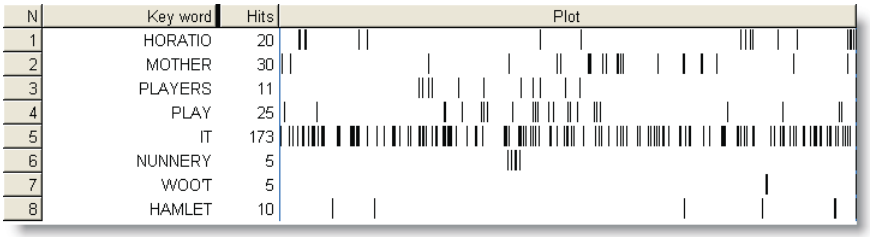


Figure 6. Plot of IT in Hamlet's own speeches

Some of Hamlet's own individual KWs such as *nunnery* are very localised and the mentions of the *play* and *players* are centred roughly within the middle sections of the play, but clearly IT is spread throughout quite globally.

Horatio's mentions of IT are much less global, by contrast, as can be seen in Figure 7.

Figure 7. Plot of IT in Horatio's speeches

I am not yet in a position to explain these apparently interesting facts, unfortunately. To do so will require more knowledge of the play than I can currently muster, so for the present all I can promise is that with a colleague who is an expert on the play, this issue will get further treatment. It is easy to concordance these cases of *it* but not to easy to see why Horatio's *it*s behave differently from those of Hamlet. How does Horatio manage to avoid the word for two considerable chunks of the play? (At default settings, *'t*, which he uses 6 times, is not key in Horatio's speeches, but the *it*-gaps are not full of *'t*s, so there is not a simple answer here.)

9.2 DO in Othello

Finally, let us examine the KW *do* in Othello. It occurs nearly twice as frequently as in the other plays, and is characteristic of Iago (who uses it nearly twice as often) and Desdemona (more than 3 times as often), as seen in Figures 8 and 9.

DOST is characteristic of Othello (more than 6 times as frequent).

	Concordance
1	<IAGO> Do thou meet me presently at the
2	knows you not. I'll not be far from you: do you find some occasion to anger
3	time, man. I'll tell you what you shall do. Our general's wife is now the general:
4	vow I here engage my words. <IAGO> Do not rise yet. Witness, you ever-burni
5	out to savage madness. Look! he stirs; Do you withdraw yourself a little while, He
6	speak with me; The which he promis'd. Do but encave yourself, And mark the
7	mind again. This night, Iago. <IAGO> Do it not with poison, strangle her in her
8	him so That I may save my speech. Do but go after And mark how he
9	I am none such. <IAGO> Do not weep, do not weep. Alas the day! <EMILIA> Has
10	I am sure I am none such. <IAGO> Do not weep, do not weep. Alas the day!

Figure 8. Commanding DO in Iago

If Iago's conspiratorial organising is signalled by the KW *do*, Desdemona is much more conditional (Figure 9):

Concordance

11	warrant of thy place. Assure thee, If I do vow a friendship, I'll perform it To the
12	go seek him. Cassio, walk hereabout; If I do find him fit, I'll move your suit And seek
13	tears, my lord? If haply you my father do suspect An instrument of this your
14	and ever did, And ever will, though he do shake me off To beggarly divorcement,
15	Good faith! how foolish are our minds! If I do die before thee, prithee, shroud me In
16	tell me, Emilia, That there be women do abuse their husbands In such gross

Figure 9. Conditional DO in Desdemona

Othello, on the other hand has DOST as a KW, expressing questioning and suspicion as can be seen in Figure 10.

Concordance

1	Ha! I like not that. <OTHELLO> What dost thou say? <IAGO> Nothing, my lord:
2	I love you. <OTHELLO> I think thou dost; And, for I know thou art full of love
3	thy brain Some horrible conceit. If thou dost love me, Show me thy thought.
4	for aught I know. <OTHELLO> What dost thou think? <IAGO> Think, my lord!
5	My noble lord,— <OTHELLO> What dost thou say, Iago? <IAGO> Did Michael
6	He did, from first to last: why dost thou ask? <IAGO> But for a
7	thought Too hideous to be shown. Thou dost mean something: I heard thee say
8	meditations lawful? <OTHELLO> Thou dost conspire against thy friend, Iago, If
9	to me as to thy thinkings, As thou dost ruminate, and give thy worst of
10	know my thoughts. <OTHELLO> What dost thou mean? <IAGO> Good name in
11	but keep 't unknown. <OTHELLO> Dost thou say so? <IAGO> She did
12	Farewell, farewell: If more thou dost perceive, let me know more; Set on
13	My noble lord,— <OTHELLO> If thou dost slander her and torture me, Never
14	you not hurt your head? <OTHELLO> Dost thou mock me? <IAGO> I mock
15	most cunning in my patience; But—dost thou hear?—most bloody. <IAGO>
16	And nothing of a man. <OTHELLO> Dost thou hear, Iago? I will be found most
17	t on the tree. O balmy breath, that dost almost persuade Justice to break her
18	in 's hand. O perjur'd woman! thou dost stone my heart, And mak'st me call

Figure 10. Othello's suspicious DOST

These patterns of KW occurrence in Shakespeare are like butterflies in the forest. Some of the time it seems fairly straightforward to decide that they are prominent because there is suitable nectar or more sunlight, but some of the time more in-depth understanding of ecology is needed.

10. Conclusion

Though we have explored some of the territory, there is clearly much more to do and see. The identification of KWs, especially ones which otherwise readers pass by unnoticed, can lead to perceptions but can also lead to the discovery that much

more needs to be understood. KWs are pointers, that is all. As such the interesting ones which would not otherwise get any attention merit chasing up and tracking down. There is so much more to do.

References

Baker, P. 2009. 'The question is, how cruel is it?' Keywords, fox hunting and the house of commons. In *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*, D. Archer (ed.), 125–136. Farnham: Ashgate.

Berber Sardinha, T. 1999. Using key words in text analysis: Practical aspects. *DIRECT Papers* 42, LAEL, Catholic University of São Paulo.

Biber, D. & Conrad, S. 1999. Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honor of Stig Johansson*, H. Hasselgard & S. Oksefjell (eds), 181–189. Amsterdam: Rodopi.

Culpeper, J. 2002. Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*. In *Conversation in Life and in Literature: Papers from the ASLA Symposium* [Association Suedoise de Linguistique Appliquée (ASLA) 15], U. Melander-Marttala, C. Östman & M. Kytö (eds), 11–30. Universitetstryckeriet: Uppsala.

Culpeper, J. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1): 29–50.

Deignan, A. 2009. Searching for metaphorical patterns in corpora. In *Contemporary Corpus Linguistics,* P. Baker (ed.), 9–31. London: Continuum.

Hoey, M. 1986. The discourse colony: A preliminary study of a neglected discourse type. *University of Birmingham*: *Discourse Analysis Monographs* 13: 1–26.

Kemppanen, H. 2004. Keywords and ideology in translated history texts: A corpus-based analysis. *Across Languages and Cultures* 5(1): 89–106.

Ku, P. & Yang, A. 1999. An analysis of key words in tourism English. The Proceedings of the 1999 English for Specific Purposes Conference, Taoyuan, 232–241. Taiwan: Ming Chuan University.

McEnery, T. 2009. Keywords and moral panics: Mary Whitehouse and media censorship. In *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*, D. Archer (ed.), 93–124. Farnham: Ashgate.

Rigotti, E. & Rocci, A. 2002. From argument analysis to cultural keywords (and back again). <http://www.ils.com.unisi.ch/articoli-rigotti-rocci-keywords-published.pdf> (May 2007). In *Proceedings of the 5th Conference of the International Society for the Study of Argumentation*, F. H. van Eemeren, A. Blair, C. A. Willard & A. F. Snoeck Henkemans, 903–908. Amsterdam: SicSat.

Scott, M. 1996. New versions in 1997, 1999, 2004, 2008. *Wordsmith Tools*, Oxford: OUP. Now Liverpool: Lexical Analysis Software.

Scott, M. 1997. PC analysis of key words – and key key words. *System* 25(1): 1–13.

Scott, M. 2006. The importance of key words for LSP. In *Information Technology in Languages for Specific Purposes: Issues and Prospects*, E. Arnó Macià, A. Soler Cervera & C. Rueda Ramos (eds), 231–243. Berlin: Springer.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Scott, M. 2008. Developing WordSmith. *International Journal of English Studies* 8(1): 153–172.

Scott, M. 2009. In search of a bad reference corpus. In *What's in a Word-list? Investigating Word Frequency and Keyword Extraction,* D. Archer (ed.), 79–92. Oxford: Ashgate.

Scott, M. Forthcoming. What can corpus software do? In *Routledge Handbook of Corpus Linguistics,* M. McCarthy & A. O'Keeffe (eds.). London: Routledge.

Scott, M. Ongoing. <http://tinyurl.com/cve22m> : <http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm>.

Seale, C., Charteris-Black, J. & Ziebland, S. 2006. Gender, cancer experience and internet use: A comparative keyword analysis of interviews and online cancer support groups. *Social Science and Medicine* 62(10): 2577–2590.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Toolan, M. 2004. Values are descriptions; Or, from literature to Linguistics and back again by way of keywords. *Belgian Journal of English Language and Literatures* (BELL New Series 2): 11–30.

Toolan, M. 2006. Top keyword abridgements of short stories: a corpus linguistic resource? *Journal of Literary Semantics* 35: 181–194.

Williams, R. 1976. *Keywords*. London: Fontana.

Zipf, G. K. 1965. *Human Behavior and the Principle of Least Effort.* New York NY: Hafner. (Facsimile of 1949 edition).

Closed-class keywords and corpus-driven discourse analysis

Nicholas Groom
University of Birmingham

Keywords belonging to closed grammatical classes (i.e. conjunctions, determiners, prepositions and pronouns) are often perceived as useful indicators of the characteristic style of a particular text or corpus, but as being of less interest to researchers interested in its semantic properties. The aim of this chapter is to propose, contrary to this mainstream view, that closed-class keywords can form a valid and even preferable basis for empirical linguistic research into specialized discourses, "discourses" being defined here as constellations of meanings and values associated with specific communities or institutions. The argument is illustrated with practical examples drawn from a keywords analysis of a 3-million-word corpus of academic journal articles representing the academic disciplinary discourse of history.

1. Introduction

The advent of computer-assisted keyword analysis has opened up a range of exciting new possibilities for corpus linguistic research into *discourses*, sets of meanings and values which are associated with specific communities or institutions, and which are produced and reproduced through characteristic and often highly conventionalised linguistic choices (Carter 1995; Stubbs 1996, 1997, 2001; Partington 2003; Baker & McEnery 2005; Teubert 2005; Baker 2006; Mautner 2007). Prior to the development of keyword-identifying software such as *WordSmith Tools* (Scott 1996) and *AntConc* (Anthony 2006), applied linguists interested in using corpora to study discourses had little option but to adopt a broadly deductive approach, in which items for concordance analysis or other form of detailed study would have to be specified in advance.

Although perfectly appropriate for some research purposes, this approach can be problematic in the context of discourse analysis. First of all, it runs the risk of producing findings which may be fatally biased by the *a priori* nature of the

selection process itself (Widdowson 2004; O'Halloran & Coffin 2004). Additionally, the analysis may be overly constrained by conventional descriptions and thus only able to provide more detail about already well-known features (Sinclair 2001, 2004). At the same time, it may also be insensitive to any genuinely new features that might be lying undiscovered in the data (Hunston & Francis 1999).

Keyword analysis provides a way of circumventing all of these problems at a stroke. First of all, delegating the initial task of identifying items for analysis to a computer algorithm ensures that this stage of the research is completely insulated from researcher bias. Secondly, the simple way in which the algorithm itself works – by comparing the frequencies of wordforms in a corpus of specialized texts against the frequencies of the same wordforms in a larger and more general reference corpus – means that it is entirely unencumbered by previous linguistic theories or descriptions. Finally, the keyword identification procedure has a well-attested knack of unearthing features and trends in corpus data that would be difficult or impossible to observe by more conventional methods, and which could not have been predicted or even imagined in advance of the analysis (Scott 1997, 2001; Hunston 2002; Baker 2004, 2006; Scott & Tribble 2006; Lee 2008). In short, the advent of keyword analysis has opened up the possibility of a genuinely inductive, "corpus-driven" approach to the analysis of specialized discourses (Tognini-Bonelli 2001).

The adoption of a keyword approach is not without its own methodological challenges, however, and the aim of this chapter is to address one of the most fundamental of these. The problem to be discussed here is that, even when extremely high statistical cut-off values are applied (e.g. p = .000001 or .0000001), a keyword test of any corpus large enough to represent the typical linguistic practices of a specialized discourse community is likely to yield a list of many hundreds or even thousands of candidate items for analysis (see e.g. Baker 2004: 348–49; Scott & Tribble 2006: 77–79). Given that it would be practically impossible to submit every keyword in such a list to detailed study, the question arises as to how a principled and yet maximally useful selection of items for closer attention might be made.

A simple and frequently applied solution involves focusing exclusively on the words at the very top of the list, and disregarding anything that falls below an arbitrary cut-off point (usually the top 20, 50 or 100 items). The justification for this top-slicing approach is reasonable enough: the words at the top of the list have the highest "keyness" scores, and are therefore statistically the most strongly associated with the discourse in question. However, this is clearly only an ad hoc solution at best, and many researchers who have applied it will have experienced the frustration of noticing keywords lurking just below their chosen cut-off point that look much more interesting than many of the words above it.

Another commonly (and in practice often jointly) applied solution is to prioritise certain categories of keywords over others, or even to discard some categories of keyword from the analysis altogether, on the grounds that they are likely to be less germane to the concerns of the discourse analyst. As Baker (2006) notes, the default option for some researchers is to discard from the outset all of the so-called "grammatical" words – the conjunctions, determiners, prepositions and pronouns that belong to (more or less) closed sets – that may feature in any given keyword list, and to focus instead on the "lexical" words – the nouns, verbs, adjectives and adverbs that form "open" (i.e. infinitely extendable) grammatical classes – that remain.

In this chapter, however, we will present a case for pursuing exactly the reverse strategy. That is, a case will be made for discarding all of the open-class items in a keywords list as a preliminary step, and focusing instead on the closed-class keywords that remain (e.g. Gledhill 2000a, 2000b; Groom 2007). The argument of the chapter is divided into two parts. The first part presents a theoretical and practical case for closed-class keywords as valid objects of study. In the second part, a detailed rationale for focusing exclusively on closed-class keywords is provided. Throughout, the argument will be illustrated with attested examples, most of which have been drawn from a corpus-driven study of the academic disciplinary discourse of history (Groom 2007), which was based entirely on precisely such a "closed-class keywords only" methodology.

2. Closed-class keywords as valid objects of analysis

Our first task is to establish that closed-class keywords constitute valid objects of semantic analysis at all, as this is an idea that runs counter to certain strands of conventional wisdom in both mainstream linguistics and keywords research alike. Traditionally, open class words such as *gift*, *drive*, *cold* or *gradually* are seen as having both semantic content and grammatical functions, while closed class words such as *and*, *the*, *of* or *its* are regarded as having only the latter. On this account, closed-class keywords would seem to be of no interest to the discourse analyst at all, given that the aim of discourse analysis (as defined in this chapter) is to identify the conventional meanings and values expressed in a corpus of texts.

Although intuitively appealing, this neat theoretical division between meaningful and meaningless words is not supported by the empirical research findings that have emerged in recent decades from the field of corpus linguistics. In fact, the weight of corpus evidence has led to the development of a different theoretical model altogether, in which it is proposed that meanings typically reside in sequences of words, and not in the individual wordforms that comprise

such sequences (Sinclair 1991, 1999, 2003, 2004; Francis 1993; Francis, Manning & Hunston 1996, 1998; Moon 1998; Partington 1998; Hunston 2002, 2003; Danielsson 2003, 2007; Groom 2005; Hoey 2005; Teubert & Čermáková 2007). As Sinclair (1991: 108) puts it, "[m]ost everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text". Consider, by way of illustration, the noun *gift* in the following three statements (all of which are attested instances from the Bank of English corpus).

(1) You do have a **gift** for understatement, Jill.

(2) England gave Romania a **gift** of a goal.

(3) I wish I could help you, but it's not in my **gift**.

For the corpus-driven linguist, *gift* in itself does not mean 'talent' in the first instance, 'unmissable and unintentionally conceded opportunity' in the second, and 'power' or 'authority' in the third. Rather, the corpus-driven view is that the 'talent' meaning is made by the whole sequence **HAVE** + ***a*** + ***gift*** + ***for*** + **noun**, the 'unmissable opportunity' meaning is made by the whole sequence ***a*** + ***gift*** + ***of*** + ***a*** + ***goal***, and the 'power/authority' meaning is made by the whole sequence **noun/pronoun** + **BE** + (***not***) + ***in*** + **possessive determiner** + ***gift***. Each meaning is the property of an extended lexical unit consisting of both open and closed class items, and cannot be reduced to the noun *gift*, or indeed to any other single constituent item. Indeed, in these examples it is clear that the supposedly meaningless closed-class words make just as important a contribution to the overall meaning of each phraseology as do the open-class items (***a*** *gift* ***for*** *understatement*; ***a*** *gift* ***of*** ***a*** *goal*; ***not in my*** *gift*), and therefore constitute equally valid starting points for a semantically-oriented analysis.

Within the specific context of keywords analysis, crude divisions between meaning and form have now largely been replaced by a more subtle distinction between open-class keywords as indicators of the "aboutness" of a text or corpus, and closed-class keywords as indicators of its characteristic "style" (Scott & Tribble 2006). While this distinction could still be (and no doubt sometimes still is) used to justify the omission of closed-class keywords from the analysis altogether, the consensus view among discourse researchers is that closed-class keywords are potentially of interest insofar as they act as indicators of style, and "the style of a text may play some role in the discourses within it" (Baker 2006: 127). However, this is not to say that closed-class keywords are widely regarded as being *equally* worthy of analytic attention. On the contrary, the standard view among keywords analysts remains very firmly that open-class keywords are the principal indicators of the "aboutness" of a corpus, and are thus "generally those which are most interesting to analyse" (Baker 2006: 127).

The argument against pronouncements such as this is not so much that they are wrong, but rather that they are missing the point. For corpus-driven discourse analysis, the crucial distinction is not between ontologically different categories of keywords (aboutness versus style, lexical versus grammatical, content versus function, etc.), but between methodologically different ways of *looking at* keywords; specifically, as endpoints of quantitative analysis on the one hand, or as starting points for qualitative analysis on the other. If we only study closed-class keywords in the decontextualised lists that constitute the on-screen output of the keywords procedure, then it is certainly true that these words can tell us almost nothing about the meanings and values expressed in a specialized corpus. However, if we take the trouble to look in detail at the phraseological behaviour of individual closed-class keywords in their typical contexts of occurrence as revealed by the concordancer, we find that these ostensibly meaningless words can tell us at least as much about the preferred meanings of a particular discourse community as they can tell us about the preferred stylistic features associated with this community.

In support of this claim, let us look in detail at a 100-line concordance sample of *of*, the highest ranking closed-class keyword (and the fifth highest ranking keyword overall in statistical terms) in HistArt, a 3.2 million-word corpus of journal articles representing the academic disciplinary discourse of history. HistArt was originally compiled and analysed as part of a broader corpus-driven study of Anglophone humanities discourses (Groom 2007), and the data reported below are derived from this larger-scale research. As the second most frequent word not only in HistArt but also in written English more generally (Leech, Rayson & Wilson 2001: 181), *of* constitutes an excellent test-bed for the claim that closed-class keywords are tractable to qualitative semantic analysis. The choice of *of* here is also intended as a modest tribute to the pioneering work of Sinclair, whose own corpus-driven analysis of *of* (Sinclair 1991) provided both the inspiration and the methodological template for the empirical research to be reported here. We will begin by describing our sample concordance (given in full as Appendix 1) in simple structural terms. Following this, we will submit the same data to a more thoroughly semantically oriented analysis.

In 89% of the concordance sample in Appendix 1, *of* functions as a central linking preposition in complex noun phrases such as those featured in Figure 1:

n	***of***	**N**
the post-Suez growth	**of**	Arab nationalism
a rejection	**of**	parliamentary and democratic values
two factions	**of**	mill workers

Figure 1. n *of* n in HistArt

Following the COBUILD metalanguage used by phraseological researchers such as Hunston and Francis (1999), these postmodified noun phrases will be referred to as **n *of* n**. (Note that in this metalanguage, grammatical categories are written in lowercase letters, particular word forms are written in italic, and the whole sequence is presented in bold typeface.)

Of the remaining eleven lines in the sample, four can be classified as prepositional phrases in which *of* occurs at the end of an initial multi-word prepositional element (Biber et al. 1999; Carter & McCarthy 2006). Again following COBUILD practice, these will be coded here as **prep n** (Figure 2):

prep	**N**
Instead **of**	a long, cumbersome list of names
because **of**	the expression often used by merchants
at the level **of**	ideology critique
on the heels **of**	the Lyons Council of 1274

Figure 2. prep n in HistArt

The bulk of the remainder consists of a small selection of verb and adjective patterns of the kind surveyed by Francis, Manning & Hunston (1996, 1998) and theorised by Hunston & Francis (1999). These are summarised in Figure 3.

Structural sequence	**Example**	**N**
v n *out of* n	deport them out **of** the region	2
adj *of* n	The sufferer's own silence is typical **of** a larger phenomenon	1
***the* adj-superl *of* pl-n**	the dubbing of a knight is the most familiar **of** the new ceremonies	1
v n *of* n	Chulaki accused them both **of** cosmopolitanism	1
v *of* n	the entire crew … consisted **of** lascars	1

Figure 3. Other structural sequences in the HistArt sample

Finally, we note the following fixed phrase:

> they facilitated the entry of respectable women into what one turn-**of**-the-century writer termed the "Night Side of London"

If there were sufficient quantities of these it would be worth subdividing them into different kinds of phrase (adjectival, nominal, etc.), but this is clearly not the case in this sample.

A question that immediately arises at this point concerns how representative the above analysis is of the structural behaviour of *of* as a whole, particularly given that *of* occurs 136,697 times in HistArt in total. An identical analysis of two further 100-line samples of *of* in HistArt conducted in Groom (2007) suggests that randomly selected concordances of *of* in HistArt are in fact surprisingly consistent – so much so in fact that it would seem fairly safe to make rough statements of proportion on the basis of three concordance samples at most, as can be seen in Table 1 below. (It is also worth noting here that the other 25 closed-class keywords studied in Groom (2007) proved equally consistent.)

Table 1. Structural analysis of three concordance samples of ***of***

Structures	Sample 1 (%)	Sample 2 (%)	Sample 3 (%)	Average
n *of* n	89	92	87	89.333
prep n	4	5	9	6.000
Other	7	3	4	4.666

The other objection that might be raised at this point is that, while interesting in itself, the above analysis is not particularly useful from a discourse analytic perspective. It is certainly true that all the foregoing observations seem to do is provide further confirmation – as if further confirmation were needed – of the standard view of academic writing across the disciplinary spectrum as being characterised by a heavy use of complex noun phrases (see e.g. Halliday 1989; Halliday & Martin 1993; Biber et al. 1999; Carter & McCarthy 2006; Scott & Tribble 2006). Even less encouragingly, the sample itself is far too small to exhibit any repeated combinations of words that might conveniently encapsulate particular aspects of the academic disciplinary discourse of history. Indeed, apart from the regular (and entirely unsurprising) appearance of the definite article *the*, there is no surface patterning in these data to speak of at all. However, if we search this concordance not for repetitions of particular wordforms but for repeated sequences of underlying meanings instead, then distinct phraseological groupings immediately begin to appear. In the analysis of the **n *of* n** data that follows, we shall follow Hunston (2006) in referring to these repeated catenations of conceptual categories as *semantic sequences*.

In 14 of the 89 **n *of* n** sequences (i.e. 14% of the whole sample), the noun that follows ***of*** describes an entity (e.g. *Florence*), a process (e.g. *military mobilisation*) or a state of affairs (e.g. *academic life*), and the noun on the left describes a property, attribute or feature that this entity, process or state of affairs has. *Florence* has *an urban population*; *military mobilisation* has *a meaning*; *academic life* has *essential values*; and so on. This catenation can be expressed formally as the semantic sequence **PROPERTY + *of* + PHENOMENON**. Examples of this sequence are presented in Figure 4 below.

PROPERTY	*of*	PHENOMENON
the essential values	**of**	academic life
the urban population	**of**	Florence
the meaning	**of**	military mobilisation

Figure 4. PROPERTY + *of* + PHENOMENON in HistArt

The second most frequent group in the concordance sample consists of eleven concordance lines in which the noun preceding ***of*** describes a process (e.g. *building*, *attribution*, *rejection*), and the noun that follows ***of*** describes the phenomenon affected by that process (e.g. *a new church, insensitivity, values*). The concordance data for this semantic sequence, which we shall parse as PROCESS + *of* + OBJECT, are presented in Figure 5.

PROCESS	*of*	OBJECT
the building	**of**	a new church
the attribution	**of**	insensitivity
a rejection	**of**	parliamentary and democratic values

Figure 5. PROCESS + *of* + OBJECT in HistArt

A simple test of whether an instance belongs to this group is to ask whether it can be expressed as an independent clause in the passive voice. For example, *the building of a new church* can be reformulated as 'a new church was built', *the attribution of insensitivity* as 'insensitivity was attributed', and *a rejection of parliamentary and democratic values* as 'parliamentary and democratic values were rejected'.

The next two groups both consist of nine lines each. In the first of these, TEXT + *of* + CONTENT (Figure 6), the noun or noun phrase to the left of *of* describes a text or a text type (or a series of texts or text types), while the noun or noun phrase to the right indicates the content of the text(s) or text type(s) in question.

TEXT	*of*	CONTENT
list	**of**	cases
books	**of**	common errors
The Cambridge History	**of**	Japan

Figure 6. TEXT + *of* + CONTENT in HistArt

The other group features an initial element describing a person or group holding a position of social power, responsibility or esteem, and a final element describing the domain over which this individual or group has authority. This sequence of elements can be parsed as AUTHORITY + *of* + DOMAIN (see Figure 7):

AUTHORITY	*of*	DOMAIN
the Emperor	**of**	Constantinople
Secretary	**of**	State
the Austin canons	**of**	St Mary's Leicester

Figure 7. AUTHORITY + *of* + DOMAIN in HistArt

A fifth phraseological group consists of seven lines which quantify phenomena. These we may refer to as instances of the semantic sequence **QUANTITY + *of* + PHENOMENON**, as illustrated in Figure 8.

QUANTITY	*of*	PHENOMENON
a fair number	**of**	dioceses
several	**of**	the dead
eight	**of**	the defendants

Figure 8. QUANTITY + *of* + PHENOMENON in HistArt

The next two groups of sequences augment the **PROCESS + of + OBJECT** phraseology discussed earlier, in that they focus on other aspects of the system of transitivity in English. The first of these, **PROCESS + *of* + ACTOR** (6 instances in the sample) first describes the process in question, and then goes on to identify the agent of that process. As can be seen in Figure 9 below, what unites these phrases is that they can all be reformulated as intransitive clauses (e.g *Arab nationalism grew*; *politics and morality converged*; "*Western Civilization*" *spread*).

PROCESS	*of*	ACTOR
the post-Suez growth	**of**	Arab nationalism
the convergence	**of**	politics and morality
the spread	**of**	"Western Civilization"

Figure 9. PROCESS + *of* + ACTOR in HistArt

The other group consists of five instances which first identify the actor of a process, and then identify the object of that process. This group, which can be parsed as **ACTOR + *of* + OBJECT**, is shown in Figure 10.

ACTOR	*of*	OBJECT
authors	**of**	non-fiction
occupants	**of**	the throne
reformers	**of**	local government

Figure 10. ACTOR + *of* + OBJECT in HistArt

As can be seen, the process itself is not stated directly at all in this sequence, but can be inferred from the meaning of the sequence as a whole. In fact, the ability to paraphase instances of this sequence as full SVO-type clause constructions constitutes an effective test of membership of this semantic group. *[A]uthors of non-fiction* can be reformulated as 'they wrote non-fiction'; *occupants of the throne* as 'they occupied the throne'; and *reformers of local government* as 'they reformed local government'.

The eighth and final group to be discussed here consists of six instances in which a phenomenon (described in the element that follows *of*) is being conceptualised in a particular way. In the examples in Figure 11, *Mothers' Day* is being conceptualised as an *institution*, *cultural stewardship* as a *complementary component*, and *Marcel Déat* and *Jacques Doriot* as *cases*. We may express this sequence as **CONCEPTUALISATION + *of* + PHENOMENON**.

CONCEPTUALISATION	*of*	PHENOMENON
the institution	**of**	Mothers' Day
the complementary component	**of**	cultural stewardship
The cases	**of**	(ex Socialist) Marcel Déat and (ex Communist) Jacques Doriot

Figure 11. **CONCEPTUALISATION + *of* + PHENOMENON** in HistArt

A test of membership of this group involves asking whether the **PHENOMENON** element could plausibly be conceptualised in another way. For example, *Mothers' Day* could be conceptualised as a *respite* (for overworked mothers), as *a guilt trip* (for lazy children), or as *a marketing ploy* (from the point of view of a cynical observer), among other possibilities.

The remaining 22 **n *of* n** phrases in our concordance sample fall into eight smaller semantic sequence groups which, for reasons of space, will not be described in detail here. The important point to note is that, even with the addition of these eight minor groups of sequences, we have only needed a total of 16 semantic sequences in order to account for 89 concordance lines for *of*. Given the very high frequency and enormous combinatory potential of *of* in English, it would have been entirely reasonable to assume (as indeed I had done prior to conducting this analysis) that 89 semantic sequences might have been required to describe 89 lines of data, even in a corpus as specialized as HistArt.

Even more surprisingly, a semantic analysis of the two further samples studied in Groom (2007) were also very consistent with the original analysis; it was only necessary to define two more semantic sequences in order to cover an additional 179 concordance lines featuring the structure **n *of* n**, and even these were only for

minor (i.e. very infrequently occurring) usages. The same major sequences dominated all three samples, and in very consistent numerical proportions overall. The only corrections worth noting concern the sequences **CONCEPTUALISATION + *of* + PHENOMENON**, which seems to have been substantially underrepresented in the first sample, and **TEXT + *of* + CONTENT**, which seems to have been correspondingly overrepresented (see Table 2). In all cases, the average figures probably present a more reliable general indication of likely proportion. (This also indicates the value of investigating more than just one 100-line sample per word).

Table 2. Semantic sequences in three concordance samples of ***of***

Semantic sequence	% of sample			
	First sample	Second sample	Third sample	Average
PROCESS + *of* + OBJECT	11	17	15	14.333
PROPERTY + *of* + PHENOMENON	14	13	16	14.333
CONCEPTUALISATION + *of* + PHENOMENON	6	15	15	12.000
QUANTITY + *of* + PHENOMENON	7	8	7	7.333
PROCESS + *of* + ACTOR	6	7	6	6.333
AUTHORITY + *of* + DOMAIN	9	4	4	5.666
TEXT + *of* + CONTENT	9	2	2	4.333
ACTOR + *of* + OBJECT	5	1	2	2.666
Other sequences	22	26	20	22.666

In summary, what the above analysis shows is that a set of concordance lines based on a single closed-class keyword can be divided into a distinct and quantifiable series of phraseological groups, each of which might yield interesting insights into the preferred meanings and values of a specialized discourse community. Space does not permit a detailed discussion of precisely what the data reviewed above might be telling us about the academic discourse of history, but the following two claims may provide some indication of their wider interpretative potential. First, we may advance the claim that around a quarter of all instances of ***of*** in HistArt participate in semantic sequences expressing nominalized processes of various kinds. Of these, the majority (c. 14%) are the nominal equivalents of passive clauses. This finding suggests that professional historical discourse is more likely to focus on social phenomena as they are affected by historical processes than it is to focus on "History" as a temporal phenomenon shaped by the agentive actions of powerful individuals (cf. Coffin 2004, 2006).

Second, we may propose that the semantic sequences **PROPERTY + *of* + PHENOMENON** and **CONCEPTUALISATION + *of* + PHENOMENON** account for a further 25% of all instances of ***of*** in HistArt. Given that the wordform *of* alone comprises

4.34% of the whole corpus, it would seem reasonable to suggest that these two sequences are also very important indicators of aspects of the epistemology of history as a humanities discipline (cf. Becher & Trowler 2001). Specifically, the prevalence of PROPERTY + *of* + PHENOMENON sequences may indicate a general preference for highly detailed and particularistic modes of analysis, while the prevalence of CONCEPTUALISATION + *of* + PHENOMENON sequences may indicate that historical knowledge-making is to a large extent reiterative in nature. Writing a book about the causes of the Crimean War, for example, will not definitively "solve" the problem of what caused the Crimean War; this problem will remain largely intact for future generations of scholars to revisit and tackle afresh. The task of the historian therefore is to deepen and broaden understandings of established disciplinary problems by reconceptualising and reinterpreting them in the light of new theoretical perspectives or fresh primary source data.

There is much more that could be said on these (and many other) observations arising from the data reviewed in this section, but to do so would be to go well beyond the scope of this article. The main point that we hope to have established thus far is that closed-class keywords may be just as useful as starting points for the analysis of the "aboutness" of a specialized corpus as they are commonly held to be as indicators of its preferred "style", and that it would therefore be wrong to regard them as irrelevant, or even as being only of secondary importance or indirect interest, to the discourse analyst.

3. Closed-class keywords as preferred objects of analysis

Having established closed-class keywords as valid objects of corpus-driven discourse analysis, we now turn to the question of whether closed-class keywords can serve as the *sole* objects of such an analysis. If it is the case that meaning typically resides in conventionalised sequences that consist of both open- and closed-class items, as was argued at the beginning of the previous section, then there would on the face of it seem to be no more justification for excluding open-class words from a keywords analysis than there is for excluding closed-class items. Yet this is precisely what is being proposed in this chapter. What, then, is the rationale for basing an analysis exclusively on closed-class keywords?

The simplest argument in favour of this approach is that the analyst who opts to focus on closed-class keywords will receive a compact and tractable list of items for analysis from the outset, and will therefore not need to resort to the arbitrary top-slicing procedures which we criticised earlier in this chapter. Gledhill's (2000a) pioneering study of the discourse of cancer research, for example, was based on an analysis of 39 closed-class keywords extracted from a 500,000-word corpus

of pharmaceutical science research articles, while the keywords comparison of HistArt against the written component of the British National Corpus described in Groom (2007) resulted in an even more compact list of just 26 closed-class items for analysis. An additional point worth making here is that closed-class keywords tend to be scattered fairly evenly throughout keyword lists which have been ranked according to keyness (rather than raw frequency) values, thereby making the choice of such items an attractive alternative to random sampling.

Focusing on closed-class keywords may also result in a more fruitful as well as a more manageable set of starting points for concordance analysis. Given that closed-class words are the commonest words in virtually all corpora, it follows that an analysis based on even a small selection of them will account for a far greater proportion of the data as a whole than can be achieved through an analysis of even a large selection of open-class items (cf. Zipf 1935; Sinclair 1991, 1999). For instance, the 26 closed-class keywords obtained for HistArt alone constitute 20.28% of the whole corpus, and the focus of the analysis carried out in this study was not on these words in isolation but on their broader phraseological environments, thereby expanding this coverage further still. If it is true that "the majority of text is made of the occurrence of common words in common patterns", as Sinclair (1991: 108) suggests, then it is arguably preferable to select the commonest of these common words as the empirical basis of a corpus-driven discourse analysis.

Concordances of single closed-class words also allow the researcher to identify and accommodate a much wider range of phraseological phenomena than might otherwise be possible. Any random concordance sample of ***of*** in HistArt, for example, will (as we saw earlier) include a number of instances of the kinds of fixed phraseological sequence that are typically highlighted by currently popular "lexical bundle" or "word cluster" modes of analysis (e.g. Biber et al. 1999; Stubbs 2003; Cortes 2004; Scott & Tribble 2006; Hyland 2008). Some examples of these are given in Figure 12 below.

1.	debate began in London,	**in advance of**	the 1907 City Council el
2.	men?"37 His defensiveness	**on behalf of**	the scholastic prerogati
3.	nds to loom ever larger	**in the eyes of**	electors, whether their
4.	y, on June 8, 1794. 66	**In the face of**	such events, the madonna
5.	foreign member states.	**As a result of**	these rulings, and relat
6.	ularly from farmers	**in the vicinity of**	the camps.31 Although th

Figure 12. Fixed 3- and 4-word sequences in a concordance sample of ***of***

The same sample is also likely to include phraseologies which, while just as conventionalised as the above, are much less rigidly formulaic, and may therefore be under-emphasized or overlooked altogether by automatic routines that search

for repetitions of sequences of exactly the same orthographic wordforms. Among these will be instances of the "grammar patterns" surveyed by Francis, Manning & Hunston (1996, 1998) and theorised by Hunston & Francis (1999). Examples of these are given in Figure 13.

1.	tion and most of us are **aware**	**of**	rules of evidence, which of ne
2.	ed, we would **know** very little	**of**	the context of these oaths of
3.	m of money. The vast **majority**	**of**	cases, however, went before th
4.	icular, the frequent **outbreak**	**of**	disease, which posed an ostens
5.	erer's own silence is **typical**	**of**	a larger phenomenon. Before th
6.	brother, he **wrote** repeatedly	**of**	his bewilderment at the eighte

Figure 13. Grammar patterns in a concordance sample of ***of***

In pattern grammar terms, lines 1 and 5 in Figure 8 would be coded as instances of the pattern **adj *of* n** belonging to the 'AWARE AND UNAWARE' and 'INDICATIVE' meaning groups respectively (Francis, Manning & Hunston 1998: 451–457). Lines 2 and 6 could be categorised as instances of the pattern **v *of* n** belonging to the 'KNOW' and 'TALK' groups (Francis et al 1996: 211–214), and lines 3 and 4 as instances of the pattern **n *of* n** belonging to Francis et al.'s 'PERCENTAGE' and 'EPISODE' meaning groups (1998: 176–199). Further samples may reveal a corpus-specific preference for certain of these meaning groups, thereby providing further information about the preferred meanings and values of professional historians (cf. Charles 2000, 2003, 2006; Groom 2005).

Most importantly of all, concordances of single closed-class keywords allow the analyst to identify semantic sequences (Hunston 2006), that is, repeated sequences of semantic elements which may have a very heterogeneous surface realisation. A good example of such a sequence is the CONCEPTUALISATION + *of* + PHENOMENON phraseology illustrated in Figure 11 earlier. Clearly, phraseologies such as this would not be captured by lexical bundle or word cluster analyses because they are not manifested in exact repetitions of the same sequences of orthographic wordforms. Neither are they classifiable as grammar patterns, because their constituent elements are not necessarily bound by grammatical dependency relationships or collocational ties.

Even more importantly in terms of the present argument, it is not easy to imagine how a semantic sequence such as CONCEPTUALISATION + *of* + PHENOMENON could be identified by studying concordances of open-class keywords either. A concordance analysis of any given open-class word may generate a useful phraseological profile for that particular word, but the analyst would have to study, sift through and cross-classify the phraseological profiles of a forbiddingly

large number of individual open-class items in order to generalise a single semantic sequence from her or his corpus. Concordancing closed-class words, on the other hand, allows the analyst to proceed directly to the identification of underlying (and frequently expressed) commonalities of meaning among superficially very different-looking sequences of linguistic elements. And as we have seen in the case of *of*, these sequences may actually account for the vast majority of the phraseological data available for any given keyword. It is for this reason above all others that we may reasonably propose closed-class keywords as productive and even perhaps preferable points of departure for corpus-driven investigations of specialized discourses.

4. Conclusion

The keywords procedure offers researchers a powerful means of gaining access to the meanings and values of specialized discourse communities. However, it is not feasible in most cases to submit every item identified by the keyword algorithm to detailed study. This chapter has presented a case for resolving this problem by focusing exclusively on closed-class keywords. We have seen that, despite their very high frequency and tendency to occur in a wide range of grammatical structures, closed-class keywords are entirely amenable to empirical semantic analysis, and that reasonably robust claims can in fact be made on the basis of the analysis of surprisingly small numbers of randomly selected concordance lines. It has also been argued that closed-class keywords may actually be preferable over their open-class counterparts as objects of corpus-driven discourse analysis because they offer much greater coverage of the phraseological data in a specialized corpus in both quantitative and qualitative terms.

For the most part, these arguments have been illustrated by reference to the analysis of a single closed-class keyword in a corpus representing a single academic discourse, but the final claim of this chapter is that precisely the same arguments would be applicable to any other closed-class word, and to any specialized discourse for which a corpus could be compiled. It will be interesting to see whether this claim is borne out by future corpus-driven studies of specialized discourses that dare to take closed-class keywords seriously, and dare to place them where this chapter has argued that they deserve to be placed: right at the heart of the analysis.

References

Anthony, L. 2006. *AntConc: A Freeware Concordance Program for Windows, Macintosh OS X, and Linux*. <http://www.antlab.sci.waseda.ac.jp/software.html> (5 April 2008).

Baker, P. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32: 346–359.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. & McEnery, T. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics* 4: 197–226.

Becher, T. & Trowler, P. 2001. *Academic Tribes and Territories: Intellectual Enquiry and the Vulture of Disciplines*, 2nd edn. Buckingham: The Society for Research into Higher Education and Open University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Carter, R. 1995. *Keywords in Language and Literacy*. London: Routledge.

Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: CUP.

Charles, M. 2000. The role of an introductory *it* pattern in constructing an appropriate academic persona. In *Patterns and Perspectives: Insights into EAP Writing Practice*, P. Thompson (ed.), 45–59. Reading: The University of Reading, CALS.

Charles, M. 2003. "This mystery…": A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* 2: 313–326.

Charles, M. 2006. The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics* 27: 492–518.

Coffin, C. 2004. Learning to write history: The role of causality. *Written Communication* 21: 261–289.

Coffin, C. 2006. *Historical Discourse: The language of time, cause and evaluation*. London: Continuum.

Cortes, V. 2004. Lexical bundles in published and student writing in history and biology. *English for Specific Purposes* 23: 397–423.

Danielsson, P. 2003. Automatic extraction of meaningful units from corpora: A corpus-driven approach using the word *stroke*. *International Journal of Corpus Linguistics* 8: 109–127.

Danielsson, P. 2007. What constitutes a unit of analysis in language? *Linguistik Online* 31: 2/07. <http://www.linguistik-online.de/31_07/danielsson.html> (Last accessed 5th April 2008).

Francis, G. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 137–156. Amsterdam: John Benjamins.

Francis, G., Manning, E. & Hunston, S. 1996. *Collins COBUILD Grammar Patterns* 1: *Verbs*. London: HarperCollins.

Francis, G., Manning, E. & Hunston, S. 1998. *Collins COBUILD Grammar Patterns* 2: *Nouns and Adjectives*. London: HarperCollins.

Gledhill, C. 2000a. *Collocations in Science Writing*. Tübingen: Gunter Narr.

Gledhill, C. 2000b. The discourse function of collocation in research article introductions. *English for Specific Purposes* 19: 115–135.

Groom, N. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4: 257–277.
Groom, N. 2007. Phraseology and Epistemology in Humanities Writing: A Corpus-driven Study. PhD dissertation, University of Birmingham.
Halliday, M. A. K. 1989. *Spoken and Written Language*. Oxford: OUP.
Halliday, M. A. K. & Martin, J. R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer.
Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP.
Hunston, S. 2003. Lexis, wordform and complementation pattern: A corpus study. *Functions of Language* 10: 31–60.
Hunston, S. 2006. Starting with the small words: Patterns, lexis and semantic sequences. Paper given at *Exploring the lexis-grammar interface*, University of Hannover, 6th October 2006.
Hunston, S. & Francis, G. 1999. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins.
Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.
Lee, D. W. Y. 2008. Corpora and discourse analysis: New ways of doing old things. In *Advances in Discourse Studies*, V. K. Bhatia, J. Flowerdew & R. H. Jones (eds), 86–99. London: Routledge.
Leech, G., Rayson, P. & Wilson, A. 2001. *Word Frequencies in Written and Spoken English*. Harlow: Pearson Education.
Mautner, G. 2007. Mining large corpora for social information: The case of elderly. *Language in Society* 36: 51–72.
Moon, R. 1998. *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.
O'Halloran, K. & Coffin, C. 2004. Checking overinterpretation and underinterpretation: Help from corpora in critical linguistics. In *Applying English Grammar: Functional and Corpus Approaches*, C. Coffin, A. Hewings & K. O'Halloran (eds), 275–297. London: Hodder Arnold.
Partington, A. 1998. *Patterns and Meanings* [Studies in Corpus Linguistics 2]. Amsterdam: John Benjamins.
Partington, A. 2003. *The Linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House*. London: Routledge.
Scott, M. 1996. *WordSmith Tools*. Oxford: OUP.
Scott, M. 1997. PC analysis of key words – and key key words. *System* 25: 233–245.
Scott, M. 2001. Comparing corpora and identifying key words, collocations and frequency distributions through the *WordSmith Tools* suite of computer software. In *Small Corpus Studies and ELT* [Studies in Corpus Linguistics 5], M. Ghadessy, A. Henry & R. Roseberry (eds), 47–67. Amsterdam: John Benjamins.
Scott, M. & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.
Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
Sinclair, J. M. 1999. A way with common words. In *Out of Corpora: Studies in Honour of Stig Johansson*, H. Hasselgard & S. Oksefjell (eds), 157–179. Amsterdam: Rodopi.
Sinclair, J. M. 2001. Review of 'The Longman Grammar of Spoken and Written English'. *International Journal of Corpus Linguistics* 6: 339–359.

Sinclair, J. M. 2003. *Reading Concordances*. Harlow: Pearson Education.
Sinclair, J. M. 2004. *Trust the Text: Language, Corpus and Discourse,* R. Carter (ed). London: Routledge.
Stubbs, M. 1996. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.
Stubbs, M. 1997. Whorf's children: critical comments on critical discourse analysis. In *Evolving Models of Language*, A. Wray & A. Ryan (eds), 100–116. Clevedon: Multilingual Matters.
Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Stubbs, M. 2003. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7: 215–244.
Teubert, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10: 1–13.
Teubert, W. & Čermáková, A. 2007. *Corpus Linguistics: A Short Introduction*. London: Continuum.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.
Widdowson, H. G. 2004. *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell.
Zipf, G. K. 1935. *The Psychobiology of Language*. New York NY: Houghton Mifflin.

Appendix 1

Random 100-line concordance sample of ***of*** in HistArt

l inheritance. The Portuguese revolution	**of**	1974 emerged from these irreconcilably
on the plague of fire-sickness and peace	**of**	994 opens with a reference from the la
d Franco, whose ascension was the result	**of**	a civil war, both had much greater roo
es. The sufferer's own silence is typical	**of**	a larger phenomenon. Before the days o
ls required of notarial practice. Instead	**of**	a long, cumbersome list of names, the
se that could be diverted to the building	**of**	a new church. But La Cour had no deep
ion between them and the essential values	**of**	academic life.[42] Ironically, o
hing number of countries allowing freedom	**of**	action to their own municipal organis
movement fuelled by the post-Suez growth	**of**	Arab nationalism. During the course of
ecurity pact at Locarno. This combination	**of**	blind spots with regard to security –
[69] See case in 819/2/28.[70] See list	**of**	cases in 59/2/8.[71] See cases in 819/
erge Bolshakoff, Secretary	**of**	the Academy of Christian Sociologists. The usual argu
istracyjnym ustroju Polski' (The Position	**of**	Cities in Polish administrative Struc
est African Students Union and the League	**of**	Coloured Peoples, the British ILD coll
he proscribed classes. Even the secretary	**of**	commerce and labor recognized the diff
ion and the Holy Office, along with books	**of**	common errors like Aldana's and litera
mage, he won such favour with the Emperor	**of**	Constantinople that the latter insiste
emperament'.[21] One particular consortium	**of**	construction companies systematically r
lengthy letter, Chulaki accused them both	**of**	cosmopolitanism and formalism, and no
y also served the complementary component	**of**	cultural stewardship: the rational use
ng goods that had different time patterns	**of**	demand, it was tempting to conclude th
charters are available for a fair number	**of**	dioceses. Careful examination of this
d by improving the internal organisation	**of**	enterprises. Inefficiency, however, is
of course, been a long and deep tradition	**of**	episcopal responsibility for the poor
h the Germans in the Vichy era. The cases	**of**	(ex-Socialist) Marcel Déat and (ex- C
r of the polyptych: the Blessed Margarita	**of**	Faenza.[76] Another possibility is the t

ties, estimated that the urban population of Florence in the decade between about 1
c use were then replaced by other sources of foreign exchange.[60] The crucial point
ientists, G. E. Shchurovskii, a professor of geology at Moscow University and presi
which was expanded by Nigel son and heir of Henry atte Wode after 1242–3 by his pr
riously deficient" precisely at the level of ideology critique, or of historically
minals, and animals.[7] But the attribution of insensitivity to the "primitive" had n
Hall, et al., eds., The Cambridge History of Japan, 6 vols. (Cambridge, 1988–99), 5
stifiably be be compared with the rescue of Jews in German-occupied Europe. Such a
and relating it to the English conception of kingship that developed over the sixte
icipated, would follow the implementation of land reform. The first account of the
except English quartermasters, consisted of lascars, while the firemen and stokers
ty of the frontier is typical of a number of literary sources. "Mi padre era de Ron
ut-and-out failure.[56] Would-be reformers of local government pointed out that the
ion of the laws acted to take weapons out of men's hands. The courts increasingly t
the early twentieth century, the meaning of military mobilisation had expanded to
dent in an ongoing feud been two factions of mill workers, each from a different re
ially favoured agrarian groups. A change of mind was brought about in the1930s, wh
mportance of establishing the institution of Mothers' Day in Greece.[58] For excel
ve historical sources and the elaboration of native histories by native historians.
ll, yet held proportion with the revenues of neighbouring Princes and States about
shness by one of the best-selling authors of non-fiction in inter-war Britain, H.
d policing. Despite the conservative aims of officials, support for novel forms of
iberal positions, engendered a rejection of parliamentary and democratic values wh
operation of reason and the convergence of politics and morality, just as the fre
o this situation through an ???? reprint of Principles of Political Economy, wher
territory beyond the scope and competence of private associations, officialdom tried
ources are no more explicit.[13] As Humbert of Romans pointed out a century later, fo
me time. The print denies the possibility of solidarity between the Japanese and
icestershire), of which the Austin canons of St Mary's, Leicester, held the advowso
ear of the poor tables began on the feast of St. John the Baptist (June 24), except
private, as Maurice Faure, then Secretary of State for Foreign Affairs, told Maudl
the PAIGC's deployment for the first time of Strella ground-to-air missiles. In a w
es was precisely this non-national nature of television coverage. The northern con
whose approach met the central challenge of the 1920s: he sought transatlantic sol
1931), 245–46; see also Mandeville, Fable of the Bees, 1: 169.[42] Adam Smith, "
alentine, ed., Manual of the Corporation of the City of New York for 1860 (New Yor
superintendance, management and control' of the company's affairs. The auditors' c
The Liber feudorum maior of the counts of Barcelona: the cartulary as an express
y of incurable delinquents', and a study of the creation of anthropological labora
ed as a powerful reminder to London women of the dangers of public life and the sup
panese soldiers took the heads of several of the dead, including the leader of the
stablished "Brigade Eight" to honor eight of the defendants: "To the eight Negroes
aley, informed ministers that the design of the `dished' end-cap had been based on
s still more illustrative. First, because of the expression often used by merchants
ring. For Jackson E. Reynolds, president of the First National Bank of New York, i
Mme. Sun Yat-Sen, American-educated widow of the first leader of the Chinese republ
he later twelfth century until the middle of the fourteenth century.[5] Sources
nge. In a sense, this was the real limit of the Giolittian strategy; the narrow co
Two articles dealing with the aftermath of the Hungarian Revolution in the Frenc
the principal Italian mercantile centres of the later Middle Ages: specifically, w
ugustine seems to have begun on the heels of the Lyons Council of 1274, lending cre
d in the patristic era, several adherents of the new apostolic life cultivated this
dubbing of a knight is the most familiar of the new ceremonies, and the presentati
m the German villages and deport them out of the region and to apply the harshest m
image of Algeria as "a small-scale model of the relationship between the underdeve
lmost exactly reproduced from ordo 40.102 of the RGP.[25] The blessing of a cross als

. The central space just opposite the top of the stair became the Governor's Room a
Hanoverian monarchs as mere 'occupants' of the throne and unlawful magistrates.
gne Vol. 143, no. 291/EU; Luard, History of the United Nations. II, 89–90; for det
rm what was different about the approach of the women in WILPF, and resulted in a
y of respectable women into what one turn- of-the-century writer termed the "Night S
ry and considered expendable – than to that of their alleged union allies.[46]
t cheering accounts are daily reaching us of their success in New York," Trench inf
were executed in the city of Ma¬laga, two of them by garrote vil, a brutal form of
15 minutes readiness. (51) The proportion of Thor missiles at 15 minutes readiness
aturgic miracles account for the majority of those attested to in canonisation proc
isitions ranged from those from the lords of Triennel who had been so instrumental
imony and traveling to estimate the value of various laundries in order to ascertai
. L. Bretton, Stresemann and the Revision of Versailles; a Fight for Reason (Stanfo
a change? If we wish to gauge the impact of war on the development of sexual ident
alcoholism in 1952 – apparently another of Waz«yk's lost souls.[93] Personal catas
ses, Ruth Benedict argued that the spread of "Western Civilization" was due only to
cial Order. The Debate over the Fertility of Women and Workers in France 1770–1920
origin – is, in effect, assigned the value of zero. Figured in relation to this init

Hyperlinks

Keywords or key words?*

Jukka Tyrkkö
University of Helsinki, Finland

The hyperlink of digital hypertext is a text organizing device which allows the creation of alternative structures through texts organized as networks. This paper examines the inherent keyness of the hyperlink by comparing hyperlinks to statistical keywords derived from text through computational means. The argument is made that because hyperlinks are selected by a human author for a discursive, coherence promoting purpose, they possess an inherent keyness and are key words, but do not necessarily correspond with the statistical keywords of the text.

1. Introduction

Over the last decade, an increasing amount of scholarly attention has been paid to the new textual features that have come about as a result of innovations in digital text technology. In addition to new means of near-instant communication (text messaging, e-mail) and the communal production of texts (blogging, discussion forums), the explosive growth of the digital media over the last decade and a half has made electronic *hypertext* a text type familiar to everyone. The most conspicuous identifying feature of hypertext is the *hyperlink*, an interactive and overtly marked discourse marker that makes it possible for texts to be read not as unilinear sequences of chunks of information, but instead as networks designed to be traversed in alternative orders. Because the successful negotiation of the coherence relationship between two fragments connected by a hyperlink depends in part on the process of forming expectations about the potential follow-up

* Research for this paper was conducted under funding by the Research Unit for Variation, Contacts, and Change in English (VARIENG) at the University of Helsinki, funded by the Academy of Finland. The issues raised in the paper will be expanded on in the author's doctoral dissertation, (in preparation).

fragment of text, the hyperlink is placed in a semantically salient position within the text. However, due both to the characteristic brevity of the typical hyperlink and the ambiguity brought about by potential discursive redirection, the referential function of the hyperlink is prone to semantic imprecision which may challenge coherence production.[1] A seemingly contradictory situation thus presents itself, as hyperlinks are at once *functional keywords* essential to the coordination of textual continuity, and *loci* where the text often appears to challenge and even lose coherence.

This paper discusses hyperlinks as *keyness* possessing discursive devices and compares their usefulness as topic indicating devices to statistical keywords. I take as a starting assumption that as overtly marked lexical items invested with the function of coordinating thematic relationships between chunks of text, hyperlinks can be assumed *a priori* to be salient elements in a hypertext. Therefore, it may be posited that hyperlinks possess a quality of semantic keyness distinct from that indicated by keywords, a position which leads to the discussion of the conceptual difference between *keywords* and *key words*, the former meaning lexical items possessing statistical keyness and the latter items perceived by human readers as key. The implications of this claim are contextualized by the means of a case study examining the hyperlinks of a work of hypertextual fiction.

2. Statistical keywords and discourse topics

Over the last few years, the combination of ever-increasing computing power and the development of software specifically designed for corpus linguistic analysis has led to new analytical paradigms and research methods. One of the most interesting of these, *keyword* analysis, has come to be viewed as an effective and useful method for identifying the discourse topic and stylistic features of texts through statistical means.[2] A potential source of confusion is that the corpus linguistic sense of *keyword* is markedly different from earlier usage by e.g. Williams (1976), who essentially picked a set of words indicative of major cultural issues

1. We shall assume the view that coherence is the product of readerly processing and not a quality or characteristic of a given text.

2. Keywords are most commonly identified from corpora using either chi squared or log likelihood methods. Depending on whether the difference in frequency of a particular lexical item is significantly higher or lower than the frequency of the same item in a larger reference corpus, lexical items are described as keywords or negative keywords (Scott & Tribble 2006: 55–88).

and then discussed each one.[3] Stubbs (1996 and 2001) employs the term 'cultural keyword' when talking about lexical keywords, suggesting that they can be used to identify topics central to the discourse of particular cultures. The related concept of *aboutness*, sometimes employed roughly synonymously to *discourse topic*, has been widely adopted in text linguistics since Phillips (1989). Some sort of relationship is often posited between statistical keywords and aboutness, the argument being that unusually frequent lexical words differentiate texts from each other and thus indicate the prevailing topic of the text.[4] Baker (2006: 127), for example, using the classification of keyword types[5] by Scott (1999), describes lexical keywords (nouns, adjectives, lexical verbs) as 'aboutness' keywords, i.e. words that can be used in identifying what the text is about. So long as the keyword list is sufficiently rich in nouns and lexical verbs, and the text under investigation is non-fiction,[6] it is certainly true that keywords – particularly if they can be construed to form lexical fields – can often be successfully used as the basis of identifying discourse topics.[7] The keyword-based analysis of the topical structure of texts can be further operationalized under the paradigm of *keyword linkage* (Scott & Tribble 2006: 73–88). A set of keywords shared by two texts or parts of texts are considered to suggest a heightened likelihood of a shared discourse topic, with the clustering of keywords, particularly of s shared lexical field, indicating a growing topical certainty.

However, it is important to bear in mind that the identification of keywords in the corpus linguistic sense is performed by processing data using statistical formulae, and it can therefore never be fully analogous to human cognitive processing of text, which relies on introspective examination based on understanding of language, text, and culture (see e.g. Andor 1989). The lists of words produced by keyword analysis are often not even close to what a human reader is likely – or even able – to identify as significant. This is particularly true when it comes to grammatical keywords (e.g. pronouns, modal verbs, and so on), which in general are both too frequent in text for a human to evaluate accurately and

---

3. Firth (1935) used the term "focal word" to discuss "sociologically important words." Like Williams, Firth did not use statistical methods or corpora to define such words.

4. See e.g. Bassi (this volume).

5. According to Scott (1999), the typical list of keywords in any text consists of proper nouns, lexical words, and grammatical words.

6. Text linguistic studies have been heavily weighed toward non-fiction texts. For arguments and explanations, see e.g. Winter (1982), Phillips (1989), and Hoey (1991).

7. Scott and Tribble (2006: 161–177), for example, show how keywords can be used for identifying the discourse topics and stylistic choices of newspaper articles.

too transparent to take notice of.[8] The strengths of human readers are quite the opposite. Humans have the ability and inclination to make interpretative, holistic observations about a text, and thereby to understand the keyness of words in a qualitative, rather than quantitative, fashion.

3. Hyperlinks: Cataphoric references – and key words?

As already noted in the introduction, hyperlinking is an exceptionally salient type of endophoric reference and yet, paradoxically, notably imprecise in its application. Under the general referential paradigm of hyperlinking, the author of a text picks a lexical string from the running text, and highlights it to communicate to the reader that another fragment of text, relevant to the link element, can be opened by choosing the link.[9] For the reader, navigating through the text means having to use the partial cues provided by the hyperlink – usually between one and four words in length – to form expectations concerning the direction the text is likely to take. Although the forming of expectations is a normal part of the human reading process (cf. Hoey 2001: 18–31), hypertext raises the bar considerably by turning it into a principal means of textual macro-coordination (Tyrkkö 2007). Also, because the choosing of links requires the reader to make a conscious choice, the very paradigm of predicting how the text is going to continue changes from an unconscious and cognitively transparent process into something considerably more conspicuous.

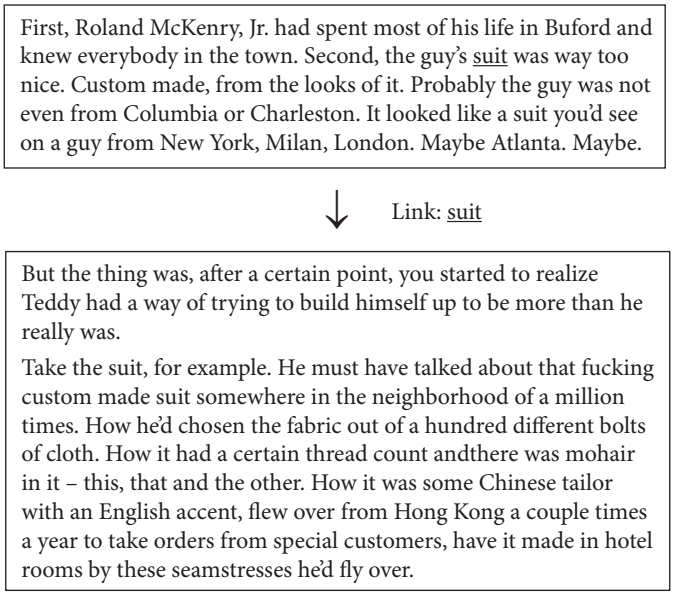
The type of referentiality where a word refers to something in the text that is yet to come is called *cataphoric*. With hyperlinks, cataphoricity is created in two primary ways. The link element – i.e. the word or words of the link – can form a lexically cohesive bridge from the source fragment to the target fragment (see Example 1) or function as a discourse label, in which case the link element describes the target fragment in some way. Significantly for coherence negotiation,

---

8. Grammatical keywords do, of course, have many intriguing uses in linguistic analysis. They can, for example, be useful in the discovery of the stylistic features a particular text or genre or the salient features of language produced by learners.

9. Not all hyperlinks follow this referential paradigm. Although considered bad web design practice, some web designers continue to use links such as click here or ". . such as this." It may be argued, however, that such 'deictic' links do in fact function much in the same way as more descriptive links, the only difference being that the referential string is not in the link itself but is instead given before it, e.g. "If you're interested in fishing, click here." Here, most readers would form the expectation that the target fragment is likely to be about fishing. My thanks to Mike Scott for bringing this point up at the *Keyness in Text* conference in Siena.

unless the author of the hypertext chooses to follow a systematic approach when it comes to the logic of linking, the reader can not know which type he or she should expect.



Example 1. Extracted parts of two fragments *The Heist*, linked by suit

The referential salience of the individual hyperlink depends greatly on the lexical item – or, in the case of a multiword link element, lexical items – of the link element. As discussed by Storrer (2002), lexical cohesion is established between hyperlinks and their target fragments more or less similarly to what is conventionally done between sentences (see also Jucker 2002: 44–45). However, because the hyperlink is principally a device for discursive redirection of one kind or another, salience of referent determination is usually even more necessary than in running text. Jucker (2002: 41–42) uses the term "semantically filled link" to describe hyperlinks that refer explicitly to an identifiable referent. Proper nouns and concrete nouns can generally be expected to form referentially more or less straightforward bridges, while abstract nouns require more mental processing: in other words, it is easier to envision the potential referent of a link that reads John or the moon than it is with links such as jealousy or quickly. Verbs, adjectives and adverbs, as well as multiword units, present ever increasing referential challenges, and are for that reason often avoided in hypertexts intended for easy readability. However, the potential for occasional miscommunication remains even with single nominal link elements, for two reasons. First, because the author's and reader's understanding of the structure of the text are necessarily different, and, consequently the author

may unintentionally (or intentionally) construct a link the proper understanding of which requires contextual knowledge that the reader does not have (see Tyrkkö 2006). Second, the author and reader may have different understandings of a particular lexical items meaning, which leads to referential ambiguity. Recent empirical work by Morris & Hirst (2005) has demonstrated the extent to which lexical cohesion is greatly contingent on the subjective interpretation of interlocutors. For example, if the author uses the lexical item *dog* to evoke a link to a sexually promiscuous male human, and the reader does not know that the word has such a connotation, a hyperlinking which builds a cohesive bridge on that particular sense would be unintelligible for that particular reader – both in terms of forming an accurate expectation prior to link activation, and possibly even after the target fragment has been revealed. To derive the likely referential meaning, readers need to interpret the link element in conjunction with the co-text that has already been read – including familiarity with the dominant linking strategies employed in the particular hypertext.

In contrast to conventional linear text, in which lexical cohesion normally serves the purpose of creating local coherence, hyperlinking utilizes the means of local cohesion to form continuities to other parts of the global textual space. In this sense, and bearing in mind their nature as particularly salient textual elements, hyperlinks can also be considered discourse markers as defined by Schiffrin (1987: 314). Although hyperlinks are not a closed set of lexical items, they share many of the distinguishing features of discourse markers. Most importantly, hyperlinks function as text organizing units and have a significant role in promoting coherent transitions between text fragments. As text organizers, hyperlinks frequently serve as discourse labels, i.e. as textual devices which provide an encapsulated description of the overall discourse topic of the target fragment, much in the same way as a chapter heading or the title of a book both describes and defines the text in question. The shifting of discourse topics, or discursive redirection, is a fundamental function of the hyperlink and a central paradigm of hypertext. Because hyperlinking is usually employed when the author wants to provide additional information on a particular topic, the link element is likely to be interpreted by the reader as signaling a point in the text at which a new discourse topic is introduced.

On the basis of these functional features, the question can now be raised whether or not hyperlinks should be considered *key words* in a hypertext? Let us begin with the notion of keyness not as a statistical feature, but as perceived by the human reader. Scott & Tribble (2006: 55) define *keyness* as "a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail." From this perspective, we may begin by noting the significance of the markedness of

hyperlinks in text. As demonstrated by Black, Wright, Black & Norman (1993), overt marking invests hyperlinks with a clear salience among the words of the running texts, pointing attention to them and signaling their prominence in the textual space. Taken together with the pragmatic reality that readers are aware of the function of the hyperlink, it is likely that they perceive each link as relevant to textual coherence; that is to say, links are where they are for a reason. Both Pajares Tosca (2000) and Tyrkkö (2006) have discussed hyperlinking from the perspective of relevance, arguing that hyperlinks play an important role in coordinating the interaction between the text and the reader. Indeed, it is implicit to the use of hyperlinks that they should be both meaningful and relevant; after all, the readers would be highly unlikely to keep reading a hypertext if the choices offered to them made no sense and kept leading to unpredictable target fragments (cf. Engebretsen 2000). This doesn't mean that a reader needs to be correct in his or her expectation about the target fragment of each and every hyperlink, but it does mean that it should be quite rare for him or her to not understand the coherence after the linking. We may also claim that the keyness of hyperlinks can be derived directly from their discourse labeling function. If the hyperlink is a word or word group which mostly saliently communicates the core information content of the target fragment, it makes sense to surmise that the hyperlinks of a text can be deemed as its key words.

All the arguments presented above would support the notion that hyperlinks are key words in the subjective, human evaluated sense.

4. Hyperlinks and keywords: A contrastive case study

To illustrate the relationship between hyperlinks and keyword in a work of hyperfiction, we'll examine a work of hypertextual fiction as a case study. *The Heist* (1996) by Walter Sorrells is a hyperfiction story about the events leading to and taking place during a bank robbery in a small country town in North Carolina. Due to the virtually endless number of permutations the reading of the text can take, *The Heist* is at once a relatively small and a surprisingly complex short story. To examine the relationship between hyperlinks and the textual fragments of *The Heist*, all the fragments were compiled as a corpus and a keyword analysis was performed on each fragment using *Wordsmith 4* with the *British National Corpus* (BNC)[10] as a reference corpus. Next, all the hyperlinks of *The Heist* were extracted from a structural map created on *Cmap Tools*, followed by the comparison of the

10. The BNC was used as a reference corpus for reasons of availability at the time the study was conducted. Ideally, an American reference corpus would have been used.

hyperlinks to the keyword lists of both the fragment in which they were found (source fragments) and the fragments they were linked to (target fragments.)

Let us begin by looking at the extent to which the hyperlinks actually function as cohesive devices. To do this, the links were compared to the full text of the target fragment to ascertain cohesion between the two. Lexical cohesion[11] was considered to exist when repetition, reiteration, co-reference or collocation was found to exist between the hyperlink and the target fragment. In the case of multiword links, cohesive relationship was deemed to exist if any of the lexical words in the multiword link formed a lexical bridge to the target fragment. The analysis shows that lexical cohesion can be found between the hyperlink and the target fragment in 88% of the cases, with a further 11% exhibiting a discourse topical, descriptive link relation. Discourse topical linking is characterized by the hyperlink functioning as a descriptive label of the events depicted in the target fragment. For example, one of the hyperlinks, incident with a punk, links to a fragment describing an altercation between a group of would-be bank robbers. Neither of the two content words of the hyperlink, *incident* and *punk*, appears in the target fragment, nor is there any apparent instance of reiteration of either item or their salient collocates. Nevertheless, a human reader will have no trouble understanding the continuity, namely, that the fragment in its entirety describes an incident and that one of the fictional characters could be described as a *punk*. Only 1% of the hyperlinks of *The Heist* show no apparent cohesive continuity at all (Figure 1).[12]

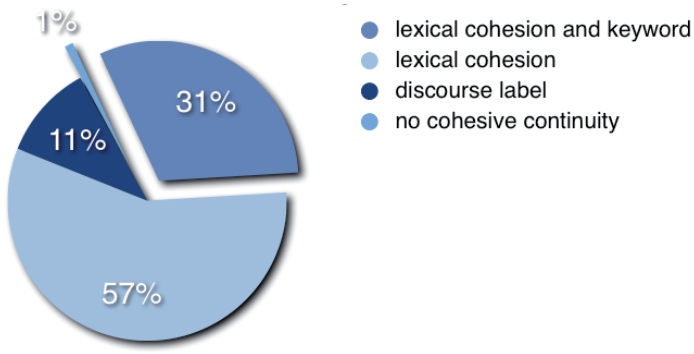


Figure 1. Hyperlinks and keywords in *The Heist*

11. Halliday and Hasan (1976); Morris & Hirst (2005).

12. The apparent lack of cohesiveness in a hyperlink does not necessarily mean lack of coherence. As discussed by e.g. Louwerse (2004), although lexical cohesion is a strong cue for coherence it is not a requirement. See Tyrkkö (2006).

On the basis of these initial findings, the hyperlinks of *The Heist* would appear to be key words, as defined in the previous section. But does this mean that the lexical field they form is similar to the statistical keywords of the text fragments? If the dual hypothesis is true that keywords identify aboutness, and that hyperlinks have a discourse topical referential function, the two lists should appear relatively similar. The list of the hyperlinks of the first part of *The Heist* (Table 1) immediately shows at least two things. First, *The Heist* makes use of a lot of multiword links. Second, and more important to the present discussion, the list would appear to make it possible to hazard relatively far reaching estimates about the discourse topic of the text in question. Apart from the proper names, which identify all of the major characters in the story, the hyperlinks evoke ideas of criminal activity, the police, and something to do with a bank. Furthermore, the repetition of certain words and word groups like *highly developed sense of irony* and *suit* suggest they may possess particular *keyness* in the text.

Table 1. Hyperlinks of *The Heist*, Part 1

rust colored suit, THE HEIST BEGINS, look sharp, highly developed sense of irony, nine hundred dollar suit, idea, The bank manager, suit, KST-464, heard anything, Mr. McKenry Sr, Roland McKenry, Jr., Roland McKenry, Sr., Spring Lake Plantation, gone off on the new guy, his daddy, paydirt, silent alarms, Columbia, Noreen, Farmer's Community Bank, gotten out the door, new guy, suit, story about Mo and the new guy, video cameras, "I'm liking this dude already", suit, highly developed sense of irony, cleft in his chin, chief of police, Roland McKenry, Sr., looking at you funny, Jim Beam, Buford, South Carolina, Bug Something-or-other, Teddy Clapp, drank like a fish, middle aged chick with big knockers, Spring Lake Plantation, Spring Lake Plantation, Roland McKenry, Farmer's Community Bank, Ed Lampier, cellmate at that Federal camp in Alabama, development, highly developed sense of humour, bright idea, Ed Lampier, suit, Chief Loy, poor dumb bastard, his dad's bank, Jew stick-up man, Buford, North Carolina, bad idea, Teddy Clapp, THE HEIST BEGINS, certain feelings, local, bank, your friend, walked away, video cameras, highly developed sense of irony, highly developed sense of irony, dad, ready, willing and able, bank, squeeze it in, tell us something, bank's layout's pretty simple, nine hundred dollar suit, nose, THE HEIST BEGINS, his son, jew bank robbers, incident with a punk, J. C. Penney, silent alarms, suit, trust, THE HEIST BEGINS, the phone rang, suit, suit, Mo Rosen, ridiculous looking overalls, Mr. McKenry Jr., suit, layout of the bank

However, if we then turn to the list of keywords (Table 2) from the same part of the story, it quickly becomes apparent that it contains a considerably higher proportion of common lexical words which do not seem to form such obviously coherent lexical fields as the hyperlinks did.[13]

13. The keyword list is considerably shorter than the hyperlink list due to statistical issues related to computing keywords from short text fragments. Allowing a higher p value would yield more keywords, but doing so would soon render the exercise almost pointless as more and more words would be returned as key.

Table 2. Keywords of *The Heist* ($p < 0.001$), Part 1

Clapp, Teddy, Guy, suit, South, Colombia, modern, farm, brochure, fat, Moe, tea, Ed, bug, from, town, decade, population, dope, Milan, Roland, thug, known, Jimmy, they, hooked, joint, Jew, jew, criminal, Dutch, Schultz, suit, tailor, bought, criminals, crackhead, drinking, goyishe, federal, that's, Jimmie, banks, percent, five, Jimmie, food, Moe, grocery, robbed, spose, mr, cabe, worked, many, system, art, state, Lampier, stick, Ed, we, Ed, we'll, grand, bird, city, council, night, police, Roland, chief, police, sharp, Gaddis, William, county, fulton, concur, probation, penitentiary, goliath, birthday, Briston, hence, nose, you, Kramer, nose, I'm, plague, big, suit, suits, blue, McKenry, mr, smart, McKenry, square, living, clubhouse, her, she, against, boring, candidate, friends, hinted, heist, crime, you

If anything, it would appear that the list compiled of the hyperlinks gives a more comprehensive impression of the text's aboutness than the statistical keywords. From the hyperlinks we can clearly see that the text is about a town in South Carolina, about something to do with a crime, and a bank. The statistical keywords, on the other hand, would also seem to suggest something to do with crime and criminals, but the list is equally filled with both lexical and grammatical words which a human reader would never pick up as significant. Of course, the difference between hyperlinks and keywords can partly be explained by the fact that the length of hyperlinks is quite often two or more lexical items,[14] whereas keywords are (by default) individual items. As demonstrated by Panunzi, Fabbri & Moneglia (2007), key phrases (three or four term strings including one noun) are clearly more descriptive than single keywords when it comes to correct assessment of aboutness. Intuitively examined, they thus resemble hyperlinks functioning as discourse labels.

Continuing with the relationship between hyperlinks and keywords, the analysis of *The Heist* shows that the lexical items of the hyperlinks match the keywords of the target fragment 31% and the source fragment 23% of the time (see Illustration 1). A match was judged successful if at least one content word of the hyperlink was found on the keyword list for the source or target fragment, respectively. In Example 1 (above), the hyperlink suit forms a cohesive bond to the target fragment by means of simple repetition, in addition to which *suit* is also a keyword of the target fragment. The fact that the hyperlinks are almost equally likely to match a keyword in the source fragment as in the target fragment would seem to suggest that at least in *The Heist*, hyperlinks are almost as likely to indicate the predominant topics of the source fragment (i.e. the fragment where the link is located) as

14. The use of hyperlinks consisting of more than one lexical item is particularly common in hypertext fictions, where the link element is sometimes intentionally ambiguous for literary effect. See Tyrkkö (2006).

they are of the discourse topic of the target fragment. Upon closer examination, we can see that this phenomenon is partly explained by the relatively high usage of proper nouns (personal names, place names) as hyperlinks; references to the characters of a story are a salient linking practice, while proper nouns are also the single most likely word group to show up as statistical keywords. Given that the function of hyperlinking is to establish global level coherence using what is essentially a local level cohesion strategy (see Storrer 1999 and Tyrkkö 2007), the need for a meaningful continuity between the hyperlink and the target fragment is paramount. Repeated instances of discontinuity between the two result in the diminishing of overall meaningfulness of the text – after all, why would a reader be motivated to select a particular hyperlink over others if the lexical content can not be taken as a reasonably reliable cataphoric reference? Although it could be argued that the hyperlink not so much functions as a discourse label but as a billboard attracting attention to the target fragment,[15] this would be to suggest that hypertext is essentially a collection of independent fragments of text rather than chunks of a single text, albeit one that is read in alternative sequencies.

Leaving hyperlinks aside for a moment, we find that keyword linkage of one or more items is to be found in only 18% of the cases between source and target fragment. Taking keywords to indicate discourse topic, we may interpret this figure to demonstrate that, at least in *The Heist,* a high incidence of topical shifting takes place from one fragment to another. Keeping in mind that the functional *raison d'ĕtre* of hyperlinks is to enact a discursive redirection, this seems a reasonable finding. Given that the overall lexical cohesion between hyperlinks and target fragments is 88%, the findings suggest that in this particular text, hyperlinks perform as reliable coherence cues, which in turn reinforces the reader's subjective impression of hyperlinks as key words. While this does not suggest that the presence of lexical cohesion alone would be sufficient grounds for successful coherence negotiation, the question is raised whether hypertextual linking may be a textual feature that emphasizes the cueness of lexical cohesion. Under this paradigm, the markedness of hyperlinks in text can arguably elevate the coherence-forming properties of the lexical items used in the cohesive bridge over the fragment boundary. Because the reader is not only aware of the lexical content of the hyperlink to a considerably greater extent than what would be typical with other items of running text, but also explicitly selects one on the basis that is suggests a potentially interesting discursive continuity, finding the other end of the lexical chain in the target fragment can be taken to be akin to a non-statistical analogue of keyword linkage.

---

15. An example of a type of hypertext where a hyperlink is often primarily intended to attract attention rather than inform is online advertising; see Janoschka (2004).

5. Conclusions

Hyperlinking is a prevalent feature of digital textuality which requires close analysis before text-linguistic and discourse analytical models can be reliably applied to them. The case study presented above does not provide enough empirical evidence to account for the vast field of hypertextual practices, but it does suggest that the concept of keyness can be successfully employed in the analysis of hypertextual structures and of the semantic implications of hyperlinking. In this paper, the suggestion has been made that hyperlinks, on the merit of their inherent markedness and functional properties, can be considered a surface level equivalent to statistically derived keywords; that is to say, that the hyperlinks of a text form a lexical field which can be taken to indicate the key issues of the hypertext in question. In light of the empirical findings of the case study, this argument is paradoxically both true and false. True, because the list of the hyperlinks of *The Heist* did indeed appear to represent the subjectively evaluated key elements of the story, but also false, because the hyperlinks of an individual text fragment did not exhibit a high level of correspondence with the statistical keywords of either the source fragment or the target fragment.

This paper has also raised a broader theoretical issue; namely, that the keyness of hyperlinks demonstrates the importance of examining, and conceptualizing, keyness on two different levels. On the one hand, the findings indicate that hyperlinks are not a reliable indicator of the statistically identified discourse topic of either the source or the target fragment. This need not discount the claim that hyperlinks are *key words*, but simply that they are not *keywords*, in the statistical sense. On the other, hyperlinks are arguably highly salient textual elements which, for the reader, are inevitably indicative of the discursively important topics; sign posts on the path through the text, one might say. Thus, hyperlinking in a sense reconceptualizes Firth's (1935) "focal words", only investing them with overt markedness and adding the redirective functionality made possible by the digital medium.

References

Andor, J. 1989. Strategies, tactics, and realistic methods of text analysis. In *Connexity and Coherence: Analysis of Text and Discourse*, W. Heydrich, F. Neubauer, J. Petöfi & E. Sözer (eds), 28–36. Berlin: de Gruyer.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Black, A.,Wright, P., Black, D. & Norman, K. 1993. Consulting on-line dictionary information while reading. *Hypermedia* 4(3):145–169.

Engebretsen, M. 2000. Hypernews and Coherence. *Journal of Digital Information* 1: 7. <http://jodi.tamu.edu/Articles/v01/i07/Engebretsen/> (21 July 2006).
Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society*, 36–72.
Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.
Hoey, M. 1991. *Patterns of Lexis in Text*. London: Routledge.
Hoey, M. 2001. *Textual Interaction. An Introduction to Written Discourse Analysis*. London: Routledge.
Janoschka, A. 2004. *Web Advertising. New Forms of Communication on the Internet* [Pragmatics & Beyond 131]. Amsterdam: John Benjamins.
Jucker, A. 2002. Hypertextlinguistics: Textuality and typology of hypertexts. In *Text Types and Corpora. Studies in Honor of Udo Fries*, A. Fischer, G. L. Tottie & H. M. Lehmann (eds), 29–53. Tübingen: Gunter Narr.
Louwerse, M. 2004. Un modelo conciso de cohesion en el texto y coherencia en la comprensión. *Revista Signos* 37: 41–58.
Morris, J. & Hirst, G. 2005. The subjectivity of lexical cohesion in text. In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu & J. Wiebe (eds). Dordrecht: Springer. <http://ftp.cs.toronto.edu/pub/gh/Morris+Hirst-2005.pdf> (21 July 2006).
Pajares Tosca, S. 2000. A pragmatics of links. *Journal of Digital Information* 1(6): 77–85.
Panunzi A., Fabbri, M. & Moneglia, M. 2007. Multilingual open domain keyword extractor proto-type. Paper read at Keyness in Text, Certosa di Pontignano, 29 June, 2007.
Phillips, M. 1989. *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham.
Schiffrin, D. 1987. *Discourse Markers*. Cambridge: CUP.
Scott, M. 1999. *Wordsmith Tools Help Manual*. Version 3. Oxford: OUP.
Scott, M. & Tribble, C. 2006. *Textual Patterns. Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.
Sorrells, W. 1996. *The Heist*. < http://www.waltersorrells.com/2.html> (15 June 2007).
Storrer, A. 1999. Was ist "hyper" am Hypertext. In *Sprache und neuen Medien, Jahrbuch des Instituts für deutsche Sprache*, 222–49. Berlin: de Gruyter.
Storrer, A. 2002. Coherence in Text and Hypertext. *Document Design* 3(2): 156–68. <http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/pdfs/as-paper.pdf> (21 November 2008).
Stubbs, M. 1996. *Text and Corpus Analysis – Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.
Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Tyrkkö, J. 2006. Reading as interaction. In *Dialogic Language Use* [Mémoires de Société Néophilologique LXVI], I. Taavitsainen, J. Härmä & J. Korhonen (eds), 123–146.
Tyrkkö, J. 2007. Making sense of digital textuality. *European Journal of English Studies* 11(2): 147–161.
Williams, R. 1976. *Keywords: A Vocabulary of Culture and Society*. London: Fontana.
Winter, E. O. 1982. *Toward a Contextual Grammar of English*. London: George Allen & Unwin.

Web Semantics vs the Semantic Web?

The problem of keyness

François Rastier
CNRS/INALCO, France

The *Semantic Web* programme aims to replace the "Web of Documents" with the "Web of Data", thus extending the classical programme of knowledge representation. In contrast, a corpus-linguistic inspired *Web Semantics* situates knowledge within texts and the documents that convey them. Data cannot therefore be abstracted without losing their contextual value and relevance. This leads to a recontextualisation of the notion of "data" and a rethinking of the relationship between data and metadata.

1. The ambitions and credibility of the Semantic Web

The Web operates according to three standards: HTTP as the protocol, URL for addresses and HTML as language. Tim Berners-Lee, the director of W3C, the body governing the destiny of the World Wide Web, has presented the Semantic Web since 1994 as an extension of the Web which would transform it into a zone of exchange of documents which allows access to their *content* and also allows *reasoning* to be carried out. This would require document content to be represented by ontologies using denotational semantics (the Semantic Web recognises no other kind); all contributors to the Semantic Web, and soon all those who put content on line, must therefore respect a common infrastructure as represented by the famous "layer-cake" presented by Tim Berners-Lee at the XML 2001 conference (Figure 1).

The normalisation of this infrastructure is now considered operative at the level of ontologies: in particular, ontologies provide the vocabulary of the metadata used to represent the content of the documents in the same way as a thesaurus, composed of terms (or concepts) and not of words. It must be noted that above the second level, all contact with texts and languages is lost, as (formal) languages of representation are used instead. In promoting the Semantic Web, the W3C intends to replace the "web of documents" with the "web of data" (Berners-Lee

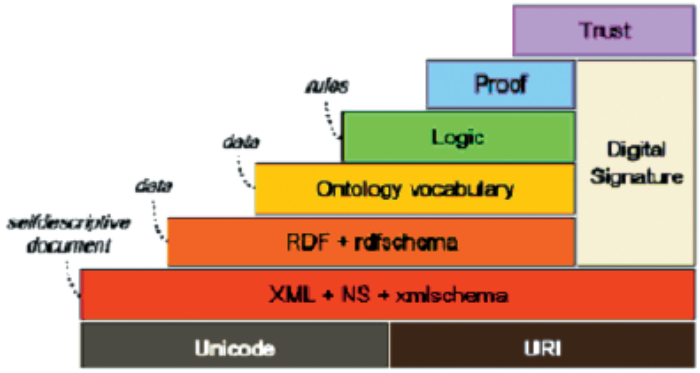


Figure 1. The Semantic Web layer-cake (as seen by the W3C)

2007). By using ontologies, the aim is freedom from the complexity of documents and their linguistic and semiotic diversity.

In accord with the objectivism of the philosophy of language which stems from logical positivism, an item of data is seen as a simple chain of characters (e.g. the French "pêche", which can be linked either to fish or fruit; Berners-Lee 2007). It would not be courteous to insist on the banal poverty of this conception of data.

The recommendations of the W3C, reassuring enough when they are presented as being purely practical, are in fact designed to become standards. However, the adoption of "low-level" standards such as HTML, or Unicode, or even XML, in no way entails that languages of representation such as RDF or OWL should be adopted as standard, unless one merely seeks to yield easily to the attempt by the W3C to force through the "Semantic Web".

It would be more discourteous still to insist on the economic aspects of the Semantic Web. It is readily understood why the US Department of Commerce supports the Semantic Web: accessible web content is normalised by means of ontologies, generally composed of English words written in capital letters, supposedly representing metadata allowing access to documents.

Metadata standards are one of the three key elements in the "Net-Centric Data Strategy" of the US Department of Defense, decided upon in December 2001 and made public in May 2003 (Stenbit 2003). This strategy consists in taking control of a network that is, by definition, non-centred, by making certain types of metadata mandatory: "to ensure that all data are visible, available, and usable", thus putting an end, by decision, to the hidden Web that escapes control. As "all posted data will have associated metadata" (Stenbit 2003: 4), the volume of private data will be drastically reduced (divided by two, while "enterprise and community

of interest" data will be multiplied by three; p. 10). The network centralisation thus established (in order to achieve Net-Centricity) will allow "a completely different approach to *warfighting and business operations*" (Appendix A, p. 2, emphasis added).

It is easy to understand why military support has been forthcoming since September 2001: the Semantic Web seeks to be a collaborative effort, with providers of content putting their databases online in a single format which will make them interoperable and thus facilitate access to them – for example, to discover new medicines (according to Tim Berners-Lee 2007). Intelligence, in the economic and military sense of the term has everything to gain from this cooperative transparency.

Political and economic considerations must not obscure the epistemological consequences of this programme. The coherence of the "layer-cake", resembling as it does the delights of Neapolitan ice-cream, may be questioned: it is probably a simple eclectic juxtaposition, but this eclecticism is guided by objectives of dubious scientific status. Anyone who has any knowledge of visual semiotics will recognise in the tiers of the "layer-cake" the steps of a *gradus ad Parnassum*, which leads from *Unicode* to *Trust* (with the meaning of *confidence*, as in "*In God we trust*", rather than with the sense of *monopoly*). In short, standards are established, and then given the status of theoretical models, a procedure characteristic of technoscience, which is not only instrumentalistic but also instrumentalised. Sir Tim Berners-Lee is an engineer; by conveniently proclaiming the creation of *Web Science* in 2007, he shrewdly avoids having scientific problems raised and debated outside the Semantic Web community, which is self-engendered and must therefore undergo only self-assessment.

Future analysts may well query the unity of thought between the W3C and the US Department of Defense. Internet sprang from a totally military contradiction between network security and information control. There had to be a distributed network in order for it to resist all attempts to destroy it; its very success and its extension to the economy and to private data has made it difficult to control. Yet any official apparatus, whether economic or military (and the US Department of Defense is merely a prime example of such a body), must maintain a hierarchy in order to exercise its authority and secure its legitimacy: it must therefore necessarily project its structure on the surrounding world, and we have elsewhere underlined the theoretical relationship, in the most metaphysical sense, between ontologies and organigrams (Rastier 2004). The W3C "layer-cake" creates its own hierarchical hierarchy (ontologies, DTD XML, etc.) and thus subsumes, by means of various levels of metadata and "data", the uncontrollable diversity of documents, and the languages and sign systems that they bring into play.

Thus the title of this study invokes in far too simple a manner two completely different conceptions which are not in direct opposition and yet remain incommensurable. No combat between David and Goliath is to be expected: the Semantic Web is a politico-technical programme while Web Semantics is a methodological project and a field of applications founded on corpus semiotics.

Benefiting as it does from influential support, the Semantic Web also commands easy assent since it embodies a series of conceptions stemming from the classical artificial intelligence tradition which will here be subjected to scrutiny.

2. Habits of thought and current limits

2.1 Postulates for the representation of knowledge

The Semantic Web transposes the paradigm of knowledge representation into a new environment at a completely new scale. Knowledge representation relies on three assumptions that link it to orthodox cognitivism.

i. Knowledge is a representation of the empirical world: the picture of the world as ontological "furniture" is generally established in those circles where logical positivism is the implicit reference. However, there is nothing to justify the assertion, put forward with that *naïve realism* to which certain influential semanticians, like John Lyons, have laid claim, that objects in the world are discrete, countable, public, and everywhere the same.
ii. Knowledge is (relatively) independent of its semiotic substrate, so that its extraction or representation never modifies its content. This axiom was the postulate of the general theory of terminology, put forward first by Wüster and upheld by the Vienna Circle: terms are considered to be language independent.[1]
iii. Knowledge is discrete and formalisable in that it can be represented by logical formalism, in general the logic of predicates. This echoes the duality of empiricism and logicism in logical empiricism: denotation theory ensures the

1. This supposes a duality which is difficult to accept, for concepts are not independent of the texts in which they are defined, organized and modified. Knowledge resides not in terms but in *texts*; moreover, scientific and technical discourse is not transparent, nor is it independent of the language in which it is expressed, even if the international norms of a discipline may facilitate its translation. Knowledge representation proceeds therefore by the collection and analysis of scientific and technical corpora.

empirical grounding of the theory while the use of the logical organon is supposed to bring conceptual productivity.[2]

A non-referential approach is therefore required in order to avoid such metaphysical postulates: the most coherent alternative for the moment seems to be differential semantics.

2.2 Current debate

In the field of knowledge representation, representative formalisms have been elaborated which are considered to be adequate, insofar as they meet the requirements of the less ambitious applications. The adoption of standards such as XML or RDF allows interoperability in principle, but does not solve the problem of the production, identification and evolution of knowledge.

Debate is centred beforehand on the problem of the reification of knowledge outside the context of use, or complementarily on the appropriateness of its mode of representation for its actual use.

Those who hold the reifying view base their claim on the growth of ontologies which radicalise the objectivist preconception of knowledge. Ontologies remain thesauri – that of Roget was explicitly used as a model for WordNet by Miller and his team. They retain all the main drawbacks of their model: a generality that will not allow them to adapt to the selective point of view of a given task, and a lack of evolutivity which entails human maintenance. They reduce language to a nomenclature which describes neither textual structures nor the considerable variation of genres and discourse.

Even for the Semantic Web, a field closely linked to ontologies, a user-centred perspective leads to resigned statements such as the following: "Semantic Web researchers […] accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility." (Berners-Lee, Hendler & Lassila 2001). The variety of user viewpoints and keyness regimes adapted to their tasks will not allow them to rely on a single, and nonetheless arbitrary, norm: but the absence of contradiction remains a firm assumption in ontologies, in accordance with the laws of identity, noncontradiction and excluded middle which underlie their logicistic conception of the world.

---

2. This logical preconception of the world mobilises only a small fragment of bivalent logic and remains equally separate from both multivalent and modal logic. A *fortiori*, it remains strangely distant from mathematics, as it cannot take into consideration the three fundamental problems of this discipline: infinity, continuity and large numbers.

More radically, those in favour of situated cognition and those ergonomists who specialise in information querying are adamant about the unpredictable diversity of applications and insist that formalisms are simply the props for interpretation pathways. If that is so, then the definitory facets of any given object cannot be set in advance: in other words, usage defines key properties for objects.

This divergence can today be resolved empirically. The study of large corpora, including technical corpora, has shown that the semantic relationships organising ontologies varied from one discourse and field to another, to such an extent that some basic semantic relationships are completely absent from certain corpora which were nevertheless of considerable size. (e.g. the Safir project run by the Crim-Lip6-Edf consortium).

The experience of WordNet and EuroWordNet has been edifying: these ontologies have proved to be pointlessly complex. Based on psychological tenets dating from Miller & Johnson-Laird (1976), they ignore even basic linguistic knowledge such as the crucial notion of the morpheme, which leads to the creation of separate subnetworks for nouns, verbs and adjectives. Notwithstanding their unprecedented cost, ontologies are of little use and are generally consulted merely as dictionaries or aids to translation. At internet-scale, the blending of knowledge from different ontologies remains problematic because, even within the same disciplinary field, ontologies are not interoperable, despite all the recommendations on standardisation.

2.3 Ontologies and the Semantic Web

In the field of communication science and NLP, the separation between cognition and communication has traditionally favoured knowledge representation, without paying any particular attention to the production, selection and transmission of knowledge. Information is extracted and communicated, the only condition being that of *information packaging*, seen simply as a way of parcelling out knowledge.

Ontologies are the outcome of the vision inherited from logical positivism: apparently created by nobody in particular for nobody in particular and therefore independent of any given point of view, they are supposed to represent an objective world, depending neither on language, nor on sign systems, nor on any given task. The nomenclature of "world" objects is not problematised, as it is based on shared evidence; for "local" or specialised ontologies, the inventory of entities depends simply on the state of the art as generally recognised.

In this type of representation, the difference between languages fades, as does the diversity of discourse and of related genres: the format of knowledge

in ontological hierarchies remains that found in semantic networks: a thesaurus in *Basic English*, enriched by stereotypical and heterogeneous semantico-logical relationships, such as hyperonymy, meronymy, etc.

The Semantic Web, entirely dependent on such arguments, is tributary to a small number of impoverished semantic universals. Yet it cannot be imagined that the keyness of a word is related to the position of its referent in an ontological hierarchy. Qualitative inequalities in a text are without any definable relationship to the hierarchical position of entities: in general, since superordinate entities are trivial, so the more superordinate a concept, the less subject to debate and the less key it is. Relevance, as a cognitive economy principle (Sperber and Wilson) thus defines not those concepts open to debate, but only the most trivial concepts.

The semiotic richness of digital documents is of little or no account in such a vision, as it is irreconcilable with the referential paradigm and does not contribute to denotation: yet expression-related clues (typography, colour-coding, etc.) may reveal themselves to be highly discriminatory.

As knowledge is never task-independent, the variety of tasks and applications means that keyness regimes should be definable and variable. Each field defines its keyness regime, and therefore a *praxeology* (not an ontology) should be used to determine "key information" in texts and corpora.

2.4 Requirements for linguistics

Internet now chiefly needs to improve search engines, by adapting strategies according to the task and the nature of the documents involved.

This requires *applicable* linguistics which can process texts, analyse their semantics and reflect their linguistic and semiotic diversity. Corpus linguistics is therefore obliged to be inventive. It stems in part from computational linguistics, which raises problems derived from Chomskyan cognitivism (phrase-structure grammar, syntactic tree construction, etc.) and also from lexicometry (rooted in mathematical and statistical linguistics).

Computational linguistics is confronted with obstacles arising from the philosophy of logical positivism (in particular the separation between syntax, semantics and pragmatics). Lexicometry, as a methodology, has no preconceived ideas about language, and is therefore more adaptable. Neither has any theoretical concept of text: for computational linguistics, a text is a series of phrases; for lexicometry, it is a set of words. Both fields are therefore at a loss when confronted with huge, multilingual, polysemiotic corpora, with multiple social and cultural demands.

It is therefore to linguistics as a science of texts, conscious of its place among the cultural sciences, that it falls to propose unification and realignment.

Linguistics must act by considering social demands and not simply by applying theories: it will only be applicable if it becomes *implicated*. Linguistics must be involved, even as an auxiliary method, at various stages: in the creation of software, in corpus building, tagging, testing corpus tools on (tagged) corpora, interpreting and discussing results. At all stages of the chain of processing, linguistic and more broadly semiotic knowledge is indispensable.

2.5 Abolishing text amnesia

The ontological paradigm of knowledge representation remains therefore tributary to an obsolete state of the art, left over from a time when access to complete texts was *impossible*. Thesauri and other formal classifications had to be employed to index texts using a static representation of their presumed content. The drawbacks of this are well known: heavy construction and maintenance costs, insufficient relevance, which could not be modulated to fit the task underlying the quest for information.

The normative point of view is based on methodological and even epistemological oversights which affect: (i) the local and global context of the information in the text; (ii) the context of the corpus within which texts and the information they contain take on meaning; (iii) the points of view which shaped that information and on which it depends; (iv) the various groups for whom the information is destined. To sum up, when context is removed, the context of use on which keyness itself depends is also removed.

These obstacles are inevitable if texts are reduced to "sets of words" without taking into consideration structure, genre, etc. Full text access now provides better solutions, provided that it is guided by the propositions of corpus linguistics. The metadata that accumulate now are obviously not there in order for the data to be forgotten!

3. Reconceptions

3.1 Dynamic elaboration

Our methodological proposition is to base all knowledge representation on the semantic and semiotic analysis of genuine corpora manifesting that knowledge: *knowledge and its "normalising" ontologies must and can be created dynamically, in response to applications and their corpora*. 'Knowledge' is in fact the objective interpretation of texts and other semiotic performances and productions.

Each application defines its own keyness regime within its corpus. No concept is key to all applications. One of the main problems with ontologies is the definition of their "nomenclature": which concepts should be included, when all words in the lexicon are potential candidates, as are multi-word units or clusters. George Miller's practice shows that he uses no other criterion than simple good sense, which means the bias of the creator of the ontology.[3]

If it is admitted that the lexis cannot be organised in a single arborescence, as each discourse and genre has its own lexis, then local zones organised by profile rather than by subsumption must replace the totalising image of the unified network: each concept is a *semantic form*, profiled against a background. Some terms lexicalise forms or parts of forms, others lexicalise backgrounds. The word "*text*" in literary criticism is part of the background; it is not a concept. It forms the basis of expressions such as "*Proustian text*", but it is never found in the context of terms such as "*notion*" or "*concept*".

Semantic forms have *values*, whereas concepts in ontologies do not: in WordNet, the closest neighbour to "caviar" could well be "fish finger". Yet it is obvious, and confirmed by corpus investigation, that the two terms are not to be found in the same contexts (Rastier & Valette 2009). The hierarchy of values outshines the ontological hierarchy which does not take values into consideration.

Concepts may be described as semantic forms belonging to theoretical texts: their diffuse or synthetic lexicalisations, their evolution, from inception to disappearance (by extinction or semantic bleaching), their semantic correlates, their expressive collocates, together form a field of research which has barely begun to be explored.

The alternative that we suggest is that of full text search-engines which take into account the progress made in textual semantics, including: (i) the definition of textual units which are not strictly bounded and sequential ("*passages*"); (ii) the extension of the differential principle of semantics to the contrast of the corpus, between discourse, genres, and text sections; (iii) the analysis of textual genres in zones of differentiated keyness.

At stake lies not the representation but the *production* of knowledge from the vast unstructured data of the web – or preferably from document databases.

The paradigm of knowledge representation must be developed within a semiotic framework. Texts are not merely chains of characters. Their segmentation, their "logical" structure, their typography, and even their tagging, are part of their semiotics. For example, in classical philosophy, the use of capital letters designated the

3. In 2002, he removed the terms "*franc*", "*lire*" and "*mark*" from his ontology, as these currencies were obsolete, and he introduced "*intifada*" and "*bacillus anthracis*" (Rastier 2004).

principle concepts. Scientific and technical texts integrate what could be referred to as "out-text": figures, tables, diagrams, and photographs all belong to the textualisation of knowledge and require, for their processing, multimodal semiotics.

3.2 Textual knowledge

An item of knowledge is a *set of passages* from texts (even multimedia): by recurring, the content of these passages (fragments) and their expression (excerpts) are in a transformational relationship, if only by changing position. Words, which result from fixing and phrase reduction, are a particular type of passage, and like other passages, cannot be interpreted without recontextualisation.

Knowledge therefore stems from the decontextualisation of certain outstanding semantic forms, and corresponding expressions, either compact (lexicalisations) or diffuse (definitions). Forms give the illusion of independence, and may even seem ideal, largely because forms are by definition extremely transposable.

Yet no single word or passage can claim to sum up a text. Identifying the singularity of a text by giving a list of keywords at the beginning of an article is a means of providing instructions as to its interpretation: this key is not however the key to unlock meaning, as meaning remains to be constructed through interpretation. Metadata should therefore track the text and the context, and allow access to them, for they can never be a valid substitute for text and context.

As the logico-grammatical paradigm cannot envisage textuality, useful metadata have no determinable logico-grammatical status. In our paradigm, metadata have philological status (to document the text) and hermeneutic status (to allow its interpretation). Information cannot therefore simply be referred to as knowledge: *knowledge* will refer to pieces of information selected for their interpretative relevance. These must still be *understood*, which means that they must be related to each other, while respecting both the structure of the text from which they are taken and the aim of the task in hand.

4. Proposals

4.1 Typology of keyness

Scientific communication is no more direct and no clearer than other types of communication. Clearness of itself does not remove the need for interpretation, even if the hermeneutics of scientific and technical texts remains under-developed.

These texts are characterised by their well-known use of indexation and a specific hierarchical structure. Keyness, by concretising qualitative inequality, valorises certain aspects: a given point will be highlighted and provide access to other points, taken therefore to be secondary. That which is branded *key* indicates paths of interpretation. Although scientific discourse is supposed to be limited to facts, keyness introduces values which apply both to the facts themselves and to their means of access. Knowledge is a cultural artefact and, as such, cannot be dissociated from values.

4.1.1 *Objective keyness*

Different sections of the text may show different types of keyness which introduce a qualitative inequality index and thus provide indications of value.

a. *Peritext* – By both function and semiotic structure (case, typeface, and boldness) it establishes qualitative inequality: titles are not merely summaries, but are interpretative guidelines.
   Included in the peritext are the explicit keywords placed at the start of a text. They are also interpretative indications to highlight outstanding semantic forms.
b. *Intratext* (*Body text*) – Units are less normalized in this section. Key passages can be words, phrases, sentences, paragraphs, etc. Corpus linguistic, and particularly quantitative, methodology may provide contrastive characterisation.
   Keyness is traditionally ascribed to individual words: a probabilistic test can identify words characteristic to a passage or a text (for example, the *theme* function in Hyperbase software).
   The most important and least studied feature is the *key passage*: because of semantic diffusion, passages link networks of correlation (partial lexicalisations of the same semantic form) which may be termed *paratopies*. Minimal *zoning* techniques must be defined: the textual unit is no longer the word, but the *passage*, meaning that a keyword is only useful if it leads to a key passage.
   Intrinsic keyness is created by three types of contrast: between passages in a text; with passages from other texts in the corpus; with the chosen corpus as a whole.
c. *Infratext* – Conventionally, the intratextual content (notes, references, etc.) is considered to be low key. Still, it is the "unconscious" part of the text, and expert study can discover crucial clues, such as simple bibliographic references, which provide context for the whole text, or allow certain passages to be reinterpreted.

4.1.2 *Subjective keyness*

The lazy reader is content to accept proposed keyness: for some personal effort is required to unravel the complex relationship between peritext and intratext. However it is important to retain some critical detachment, as scientific communication is also "indirect". Beyond "overt" keyness, may be found concealed keyness: science and technology are also domains where there can be a hidden agenda.

Texts are part of social practices; their production and their interpretation both depend on differentiated tasks and strategies. Besides objective keyness, another keyness regime depends on interpretation and the task that it concretises: this may be termed "subjective" keyness.

4.1.3 *Towards dynamic keyness*

The distinction between objective and subjective keyness is only temporary. The author proposes, the reader disposes: selecting only the keywords or key passages that correspond to the task at hand, underlining those words or passages seen as key for the task at hand. Neither subjective nor objective, keyness is thus constructed dynamically in relation to

i. document structure,
ii. specificity that may be defined in contrast to the reference corpus,
iii. current practice.

4.1.4 *Consequences for textual concept redefinition*

Textual concept characterisation relies on local and global contexts.

1. Local contextual indices include: (i) collocations, adjacent named entities (author names, in particular), morphemes, punctuation; (ii) indices of expression: typography, tags.
2. Global contextual indices include: (i) the position of concepts in the text; (ii) the specificity of concepts and their immediate context, to characterise a text; (iii) the specificity of the text in the reference corpus (at expert level, a text may also be characterised by absent concepts).
3. The temporal position of concepts: the evolutivity of concepts requires diachronic studies (e.g. the work of Mathieu Valette on a corpus of writings by Gustave Guillaume over 40 years) (Valette 2006).

4.2 Towards Web Semantics

Web Semantics, developed from textual semantics and digital philology, may draw on both to adapt a variety of queries to diverse responses, which will be key if they reflect textual diversity. It is in turn merely a waystage in the development of a *comparative semiotics* of digital documents.

The sources of diversity that cannot satisfactorily be processed by the current Semantic Web paradigm must be taken into account both epistemologically and methodologically.

4.2.1 *Language diversity*

The Web is multilingual and will become ever more so. The initial hegemony of English has been overtaken by the rise of other major languages. Search engines must therefore respond to growing multilingualism, which they cannot yet do satisfactorily.

Knowledge representation should vary according to language: it is not merely a case of slicing up the same fields of reality in different ways, but even of defining fields of reality differently (for example, consider the "ontological" contrasts between Chinese and English).

4.2.2 *Discourse and genre diversity*

"Ontologies" are "populated" in relation to discourse and genre. The existence of internationally structured discourse communities has favoured the creation of plurilingual disciplinary discourse, and the diffusion of comparable genres despite language differences: this may lead to terminological calques, but also to modes of textual structuring, for both content and expression. The adoption of international norms may limit linguistic diversity but cannot eradicate it.

4.2.3 *Stylistic diversity*

The formation and evolution of concepts are the object of major differences not only from one discipline to another, but even from one author to another. Deleuze's philosophical style defines a regime of conceptual transformations that is completely different from that of Bourdieu. Corpus linguistic methods have, in this particular area, produced results that confirm the value of a comparative programme (Loiseau 2006).

4.2.4 *Qualitative inequalities within documents*

Each genre, every single text, defines a keyness regime that favours certain semiotic forms rather than others. This calls for techniques to detect qualitative inequality to be defined, again using corpus linguistic methodology.

4.2.5 *Intrinsic semiotic variety in documents*

The distinction between texts (multimedia) and documents must be reduced, or even eliminated, since the text has no content independent of its expression, and the document cannot truly be described without referring to its content. Epistemologically, the divergences between linguistics and philology must be reconsidered within a general semiotics of communication. Web semantics to some extent calls for a *comparative semiotics* of digital documents.

4.2.6 *The diversity of tasks*

Despite the nagging but distorted problem of reusability, knowledge representation that is not established for a specific application is generally not very usable and almost never reusable. Constructing such representation with the ambition of generality remains an indefinite or even infinite task, for the task determines the keyness regime.

In contrast, the junction between the *keyness horizon*, determined by the task, and the *semiotically prominent forms*, detected by contrastive analysis of the working corpus, allows the essential passages to be qualified thus restricting drastically the number of responses obtained when seeking information.

4.2.7 *Diversity of reliability status*

For both practical and ethical reasons, the question of document reliability must not be neglected, as the Web swarms with apocryphal manuscripts, of counterfeits, to say nothing of diversely revisionist texts. A non-authentic manuscript relies solely on usurped authority.

This question will be discussed within the framework of a reflection on communication types, the fields of destination and address, and finally of authority and authenticity. The number of links and page-ranking define no more than a conformist metric of authority. One truly cannot succeed in reliable information research if the degree of confidence that can be attributed to a document is not taken into account: this is one of the weak points of Web 2, when it makes anonymity a principle – as can be seen with Wikipedia.

4.3 The complexity of all data

Here a less sketchy model of data is proposed, which respects the incompressible semiotic duality between expression and content, or more generally between *vehicle* and *value*. This extends to any chain of characters, from a punctuation mark to a chapter – disregarding the apocryphal model of the sign, attributed to Saussure by the editors of the *Course in General Linguistics* and contradicted by his original manuscripts.

This vehicle/value duality, which is the semiotic substance of the data item, is controlled by a higher order duality between *point of view* and *guarantor*. The *point of view* is not merely a point of observation: it is determined by practice and by a collective or individual agent; in data-processing, it therefore depends on the application. The *guarantor* is the instance of validation that is the basis of data evaluation: this instance is a social norm that may be legal, scientific, religious, or merely endoxal. In corpus linguistics, the guarantor is the authority presiding over corpus construction; documentary metadata, such as author or editor, are part of this instance.

Point of view is "subjective" in that it is occasional; the guarantor is "objective" in that this position is constitutional or at least constituent. The duality between point of view and guarantor defines two keyness regimes: specific keyness for point of view, and general keyness for the guarantor. As data are "constructed", they are simply the initial results of a process of elaboration – processing produces further results, in a potentially recursive cycle.

In terms of the semiotics of anthropic zones (Rastier 1996, 2001, 2002), the substance (vehicle and value) of data, as objectivised, belongs to the proximal zone of the environment; the point of view belongs to the identitary zone; the guarantor belongs to the distal zone, where the instances of normativity reside. The axis here is that of symbolic mediation, while the subordinate axis linking vehicle and value is that of semiotic mediation (Rastier 2001). This is summarized in Figure 2, below.

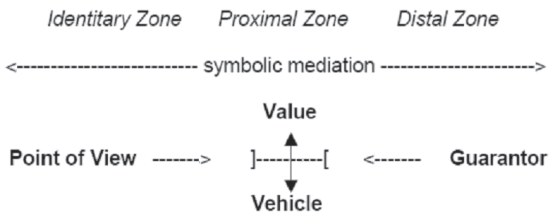


Figure 2. The four instances and the three zones of the data item

By failing to recognise the fundamental character of *value*, *point of view* and *guarantor*, by reducing data to the simple instance of *vehicle*, positivism at its most basic eludes all critical and epistemological dimensions. A database thus impoverished, a "corpus" with neither point of view nor guarantor is never a scientific object, but rather a shapeless digital mass which cannot be exploited as it stands; this is the case with "pseudo-corpora" harvested by haphazard web-crawling.

N.B.: The structure of data presented here goes beyond merely linguistic values and is apposite for other semiotic values. More generally, it is suitable for any cultural object – and a corpus is obviously a cultural object. Cultural objects

spring from the interzone connection which is part of their complexity and allows them to mediate between individuals (or groups) and their environment.

4.4 Suspicious metadata

Elementary philological indications, if retained, are today categorised as "metadata". This notion transposes, and atomises, that of metalanguage, which springs from Russellian logic. Ontologies are considered as metalanguage descriptions of documents, conditioning access to them. Metadata are to be found in the document *header*, while the data themselves form the *body* of the text, or more precisely the *intratext*. From the "Web of Documents", the transition is thus made to the "Web of Data", and then even to the "Web of Metadata": this is the conception currently in force with the Semantic Web.

There is great confusion, however, as all sorts of data are classified as metadata although they are incompatible with the impoverished text theory that generally prevails: simple bibliographic information such as author, editor, ISBN, place of publication; documentary information such as an abstract or keywords; global textual characterisations such as genre.

Linguistic theories of peritext, limiting text to intratext, separating out title, or even notes, etc., have simply heightened the confusion: for example, the title will be considered part of the metadata, although it is an integral part of the text. As a general rule, data are part of *internal linguistics* and metadata are part of *external linguistics*, and it is impossible to theorise the relationship between the two unless this duality is taken into account. Neglected problems return in concrete form as metadata; for example, in the field of multimedia, texts become the metadata of images.

Without going as far as to place an embargo on metadata, it must nevertheless be stressed that metadata are global criteria and the local-size data that depend on them: instead of separating them a priori, an elaborate theory of textuality is required to establish the correlation between metadata and data, in order to render fully the complexity of texts.

The notion of metadata must therefore be criticised and rethought: it is not a concept, but rather a class of heterogeneous problems; for example, to be more technical, would be defined as metadata the column tags for a relational database, or a class attribute in an object language, or a variable in a predicative language.

The success of Google can be explained by the introduction of a new type of metadata (links to documents) and by an implicit praxeological perspective which represents a document both from a point of view (that which selects the link) and as an guarantor (the linkmaker which will bring evaluation). This reinforces the rethinking of the notion of data which shall be pursued here.

A text is not a reservoir of knowledge that can be extracted by indexation and condensed into data summarising its information content; indexation has only a relative search and classification role. Consider the following example: in the military information service of a major European country, staff extract from Word documents certain words and expressions which are then transferred to Excel spreadsheets where they are sorted into "ontologies". These files are then transferred to analysts who synthesise them as Powerpoint presentations which are projected to the military authorities.[4] Military eloquence prizes laconism, but systematic modification of a text changes its genre and therefore its interpretation. The selection of minimal passages, mere words and expressions, is beyond control, as the degree of similarity between two indexations of the same text by the same person is on average 40%. Delinearisation and "compression" increase equivocity and create ambiguity.

The Semantic Web must inevitably be replaced by Web Semantics, because the social requirements for information seeking, the improvement of search engines, and data-mining can only be satisfied by corpus linguistics and semiotics which alone will allow the analysis of textual and documentary data.

N.B.: Grateful thanks to Évelyne Bourion, Carmela Chateau and Christian Mauceri. Translated by Carmela Chateau.

References

Berners-Lee, T. 2007. Le Web va changer de dimension. *La Recherche* 413: 34–38.

Berners-Lee, T., Hendler, J. & Lassila, O. 2001. The semantic web. A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* May 17.

Loiseau, S. 2006. Sémantique du discours philosophique: du corpus aux normes. PhD dissertation, Université Paris X.

Miller, G. & Johnson-Laird, P. 1976. *Language and Perception*. Cambridge: CUP.

Rastier, F. 1996. Représentation ou interprétation? Une perspective herméneutique sur la médiation sémiotique. In *Penser l'esprit: Des sciences de la cognition à une philosophie de l'esprit*, V. Rialle & D. Fisette (eds.), 219–239. Grenoble: Presses Universitaires de Grenoble.

Rastier, F. 2001. L'action et le sens. Pour une sémiotique des cultures. *Journal des Anthropologues* 85–86: 183–219.

Rastier, F. 2002. Anthropologie linguistique et sémiotique des cultures. In *Une introduction aux sciences de la culture*, Paris, PUF, Ch. 14, 243–267.

Rastier, F. 2004. Ontologie(s). *Revue de l'Intelligence Artificielle* (Numéro spécial Informatique et terminologies) 18: 16–39.

4. This is a true story: from Word, via Excel to Powerpoint is now the "state of the art" strategy for knowledge mining and exploitation.

Rastier, F. & Valette, M. 2009. De la polysémie à la néosémie. *Le français moderne* 77: 97–116.

Stenbit, P. (ed.). 2003. *Department of Defense Net-Centric Data Strategy*, memorandum, Washington, Defense Pentagon, 30 pp. <http://www.defenselink.mil/cio-nii/docs/Net-Centric-Data-Strategy-2003-05-092.pdf>.

Valette, M. 2006. La genèse textuelle des concepts scientifiques. Étude sémantique sur l'oeuvre du linguiste Gustave Guillaume. *Cahiers de lexicologie* 89: 125–142. Also online in Texto !: <http://www.revue-texto.net>.

SECTION II

Keyness in specialised discourse

Identifying aboutgrams in engineering texts

Martin Warren
The Hong Kong Polytechnic University, Hong Kong

This paper uses a computer-mediated methodology, concgramming, to identify the aboutness of a text. Concgrams are the raw products of the concgramming process and consist of co-occurring words irrespective of whether constituency variation (i.e. AB, A*B where * represents an intervening word) and/or positional variation (i.e. AB, BA) is present. Two engineering research articles are concgrammed to identify the most frequently occurring two-word lexical concgrams. The most frequent two-word lexical concgrams for each text are examined to determine whether the words simply co-occur or are meaningfully associated. Once this has been done, a provisional list of "aboutgrams" is drawn up which is tentatively taken to represent the aboutness of each text. These lists are then referred to a specialised corpus of engineering texts, and then a general reference corpus. Those aboutgrams on the lists which are consistently more frequent in the text than in the two corpora are then put forward as representing the aboutness of the text. In the study, the lists of aboutgrams are compared with single word frequency lists to evaluate the advantages to be gained from determining aboutness by means of phraseology rather than key words. The conclusion is that aboutgrams are a better means for uncovering the aboutness of the texts.

1. Introduction

Highlighted in the 1960s (Sinclair, Jones & Daily 1970) as fundamental to the English language, corpus linguistics studies of phraseology have increased in number since the late 1980s. Studies in, for example, lexical semantics (see, for example, Sinclair 1987, 1996, 2004a; Stubbs 2001; Tognini-Bonelli 2001), pattern grammar (see, for example, Hunston & Francis 2000), textual patterns (see, for example, Scott & Tribble 2006; O'Keeffe, McCarthy & Carter 2007) and semantic prosody (see, for example, Louw 1993; Sinclair 1991) have all contributed to a better understanding of the phraseological character of natural language.

A number of studies have focussed on contiguous word co-occurrences known as n-grams, which are also termed "bundles" (see, for example, Biber et al. 1999; Biber, Conrad & Cortes 2004), "chunks" (see, for example, O'Keeffe, McCarthy & Carter 2007) or "clusters" (see, for example, Carter & McCarthy 2006; Scott & Tribble 2006). These studies have all served to support the claim by Sinclair (1987, 1991) that it is lexis and not syntax that has the main role in organising language and meaning-making based on what he terms the "idiom principle" (1987), which is at the centre of the phraseological tendency in language. The idiom principle refers to the way in which speakers and writers co-select words to create meaning.

Phraseology is not only of interest to researchers, the pedagogical treatment of phraseology also presents new challenges for learners and teachers of the English language because currently it is rarely foregrounded in the learning and teaching of the English language and applied English language studies. Exceptions to this observation include a small number of recent textbooks that cover aspects of phraseology, collocation and extended collocational associations (see for example, Carter & McCarthy 2006; Sinclair 2003; Stubbs 2005; Scott & Tribble 2006; O'Keeffe et al. 2007). This study, therefore, builds on the work of researchers and language educators such as these by taking the notion of phraseology further by emphasising the importance of including phraseological variation in studies of phraseology. It is argued that the learning and teaching of phraseology, especially phraseological variation, merits greater emphasis. Also, this study applies a computer-mediated research methodology, "concgramming" (Cheng, Greaves & Warren 2006; Greaves & Warren 2007), aimed at achieving this objective. Scott & Tribble (2006: 55–88) demonstrate how individual key words can help to reveal the aboutness of a specific text, or genre, by comparing key word frequencies with those of different reference corpora. In this paper, it is shown that a similar procedure can be used based on the phraseology most frequently found in a text or corpus. Indeed, both Scott and Tribble (2006: 131–159) and O'Keeffe et al. (2007: 64–79) illustrate this, although in a more limited way than proposed in this paper, by comparing the relative frequencies of n-grams across genres and varieties of English. Scott and Tribble (2006: 131) state that clusters offer an alternative to key words if one wishes to differentiate texts in different corpora because clusters provide insights into the aspects of phraseology used in specific contexts (ibid: 132). In their study of academic writing, Scott and Tribble focus on the most frequent clusters, such as *one of the*, *the end of the*, *as well as*, *part of the*, and *out of the*, which they argue are important if we are to better understand how texts are formed (ibid: 132). In the present study, on the other hand, the focus is on the determination of aboutness and so the associations of lexically-rich words are investigated as these are presumed to be the primary source of the aboutness of a text or corpus. Before describing the procedure adopted in this study and the preliminary findings, concgramming and its products are described.

2. Concgrams

Uncovering the full extent of word associations in a text or corpus has posed problems in the past. Researchers in natural language processing, computational linguistics and corpus linguistics have concentrated on n-grams. Searches for n-grams find contiguous word associations, such as *take part*, but miss instances of the same phraseological pattern when they are realised in instances which contain constituency variation such as *take no part* or *take an active part*. Searches for n-grams, therefore, are good at finding instances of co-occurring words that are simply contiguous, but many other instances may be overlooked and those that are always, frequently, or sometimes, in non-contiguous sequences (i.e. AB, A*B, where '*' represents an intervening word or words) go undiscovered. In addition, other existing searches for non-contiguous word associations typically require the input of a formula which can be user-unfriendly or they require the user to input search items, which limits the likelihood of finding new phraseologies.

The limitations of n-gram searches have resulted in the development of searches for gapped n-grams or "skipgrams" in Natural Language Processing (see Wilks 2005). Skipgrams allow for a certain amount of constituency variation (i.e. AB and A*B), but they are limited to two- or three-word skipgrams and four "skips" (Wilks 2005), and so miss associated words more than four words apart. A further limitation of these searches is that they do not find instances of positional variation (i.e. AB, BA).

Another example of an automated search engine for non-contiguous co-occurring words is Fletcher's KfNgram program (2006) which is designed to find "phrase-frames". Phrase-frames are "sets of variants of an n-gram identical except for one word" (ibid, 2006) and are derived from an initial automated search for n-grams of up to 8 running words. Phrase-frames, therefore, can be described as a form of skipgram, but are constrained by a narrower search parameter (i.e. one intervening word rather than four in the case of skipgrams), with the result that non-contiguous co-occurring words, comprised of the same words, remain undiscovered if they differ by more than one word. Also, as is the case with skipgram searches, co-occurring words with positional variation are not found.

Cheng, Greaves & Warren (2006) describe the contribution that a search engine, ConcGram©[1] (Greaves 2009), can make to identifying units of meaning in a text or corpus. It extracts recurrent concgrams (i.e. sets of between two and five co-occurring words) fully automatically, within a wide span of up to 12 words

1. ConcGram© is a search engine written and developed by Chris Greaves, Senior Project Fellow, English Department, The Hong Kong Polytechnic University.

on either side of the origin[2]. Concgrams include all of the configurations of the co-occurring words irrespective of any constituency and/or positional variation. Cheng, Greaves & Warren (2006) argue that the identification of concgrams further facilitates an appreciation and understanding of Sinclair's idiom principle. The rationale for this claim is that concgrams are a useful source of raw data which, when analysed, can reveal the co-selections made by the speakers and writers represented in a text or corpus. Concgrams are thus a starting point for quantifying the extent of phraseology in a text or corpus and hence determining the phraseological profile of the language contained within it.

In this paper a distinction is made between concgrams, or specific concgram configurations, which are comprised of words which simply co-occur and those which comprise meaningfully associated words. This point is best illustrated by the example of a concgram's concordance lines described below.

The concordance lines are of the two-word concgram, *design/structural*, with the span set at 4 (i.e. four words either side of the origin) based on a search of an engineering article (Chang & Zhang 2003). A sample of ten of the concordance lines are provided in Figure 1 to illustrate what a concgram is and to explain briefly the process for determining whether the words in the concgram are co-occurring or associated.

1 (or computer) modelling of a building, framing **design**, **structural** analysis, component **design**, design
2 rocess framing plan (d) represents the final **design** of **structural** framing. One may further carry out
3 and cost-effectiveness. The **structural design** of a tall building involves several rather
4 instance, considering the preliminary **structural design** of a building, after the structure model is
5 engineers to achieve not only a safe **structural design**, but also a cost-effective design in terns of
6 has actually been applied to the **structural design** of more than 25 building projects with the
7 this model, one may perform **structural** framing **design** with assigned initial dimensions for all the
8 objectives of the preliminary **structural** framing **design** are: 1) To find which partitions are efficie
9 reduced occupied **structural** space, and shorter **design** time, have been realised. Acknowledgements T
10 **structural** analysis, optimisation, automated **design** check, and cost analysis, one may easily

Figure 1. Sample concordances for the two-word concgram *design/structural*

The primary purpose of concgram searches is to focus on the patterns of co-selection. There is therefore an equal emphasis placed on all of the co-occurring words instead of primarily focussing on the node, which is the tendency in a traditional KWIC (i.e. key word in context) display. This important feature of a concgram concordance can be seen in Figure 1. When a concgram concordance is displayed, all of the words in the concgram are highlighted in bold font with each word in

2. The rationale for using the term *origin* instead of *node* is discussed later in this paper.

the concgram assigned a different colour. This has the effect of shifting the focus away from only the centred word to all of the co-occurring words. Hence the use of the term "origin" rather than "node" for the word, or words, which forms the basis of the automated concgram search in order to emphasise the difference between a concgram search and traditional node-based searches.

The sequencing of the concordance lines in the concgram display is designed to facilitate the identification of patterns of constituency and positional variation in a concgram. In Figure 1, the lines begin with co-occurrences of *structural* positioned to the right of the origin, *design*, and they also begin with any instances with no intervening word, and then one intervening word, two words, and so on. Once all of the constituency variation to the right of the origin has been listed, the same procedure is repeated to order the co-occurring word(s) to the left of the origin and, in Figure 1, these begin in line 3. Once a concgram concordance is displayed, the user has to determine whether the words in the concgram are simply co-occurring or whether they are meaningfully associated. This point is illustrated by the above concordance lines in which lines 1, 9 and 10 are examples of *design* and *structural* not being meaningfully associated with each other, unlike the instances in lines 2–8 which are meaningfully associated in a "meaning shift unit"[3] (Sinclair 2007). The canonical form of the MSU, based on frequency of occurrence, is *structural design* with the variations to this form, in lines 2 (***design** of **structural** framing*), 7 and 8 (***structural** framing **design***), exhibiting slight turbulence relative to it. The need to distinguish between co-occurrence and association is important because it impacts the frequency count ascribed to a particular combination of words because instances deemed to be simply co-occurrences are excluded.

The fully-automated capability of the search engine means that concgramming is a truly "corpus-driven" methodology (Tognini-Bonelli 2001: 11), and this is the defining feature of the search engine. The result is that concgram searches provide the raw data of all of the co-occurring words to then enable the user to arrive at a more extensive description of associated words and their meanings, and also, and more importantly, to uncover new phraseological patterns of language use.

3. From concgrams to aboutness and aboutgrams

Phillips (1989) suggests a means to determine the aboutness, or topic, of a text using an objective, quantitative, distributional methodology. According to Phillips, aboutness is a product of the global patternings of a text, which he terms the

3. Sinclair (2007) expresses a preference for the term *meaning shift unit* rather than *lexical item* (Sinclair 1996, 1998).

"macrostructure" of the text, and these should be arrived by computational means, so that they are derived from the text itself rather than from external features. The notion of a text's profile is linked to what Phillips refers to as the "aboutness" of a text. The term "phraseological profile" refers to all of the word associations in a text or corpus, and "aboutness" can be ascertained from the word associations that are specific to that particular text or corpus. In this paper those word associations which are specific to a text or corpus are termed "aboutgrams" (Sinclair, personal communication; Sinclair and Tognini-Bonelli, in press).

An important assumption in Philips' position is that meanings in language are primarily constructed by lexical words, or the associations of lexical words. For this reason, in this paper, the concgrams that are analysed are all two-word lexical concgrams. In other words, the words in the concgrams are both lexical. It has to be pointed out that concgrams are, of course, not confined to the co-occurrences of lexical words and the most frequently occurring concgrams in a text or corpus are combinations of grammatical words, termed "collocational frameworks" (Renouf & Sinclair 1991). Studies are needed to examine whether these frequently occurring collocational framework type concgrams (i.e. concgrams comprising grammatical words which frame lexical collocates), and the less frequently occurring "organisational frameworks" (i.e. concgrams comprised of words which are organisationally-oriented rather than propositionally-oriented, such as *because ..... so*), are also text or genre specific. There is evidence to suggest that n-grams, including those made up entirely of grammatical words, can be genre-sensitive (Carter & McCarthy 2006: 828–837; Scott & Tribble 2006: 131–159; O'Keeffe, McCarthy & Carter 2007: 68) and there is every reason to suppose that this is also the case for concgrams.

The methodology for arriving at a list of aboutgrams, which in turn represent the aboutness of a text, outlined here is based on a procedure outlined by Sinclair (2006). While the phraseological profile of a text is arrived at by identifying all of the word associations in the text, the aboutness of a text is determined by a process of distillation in which the most frequently occurring lexical phrases in an engineering text, for example, are placed on a provisional aboutgram list and are referred to a specialised corpus of engineering texts. Those which are found to occur equally frequently in both the text and the specialised corpus, or more frequently in the specialised corpus, are removed from the provisional aboutgram list. The same process is then repeated for the list of aboutgrams in the text using a general reference corpus and, again, based on the same criteria, the list is further refined. The end result is a list of aboutgrams which collectively represent the aboutness of the text.

Before we look at the findings of this preliminary study, the data used are briefly described.

4. Data

The data used in this study are two engineering research articles, a specialised corpus of engineering English and a general reference corpus. The two engineering research articles are referred to as Article A (Chang & Zhang 2003) and Article B (Xu et al. 2003) and are both taken from an engineering journal, Transactions of the Hong Kong Institution of Engineers. Article A contains 5,228 words, and Article B 4,810 words. These articles are part of the Hong Kong Engineering Corpus (HKEC) which, at the time of the study, contains 750,000 words of English Engineering texts collected in Hong Kong. At the time of writing, the HKEC is mainly comprised of research papers and symposium papers written by engineering professionals and academics, most of whom are members of the Hong Kong Institution of Engineers. As with the majority of professional associations in Hong Kong, the Hong Kong Institution of Engineers uses the medium of English in all of its research papers and conference presentations. The British National Corpus (BNC), comprising 100 million words of spoken and written texts, is used as the general reference corpus.

5. Analysis of data

Using the search engine, ConcGram©, single lexical word frequency lists for each of the two articles are generated in order to compare the most frequent lexical words in the two articles (see Tables 1 and 3) with the most frequently occurring aboutgrams in the articles (see Tables 2 and 4). The reason for comparing word frequency lists with aboutgrams frequency lists is to attempt to illustrate how the most frequent lexical word associations in a text present a fuller picture of its aboutness than the most frequent single lexical words. The lists of two-word aboutgrams are generated (Tables 3 and 4) by first identifying the most frequent two-word lexical associations in each article. These rank ordered lists are tentatively put forward as the set of aboutgrams which represent the aboutness of the article from which they originate. The iterative process of then determining whether or not each list captures the "aboutness" of each text requires that a search for each one is conducted in the specialised corpus, HKEC, and then again in the general reference corpus, BNC.

Those that are not at least matched in terms of relative frequencies in the HKEC or the BNC are dropped, and those which remain can be said to be aboutgrams because they represent the language of the article based on its distinctive aboutness.

Table 1.[4] Most frequent lexical words in Article A

Lexical word	Frequency
design(s)	133
structural	116
model(s)	64
building(s)	63
architectural	34
optimisation	32
CAD	29
member(s)	29
analysis	28
layer	22

Table 2.[5] Most frequent two-word aboutgrams in Article A

Aboutgram	Article A	HKEC	BNC
design(s)/structural	37	71	13
structural model(s)	34	38	15
building(s)/design	26	70	115
architectural/model(s)	14	14	2
structural analysis	12	16	53
design/tall	11	12	2
data capture	11	16	28
structural optimisation	10	14	0
analysis/design	9	20	101
form/structural	8	9	22

Before we look at which of the above tables best captures the aboutness of Article A, it is interesting to compare the contents of the Tables 1 and 2 to see to what extent the words in Table 1 are also to be found in Table 2. In all, seven of the ten words listed in Table 1 are to be found in the aboutgrams in Table 2, the three exceptions are the single lexical words *CAD*, *member(s)*, and *layer*. Conversely, the aboutgrams in Table 2 contain four words not in Table 1, namely, *data*, *capture*, *tall*, and *form*. This observation in itself suggests that the two lists highlight aboutness differently.

4. In the tables, singular and plural forms are combined when calculating frequencies.

5. In Tables 2 and 4, contiguous aboutgrams are written as they appear in the texts and aboutgrams which exhibit constituency and/or positional variation are written in alphabetical order separated by a forward slash.

Earlier, the importance of capturing instances of phraseological variation was emphasised and, when Table 2 is examined from this perspective, it is seen that, of the ten aboutgrams, six have constituency and/or positional variation. These are *design(s)/structural*, *building(s)/design*, *analysis/design*, *architectural/model(s)*, *design/tall*, and *form/structural*. The remaining four aboutgrams are contiguous, *structural model(s)*, *structural analysis*, *data capture*, and *structural optimisation*. This observation underlines the importance of including instances of phraseological variation when compiling the most frequent aboutgrams in a text.

In terms of whether Table 1 or Table 2 best represents the aboutness of article A, we can examine the top six lexical words, *design(s)*, *structural*, *model(s)*, *building(s)*, *architectural*, and *optimisation*, all of which are also found in the aboutgrams in Table 2. When we are able to see what words each of these single words associate with in the two-word aboutgrams, we can better appreciate what Article A is about. The ways in which these six single words meaningfully combine with other lexical words bring us closer to the aboutness of the article. For example, from Table 1 it is not possible to know what kind of *design(s)*, *building(s)*, or *model(s)* the text is about, what it is that *structural* and *architectural* modify, or what form *optimisation* takes in the article. However, the aboutness of the article is much clearer when we examine the two-word aboutgrams in Table 2 because these uncertainties are removed. Now we can see that *structural* combines with *design(s)*, *model(s)*, *analysis* and *optimisation*, and we can see *design(s)* combines with *structural*, *building(s)*, *tall* and *analysis*. Similarly, *model(s)* combines with *structural* and *architectural*. Thus the phraseology exhibited in the aboutgrams in Table 2 provides us with a much less ambiguous representation of the aboutness of Article A.

In this study, we limit our investigation to two-word aboutgrams, but a search for three-word aboutgrams would result in some of the two-word aboutgrams being seen as components of larger aboutgrams, and this leads on to another finding. Some of the aboutgrams in Table 2 provide evidence of the process by which the concgramming software can help to build a profile of the intercollocation of collocates (Sinclair 2005). After establishing the unique words (i.e. "types") in the text, the search engine searches for words associated with each unique word and lists all of the two-word concgrams found. The two-word concgrams then become the origin for the next search which finds all the words associated with the two-word concgrams. This iterative process, by which concgrams are built up automatically, is the same process as that described by Sinclair (2005) for determining the aboutness of documents. Sinclair (2005) states that the "process (called the *intercollocation of collocates*) also disambiguates words, and a group gives a strong sense of the content, scope and argument of the document". For example, two of the most frequent aboutgrams in Article A are *building(s)/design* (26 instances),

and *design/tall* (11 instances) and a frequent three-word aboutgram in Article A is *design [of a] [for] tall building(s)* (11 instances). The latter is the product of the intercollocation of the collocates of two of the most frequent two-word aboutgrams in the Article A and provides evidence for Sinclair's claim.

Let us now examine the findings for Article B listed in Tables 3 and 4 below.

Table 3. Most frequent lexical words in Article B

Lexical word	Frequency
control	119
building(s)	117
damper(s)	108
MR[6]	79
storey	66
semi-active	41
logic	36
system	34
response(s)	34
current	26

Table 4. Most frequent two-word aboutgrams in Article B

Aboutgram	Instances in article	Instances in HKEC	Instances in BNC
MR damper(s)	72	72	0
building(s)/storey	64	84	157
control/damper(s)	53	53	0
logic control	40	46	0
semi-active/control	31	32	50
control/MR	29	33	1
control algorithm	24	30	0
control/passive	18	26	0
building/response(s)	17	34	0
semi-active logic	17	18	3

Again, Tables 3 and 4 contain many of the same words, but Table 3 contains two words not found in Table 4: *system* and *current*. Table 4 has two words, *algorithm* and *passive*, which are not among the most frequent single lexical words. It is interesting to note that the aboutgrams in Table 4 have a similar preponderance of phraseological variation with six aboutgrams exhibiting constituency and/or positional variation: *building(s)/storey*, *control/damper(s)*, *semi-active/control*,

6. MR is used throughout Article B and is the abbreviation of magnetorheological.

control/MR, *control/passive*, and *building/passive*. The remaining four aboutgrams, *MR damper(s)*, *logic control*, *control algorithm* and *semi-active logic*, are contiguous.

A comparison of the six most frequent single lexical words in Table 3, which also feature in Table 4, with the aboutgrams that these words are members of illustrates the extent to which the aboutness of Article B can be revealed by only looking at Table 3. The words *control*, *building(s)*, *damper(s)*, *MR*, *storey* and *semi-active* leave much unanswered, but, when we look at how they frequently combine meaningfully with other words in Article B, we have a much fuller picture of the article's aboutness. For example, *control* combines with six of the top ten aboutgrams in Table 4, *control/damper(s)*, *logic control*, *semi-active/control*, *control/MR*, *control algorithm* and *control/passive*, a fact which underlines how these aboutgrams clarify the aboutness of the article. Similarly, other aboutgrams show how other frequently occurring single words meaningfully combine in this article. These better illustrate the aboutness of Article B rather than relying on single words, such as *MR damper(s)*, *building(s)/storey* and *building(s)/responses*. There are also aboutgrams such as *algorithm control* and *control/passive* which contain words not in Table 3 and which also take us closer to the aboutness of the article.

As was discussed earlier, the process of identifying the intercollocation of collocates helps to disambiguate words, and, once identified, such a group provides a good indication of a text's aboutness. In Article B we have four aboutgrams, *logic control* (40 instances), *semi-active/control* (31 instances), *control algorithm* (24 instances), and *semi-active logic* (17 instances), which on occasion combine to become the four-word aboutgram *semi-active logic control algorithm* (13 instances). This four-word aboutgram is the product of the intercollocation of the collocates of four of the most frequent two-word aboutgrams in Article B.

In the above analyses it is shown that a search for "key phrases", that is aboutgrams, rather than "key words" brings us closer to an unambiguous understanding of the aboutness of the two articles. Once the criteria for aboutness have been established, in future studies the most frequent aboutgrams in both of these articles could be compared with other engineering articles, and their "aboutness distance" (Sinclair 2005) could then be calculated. Future studies of aboutgrams might also help to determine the genre to which a discourse belongs. Just as there are text-specific aboutgrams, we might expect to find genre-specific and register-specific aboutgrams using a methodology similar to the one outlined here. Another line of research would be to see whether the nature of aboutgrams varies from one register to another. For example, it would be interesting to investigate whether there are more compound nouns in a technical register and more complex prepositional phrases in fiction.

6. Conclusions

Sinclair (2004b: 148) observes that "the word is not the best starting-point for a description of meaning, because meaning arises from words in particular combinations". He further elaborates this important point (Sinclair 2005: 3) by stating that "a word on its own is usually not distinctive enough to deliver a stable and precise meaning (outside the protected words which are recognised as technical terms – and even they are always at risk)". Hence this study argues that an over-reliance on only key words means that the most important information regarding the aboutness of an individual text is not utilised, namely its phraseology. This study has illustrated how an examination of a text's phraseology, and, importantly, this includes phraseological variation, makes it possible to identify the tentative aboutgrams in a text. These tentative aboutgrams are confirmed with reference to both a specialised corpus containing texts from the same field as the text under investigation, and a general reference corpus. Furthermore, this methodology could help to uncover the aboutgrams not only in specific texts, but also those of particular genres and specialised corpora, such as the HKEC.

Since currently the initial lists of concgrams cannot be automatically classified into those which contain meaningful word associations and those which do not, human intervention is required to examine the concgrams in their concordance lines. This means that the raw frequencies may need to be adjusted in cases of repeats in the same concordance line, and that it is necessary to remove instances where the words in a concgram are not associated or do not conform to the canonical form of the concgram, with canonical form being defined here as the most frequent form with the prototypical meaning.

The methodology described in this study can be used in learning and teaching contexts with students of English to raise their awareness of the centrality of phraseology and its possible variations. Its adoption in such contexts would also help to develop in students the critical skills needed for identifying the phraseology which best captures the aboutness of a text, genre or specialised corpus.

Acknowledgements

The work described in this paper is a product of collaborative research examining the aboutness of texts and specialised corpora involving Elena Tognini-Bonelli, University of Siena, Italy, John McH. Sinclair, the Tuscan Word Centre, Italy, and my colleagues here in Hong Kong, Winnie Cheng and Chris Greaves. Sadly, John's death in 2007 ended his invaluable input, but his contribution to our discussions on aboutness, and how best to determine it, shaped much of what is described and discussed in this paper.

My thanks to the editors and anonymous reviewers who gave me a lot of useful feedback and suggestions. Thanks are also due to the Hong Kong Institution of Engineers for generously making available a large number of engineering texts which are now housed in the Hong Kong Engineering Corpus. The research study described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. G-YF39).

References

Biber, D., Conrad, S., Finegan, E., Johansson, S. & Leech, G. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D., Conrad, S. & Cortes V. 2004. *If you look at …*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371–405.

Carter R. & McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: CUP.

Chang T. Y. & Zhang, N. 2003. An innovative approach to the structural design of tall buildings. *Transactions of the Hong Kong Institution of Engineers* 10(4): 14–21.

Cheng, W., Greaves, C. & Warren, M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4): 411–433.

Fletcher, W. H. 2006. "Phrases in English" Home. <http://pie.usna.edu/> (15 February 2006).

Greaves, C. 2009. *ConcGram* 1.0: *A Phraseological Search Engine.* Amsterdam: John Benjamins.

Greaves, C. & Warren, M. 2007. Concgramming: A Computer-driven Approach to Learning the Phraseology of English. *ReCALL Journal* 17(3): 287–306.

Hunston, S. & Francis, G. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [Studies in Corpus Linguistics 4] Amsterdam: John Benjamins.

Louw, W. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology: In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 157–176. Amsterdam: John Benjamins.

O'Keeffe, A., McCarthy, M. & Carter, R. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: CUP.

Phillips, M. 1989. *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham.

Renouf, A. & Sinclair, J. McH. 1991. Collocational frameworks in English. Reprinted in J. McH. Sinclair *Lexis and Lexicography*, 55–71. Singapore: National University of Singapore, Unipress.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Sinclair, J. McH. 1987. Collocation: A progress report. In *Language Topics: An International Collection of Papers by Colleagues, Students and Admirers of Professor Michael Halliday to Honour him on his Retirement*, Vol. 3, R. Steele & T. Threadgold (eds), 319–333. Amsterdam: John Benjamins.

Sinclair, J. McH. 1991. *Corpus Concordance Collocation*. Oxford: OUP.

Sinclair, J. McH. 1996. The search for units of meaning. *Textus* 9(1): 75–106.

Sinclair, J. McH. 1998. The lexical item. In *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 171], E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.

Sinclair, J. McH. 2003. *Reading Concordances*. London: Longman.

Sinclair, J. McH. 2004a. *English Collocation Studies*. London: Continuum.
Sinclair, J. McH. 2004b. *Trust the Text*. London: Routledge.
Sinclair, J. McH. 2005. Document relativity. Ms, Tuscan Word Centre, Italy.
Sinclair, J. McH. 2006. Aboutness 2. Ms, Tuscan Word Centre, Italy.
Sinclair, J. McH. 2007. Collocation reviewed. Ms, Tuscan Word Centre, Italy.
Sinclair, J. McH., Jones, S. & Daley, R. 1970. *English Lexical Studies*. Report to the Office of Scientific and Technical Information.
Sinclair, J. McH. & Tognini-Bonelli, E. In press. *Essential Corpus Linguistics*. London: Routledge.
Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Stubbs, M. 2005. The most natural thing in the world: Quantitative data on multi-word sequences in English. Paper presented at Phraseology 2005, Louvain-la-Neuve, Belgium, 13–15 October 2005.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.
Wilks, Y. 2005. 'REVEAL: The notion of anomalous texts in a very large corpus.' Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 30 June–3 July 2005.
Xu, Y., Ng, C. L., Chen, J. & Qu, W. 2003. Innovative technology for seismic response reduction of tall buildings with podium structures. *Transactions of the Hong Kong Institution of Engineers* 10(4): 88–94.

Keywords and phrases in political speeches

Denise Milizia
University of Bari, Italy

This paper analyses n-grams and concgrams in the speeches of Tony Blair and George W. Bush delivered from the beginning of 2005 till June 2007. The focus of the study is first on the single word *climate*, which is found to figure high in the now former Prime Minister's keywords, and then on the cluster *climate change*, in that *change* was also picked up as key and was always found in company with *climate*. This work is an attempt to corroborate the assumption that phraseological combinations, in the form of n-grams and concgrams, i,e. in contiguous and non-contiguous sequences, are usually much better at revealing the "aboutness" of the text than individual words.

1. Introduction

Many linguists agree today that the "idiom principle" (Sinclair 1991: 109) is the norm in the organization and interpretation of language use while the "open choice" principle is the exception, and that our lexicon does not consist mainly of single words, representing atomistic concepts with syntactic rules licensing them (Cf. Erman 2007: 26).

Indeed, learning (and teaching) a mere list of individual words hardly helps: our knowledge of a language is not only a knowledge of single words but of their predictable combinations, and words should be studied not in isolation but in collocations (Stubbs 2001: 45). We have been brought up to be aware of idioms but not of collocations (Teubert cited in Sinclair, Jones, Daley & Krishnamurthy 2004): words attract each other also beyond idioms. Unlike idioms, collocations are often transparent but, like idioms, they are usually non-compositional, since they have taken on quite specific meanings in discourse and interaction (Erman 2007: 33).

Folk linguistics has taught us that idioms are only those groups of words which are intuitively felt to be idiomatic; these abound in the spoken political corpus analysed in this study: e.g. *cut and run, see eye to eye, connect the dots, stand shoulder to shoulder*, *from all walks of life*, to mention just a few. These idioms are

certainly perceived as such and it is apparent that "it is not possible to guess the meaning of the phrase from the individual meaning of the words that make it up (Sinclair 2003: 176).

In the literature, words in habitual collocation have been labelled in different ways: phrase (Sinclair 1987), idiom (ibid.), lexical bundle (Biber, Johansson, Leech, Conrad & Finegan 1999), lexical item (Sinclair 1998), extended unit of meaning (Sinclair 1996), cluster (Scott 2008), multi-word unit (Sinclair 1996), concgram (Cheng, Greaves & Warren 2006), skipgram (Wilks 2005), phrase-frame (Fletcher 2002–2007). All these labels – while referring to different types of lexical relations – comprise two or more words which are not necessarily felt to be idiomatic and which co-occur more often than chance would predict for no obvious reason other than habit. It is assumed that lexical attraction is a matter of convention (cf. Renouf & Banerjee 2007a: 419) and personal choice, so we find ourselves saying *the issue of climate change*, *the climate change issue*, *the problem of climate change* but not *the climate change problem*.[1] Collocations differ from language to language, and it was interesting to notice that in the corpus assembled in Italy in the same period of time the phrase *cambiamenti climatici* (*climate changes*: Italians seem to prefer the plural form) is never found in company with *problema* (although it certainly is a problem).

Thus, the focus of this paper is on words and phrases, of both contiguous and discontinuous word relations. For this reason, two pieces of software have been used to interrogate the corpus: *WordSmith Tools* 5.0 (Scott 2008) and *ConcGram* 1.0 (Greaves 2009). The latter is able to handle both positional (AB, BA) and constituency (AB, ACB) variation, hence patterns like *changes in the climate, the climate is changing, the climate carries on changing* were identified, showing that the relationships between words go beyond their immediate neighbourhood (Scott & Tribble 2006: vii).

The reason why *climate+change* was chosen as the focus of this study is because, after avoiding words which offer few surprises such as proper nouns and style markers, both *climate* (ranking 22nd) and *change* (ranking 40th) emerged as outstanding, therefore key, in the comparison between Tony Blair's speeches delivered from 2005 to 2007 and George W. Bush's speeches delivered over the same period. A closer look at both words has also shown that in the corpus *climate* seems to enjoy almost exclusively the company of *change* (with the only exception of ten occurrences of *climate chaos*): indeed, *climate* seems to collocate

1. Although the analysis of this paper is restricted to the years 2005–2006 and the first six months of 2007, we have looked at the whole corpus of Tony Blair's speeches (1997–2007) and of George W. Bush's speeches (2001–2007), and *the climate change problem* never occurs. It is important to underline, though, that these words may attract each other in other people's idiolects; thus, these findings are not a statement about language as a whole but about the corpus which I studied.

with *change* as strongly as *night* and *milk* collocate respectively with *dark* and *cow* (Firth, cited in Sinclair et al. 2004).

The present study investigates the keywords obtained from the comparison and the phraseology that these keywords create, showing how phraseological combinations are usually much better at revealing the "aboutness" of the text than individual words (Warren & Greaves 2007: 297–299).

The corpus and the method used for this analysis and the ideal reference corpus size are presented in Section 2. Section 3 considers the keywords emerged by referencing Tony Blair's speeches against George Bush's speeches, and Section 4 analyzes the main findings, both of the single word emerging as key and then of the n-grams and concgrams that it creates. Section 5 summarizes the results.

2. Data and method

In this paper keywords are regarded as the most outstanding words in a corpus, not simply the ones with highest frequency of occurrence (otherwise a wordlist would suffice). Keywords may be identified in a variety of ways (Scott 2009): some methods are based on frequency of occurrence in comparison with a suitable reference, others on human identification. With regard to problems in human detection of keyness, Phillips (1989) talks of the text's "macrostructure". He argues that one should determine the macrostructure of texts by computational means, to ensure that the results are derived from the text itself and not from external features.

The procedure used for identifying keywords in the present research is the one devised for use in *WordSmith Tools* (Scott 2008), and is based on simple verbatim repetition: a wordlist based on the foreground corpus is compared with a wordlist based on a background corpus. The former is automatically assumed to be the smaller of the two corpora chosen, and the larger provides background data for reference comparison.

The corpus data used in this investigation comprise one million words from speeches by Tony Blair from the beginning of 2005 until the end of his term (June 2007), and five million words from speeches by George W. Bush over the same period. The data, collected from their official websites, <www.number10.gov.uk> and <www.whitehouse.gov>, include not only speeches in the proper sense, but also statements, interviews, press conferences and press briefings, debates, and questions to the Prime Minister included into the House of Commons Debates.[2]

---

2. We will not report here details on the two corpora, partly for reasons of space and partly because detailed evidence has been presented in previous works (Milizia 2006; Milizia & Spinzi 2008; Milizia 2009).

The two wordlists generated in both politicians' corpora were analysed and, as Table 1 shows, a top lexical word in both lists was *people*, ranking 25th in Blair and 23rd in Bush with exactly the same percentage (0.56%).

Table 1. First 30 words in Blair's and Bush's word lists

N	Word	Freq.	%	Word	Freq.	%
1	THE	53.347	5,91	THE	277.164	5,53
2	TO	29.927	3,31	TO	189.102	3,78
3	AND	29.704	3,29	AND	171.647	3,43
4	OF	25.388	2,81	OF	131.237	2,62
5	THAT	22.652	2,51	A	107.014	2,14
6	IS	21.098	2,34	IN	93.623	1,87
7	IN	19.783	2,19	THAT	77.876	1,56
8	A	16.517	1,83	I	73.267	1,46
9	WE	14.627	1,62	IS	62.294	1,24
10	I	14.345	1,59	WE	60.066	1,20
11	IT	12.304	1,36	YOU	56.522	1,13
12	YOU	10.192	1,13	FOR	54.686	1,09
13	HAVE	9.767	1,08	#	42.906	0,86
14	ARE	9.121	1,01	OUR	40.054	0,80
15	FOR	8.348	0,92	IT	35.760	0,71
16	THIS	7.537	0,83	THIS	32.873	0,66
17	#	6.935	0,77	ON	31.834	0,64
18	ON	6.857	0,76	HAVE	31.459	0,63
19	BUT	6.407	0,71	ARE	30.619	0,61
20	BE	6.310	0,70	APPLAUSE	29.795	0,59
21	WITH	5.612	0,62	WITH	29.206	0,58
22	AS	5.568	0,62	PRESIDENT	28.753	0,57
23	WILL	5.352	0,59	PEOPLE	28.226	0,56
24	THERE	5.314	0,59	BE	28.045	0,56
25	PEOPLE	5.020	0,56	WILL	24.721	0,49
26	WHAT	4.771	0,53	AS	22.384	0,45
27	THINK	4.722	0,52	THEY	21.135	0,42
28	THEY	4.569	0,51	NOT	19.598	0,39
29	NOT	4.549	0,50	SO	19.390	0,39
30	DO	4.227	0,47	ABOUT	19.277	0,38

Despite the fact that *people* is heavily used by both politicians, the clusters within which this word is embedded vary across the two cultures: looking at the first 50 three-word clusters in Table 2, the *Clusters* function in *WordSmith Tools* showed that in the British[3] list *the Iraqi people*, ranking third in both lists, is more frequently

3. In this paper, for the sake of convenience, Tony Blair's speeches are often referred to as the British corpus, and George W. Bush's as the American corpus. The two corpora analysed in this work, however, contain only the speeches delivered by these two politicians.

used than *the British people*, ranking 7th, whereas *the American people* ranks first in the American corpus. Other British three-grams are *people in Europe*, *people in Iraq*, *the Afghan people*, *the Palestinian people*, *innocent people in*. None of these occurs in Bush's list, and with the exception of *people of Iraq* (ranking 40th), most of the three-word clusters contain the adjective *American*.

Table 2. First 30 three-word clusters around *people* in Blair and Bush

N	Cluster	Freq.	N	Cluster	Freq.
1	PEOPLE WHO ARE	130	1	THE AMERICAN PEOPLE	2.195
2	THE PEOPLE OF	121	2	THE PEOPLE OF	1.121
3	THE IRAQI PEOPLE	97	3	THE IRAQI PEOPLE	661
4	OF THE PEOPLE	90	4	A LOT OF	638
5	PEOPLE IN THE	77	5	LOT OF PEOPLE	505
6	THE PEOPLE WHO	71	6	FOR PEOPLE TO	400
7	THE BRITISH PEOPLE	59	7	PROTECT THE AMERICAN	346
8	I THINK PEOPLE	58	8	TO PROTECT THE	345
9	FOR PEOPLE TO	58	9	PEOPLE WHO ARE	342
10	OF PEOPLE WHO	56	10	OF THE PEOPLE	307
11	PEOPLE WANT TO	54	11	OF THE AMERICAN	274
12	PEOPLE WHO HAVE	51	12	PEOPLE IN THE	269
13	OF PEOPLE IN	51	13	TO THE AMERICAN	269
14	SAY TO PEOPLE	48	14	AMERICAN PEOPLE AND	269
15	THOSE PEOPLE WHO	47	15	TO THE PEOPLE	264
16	A LOT OF	46	16	PEOPLE OF THE	246
17	THE PEOPLE THAT	44	17	FOR THE PEOPLE	218
18	FOR THE PEOPLE	41	18	PEOPLE WHO HAVE	202
19	THAT PEOPLE ARE	40	19	AND THE PEOPLE	194
20	TO THE PEOPLE	37	20	FOR THE AMERICAN	181
21	THAT PEOPLE HAVE	37	21	THE PEOPLE WHO	178
22	PEOPLE IN EUROPE	37	22	THE PEOPLE IN	165
23	LOT OF PEOPLE	37	23	TO HELP PEOPLE	163
24	THE PEOPLE IN	36	24	OF PEOPLE WHO	146
25	THERE ARE PEOPLE	34	25	PEOPLE OF THIS	144
26	PEOPLE IN IRAQ	31	26	AND THE AMERICAN	135
27	IS THAT PEOPLE	30	27	AMERICAN PEOPLE TO	128
28	SO MANY PEOPLE	29	28	PEOPLE TO UNDERSTAND	125
29	PEOPLE WHO WANT	29	29	OF PEOPLE IN	123
30	OF THE IRAQI	29	30	OF THE UNITED	122

In both four-word clusters lists *a lot of people* ranks first but, if we scroll down, it is clear that the main concerns of the two politicians are different, and this is confirmed in the five-word clusters list, where *to protect the American people* ranks first in Bush's speeches, followed by *important for the American people* and *security of the American people*, versus *the vast majority of people* in Blair's, followed by *will of the Iraqi people*.

If word lists, which Scott (2001: 47) compares to a Swiss army knife's scissors, appear to be good indicators of what is important in a text, it is, however, the keywords function that allows us to arrive at what the text is about, "avoiding any trivia and insignificant detail. What the text boils down to is its keyness, once we have steamed off the verbiage, the adornment, the blah blah blah" (Scott & Tribble 2006: 55–56).

Keyness in this paper indicates two qualities: "aboutness" and importance. The keywords function allows us to compare our node corpus wordlist to a reference corpus, and the items that emerge from the comparison – the keywords – are those that have generated the greatest statistical prominence when compared with the reference corpus. The outstanding items will be analysed in the next sections.

Table 3. First fifty key words obtained by referencing one million words uttered by Blair against 5 million words uttered by Bush

N	Key word	Freq.	%	RC freq.	RC %
1	IS	21.098	2,34	62.294	1,24
2	PRIME	3.217	0,36	2.582	0,05
3	MINISTER	3.222	0,36	2.606	0,05
4	THAT	22.652	2,51	77.876	1,56
5	IT	12.304	1,36	35.760	0,71
6	BLAIR	1.108	0,12	248	
7	EUROPEAN	1.242	0,14	429	
8	THINK	4.722	0,52	9.332	0,19
9	BUT	6.407	0,71	16.244	0,32
10	BRITISH	817	0,09	113	
11	EUROPE	1.322	0,15	847	0,02
12	QUESTION	2.361	0,26	3.244	0,06
13	THERE	5.314	0,59	12.821	0,26
14	TONY	910	0,10	266	
15	HAVE	9.767	1,08	31.459	0,63
16	UK	523	0,06	9	
17	AM	1.560	0,17	1.800	0,04
18	BRITAIN	722	0,08	226	
19	ACTUALLY	1.290	0,14	1.279	0,03
20	VERY	3.460	0,38	7.888	0,16
21	ARE	9.121	1,01	30.619	0,61
22	AFRICA	809	0,09	718	0,01
23	CLIMATE	457	0,05	131	
24	WE	14.627	1,62	60.666	1,20
25	COUNTRIES	1.438	0,16	2.638	0,05
26	PROGRAMME	262	0,03	0	
27	INDEED	454	0,05	194	
28	UN	276	0,03	14	

29	IRELAND	335	0,04	73	
30	LONDON	364	0,04	115	
31	UNION	688	0,08	743	0,01
32	WHAT	4.771	0,53	16.186	0,32
33	SHOULD	1.360	0,15	2.760	0,06
34	POINT	889	0,10	1.329	0,03
35	NHS	209	0,02	0	
36	SITUATION	634	0,07	713	0,01
37	SAY	2.301	0,25	6.276	0,13
38	EU	347	0,04	154	
39	LABOUR	195	0,02	0	
40	CHANGE	1.202	0,13	2.493	0,05
41	BEHAVIOUR	180	0,02	0	
42	WHICH	2.466	0,27	72.202	0,14
43	INCIDENTALLY	190	0,02	5	
44	AGREEMENT	632	0,07	817	0,02
45	NORTHERN	295	0,03	115	
46	TERRORISM	701	0,06	1.011	0,02
47	THING	1.133	0,13	2.363	0,05
48	WHOLE	628	0,07	641	0,02
49	OBVIOUSLY	750	0,06	1.178	0,02
50	SERVICES	538	0,06	635	0,01

Berber-Sardinha (2004: 101–103) indicates that similar results may be expected with a reference corpus larger than five times the size of the corpus under analysis or much larger, whereas one that is less than five times the size of the node corpus may not be reliable. His assumptions are confirmed in the present investigation, where the one million words uttered by Blair were compared to a ten million-word corpus spoken by Bush: very much the same keywords were generated with a very slight difference in the percentage of occurrence. Furthermore, when the same one million words were compared to 2.5 million words spoken by Bush, the difference was again in the percentage of the items, and the keywords yielded were largely the same.

In general three claims can be made, according to Scott (2009): (1) the choice of the reference corpus will affect the results; (2) features which are similar in the reference corpus and the node corpus will not surface in the comparison; (3) only features where there is a significant departure from the reference corpus norm will become prominent for inspection.

The assumption that the choice of the background corpus will affect the results is further supported by comparing the same foreground corpus with a corpus of general written English: this was obvious and highly expected, in that, as evidence in the previous findings shows, by comparing spoken with spoken we lose the "speechiness" that emerges when comparing spoken with written, where

words like *we, I, you, that, think* predictably come top of the list. Furthermore, the word *people*, which was the most common word in both Blair's corpus and Bush's corpus, and therefore was not generated in the keywords list, ranks 9th when comparing spoken to written, suggesting perhaps that the collective noun *people* belongs more to the spoken than to the written medium.

3. One million words vs five million words

Let us examine Table 3 in more detail. This procedure has thrown up a set of keywords which can build an understanding of the major themes addressed by the former Prime Minister.

Looking at the list, we soon notice that the top lexical words offer few surprises. Proper nouns and countries are up-played: *Prime, Minister, Blair, European, British, Europe, Tony, UK, Britain*. It is not surprising that they figure high in the list, together with other words typical of British culture (and British spelling). *Actually*, *indeed* and *incidentally* are regarded more as indicators of style rather than as key indicating "aboutness" or "ofness" (Milizia 2006: 47).

The first item that caught my eye is *climate*. The position of *climate* in the Keywords list is reported below:

Table 4. Ranking of CLIMATE in the Keywords list

N	Key word	Freq.	%	RC. freq.	RC. %
23	CLIMATE	457	0.05	131	0.00007

The word is pronounced on 457 occasions (0.05% of the token count) by Blair and on 131 occasions (0.00007%) by Bush. The word *climate* was found to be key in Blair's speeches, as a good indicator that *climate* is, for him, the dominant theme of the period in question. For Bush, though, the number was too small to show any significance in the five million-word corpus.

What is worth highlighting is that in all the keywords lists generated so far (except the one that emerges by comparing spoken with written), the word *climate* was consistently found to be key, and this is regarded as a good indicator that *climate* is the dominant theme of the period in question. Adopting a Firthian (Phillips 1989) perspective, the aboutness of a text depends on the context in which it is embedded: KW analysis offers a very effective way of building up not only what is going on in a text but also in a given context. Thus, evidence

corroborates the well-known commitment of Tony Blair in the years for which we have data for the climate.[4]

The immediate, and also the not-so-immediate, environment of *climate* was investigated, and by means of the *Clusters* facility provided by *WordSmith Tools* it emerged that *climate* was almost always found in company of *change*, which was also picked up as key.

Table 5. Ranking of CHANGE in the Keywords list

N	Key word	Freq.	%	RC. freq.	RC. %
40	CHANGE	1.202	0.13	2.493	0.05

Thus, the analysis was based on the association of these two words, *climate change*.

3.1 The main concern of Tony Blair in 2005–2007?

A quick look at the behaviour of *climate* in Blair's speeches delivered before 2005, on the other hand, shows that *change* is obviously there but the two words hardly attract each other and *climate* is embedded in phrases which never occur from 2005 to 2007, such as the *right climate for investment*, *climate of fear*, *climate of trust*, *creating a climate that is conducive to*, *climate of confidence*, *a climate in which people are ready*. This seems to suggest that *climate change* is prioritized in the years for which we have data more than it was in the years from 1997 to 2005.

If climate change does not appear to be an issue in the period prior to this investigation, it certainly is in the years from 2005 to 2007. It has been noted (Duguid 2004) that *issue* is a key noun which seems to be omnipresent in the world of Number 10: in Blair's speeches the item *issue* is used twice as much as in Bush's, and its salience in terms of frequency amounts to 0.11% of occurrences versus 0.06%. Yet, relying on the *Clusters* facility, the phrases emerging in the years prior to this analysis include neither *issue* nor *challenge*. Conversely, *the issue of climate change* and *the challenge of climate change* are the most frequently used phrases in the years under investigation. The table below shows 15 five-word clusters around the word *climate*, which always keeps company with *change*:

4. Together with Africa, which ranks top of the list in all keywords lists.

Table 6. 15 five-word clusters around the word *climate*

N	Cluster	Freq.	Length
1	THE ISSUE OF CLIMATE CHANGE	36	5
2	IN RESPECT OF CLIMATE CHANGE	16	5
3	THE CHALLENGE OF CLIMATE CHANGE	16	5
4	THE EFFECTS OF CLIMATE CHANGE	8	5
5	THE THREAT OF CLIMATE CHANGE	8	5
6	THE ENVIRONMENT AND CLIMATE CHANGE	8	5
7	THE JOINT DECLARATION ON CLIMATE	8	5
8	THE FIGHT AGAINST CLIMATE CHANGE	8	5
9	ON THE ISSUE OF CLIMATE	8	5
10	FOR REASONS OF CLIMATE CHANGE	8	5
11	AND THE JOINT DECLARATION ON	8	5
12	AFRICA AND ON CLIMATE CHANGE	8	5
13	IN THE FIGHT AGAINST CLIMATE	8	5
14	JOINT DECLARATION ON CLIMATE CHANGE	8	5
15	ISSUES TO DO WITH CLIMATE	8	5

It is apparent, though, that the relationship between the two words is not symmetrical but to some degree one-sided, "just as one person may be desperately in love with another but the other may not return that love with the same intensity, so it is with words" (Scott & Tribble 2006: 37). As we will see, *change* is definitely happy with other words, whereas *climate* seems to be happy almost exclusively with *change*.

To illustrate, there are 457 instances of *climate* and 1202 of *change* in the corpus and 402 occasions where they are found together. As *change* itself is more than twice as frequent as *climate*, these numbers suggest that *climate* likes *change* much more than *change* likes *climate*. More precisely, in his speeches, Blair was more than twice as likely to refer to change when discussing climate than he was to mention climate if discussing change.

One might presume that climate change was less of an issue in the first years of his term but that it was at the heart of his agenda in the last three years. Relying on the importance of frequency as a guide to what is going on in a text, hence in a context, it seems to us that frequency might signal a change in priorities (Warren & Greaves 2007) and reflect the main concerns of the time.

The fact that climate change is a concern in Blair's speeches is very clear from the context in which the phrase is found. *Climate chaos* occurs on 10 occasions; words like *evil, damage, problem, threat, tackle, levy, menace, struggle, enemy, fight, disaster, deal with*[5] clearly suggest that *climate change* is found only in a context

5. Partington (2003: 18–19) analyses, among others, the semantic prosody of *deal with* in political press briefings, showing that it very often collocates with unpleasant items.

of negativity and that this bigram tends to carry a negative pragmatic load. At the launch of the Clinton Climate Initiative, on 1 August 2006, Tony Blair compares climate change to fascism:

> Every now and then a generation is called upon to make enormous efforts to defeat something appalling. For my parents' generation that was a war to defeat fascism, a war that consumed 60 million lives and consumed the best years of their lives, and it required a vast effort, bringing together America and the former Soviet Union, and Great Britain, in an alliance in which other differences were put aside because of the scale of the evil they faced that had to be defeated. Climate change is not an evil, it is not a conscious force, it is not someone's plan, but it threatens life on this planet every bit as much as the threat of fascism threatened our parents' generation, and we have to do what they did to try to advance on every front, to use all our energy to mobilise the resources that are necessary to create a sustainable world economy in which not just our prosperity can be secured, but the emerging nations can join in that prosperity as well.

Climate change is seen not only as an environmental challenge, but also as an economic challenge, a social challenge, and it actually represents a major challenge to the overall question of national security: climate change seems now to be beyond politics.

4. Uncovering n-grams and concgrams

As mentioned earlier, the other search engine used to interrogate the corpus is *ConcGram* 1.0 (Greaves 2009), able to identify not only collocations that are strictly consecutive in sequence but also non-consecutive linkages. Thus, the software finds n-grams such as *a lot of people*, but also the same pattern in concgrams such as *a lot of local people* or *a lot of different people* (Cheng, Greaves & Warren 2006: 412).

To illustrate, the bi-gram *hard word* and *work hard*, and the concgrams created around these two words, already investigated in the OSTI Report (Sinclair et al. 2004), are extensively analyzed in Cheng, Greaves, Sinclair & Warren (2009), displaying concgrams such as *work so hard*, *hard place to work*. Other examples that Warren provides are n-grams such as *role play* and *play a role*, exhibiting concgrams such as *play a minor role*, *played only a cameo role*, *plays a much less important role*.

In the spoken political corpus used for the present investigation, looking at the tri-gram *fight against terrorism*, concgrams such as *fight against the menace of terrorism*, or *fight against the ugly scourge of terrorism* are also uncovered (Milizia 2009). *ConcGram* identifies not only constituency variants (AB, ACB), but also positional variants (AB, BA), such as *a valuable ally in the war on terror* and *in the*

war on terror we have no better ally (Milizia & Spinzi 2008), or *a failure to connect the dots* and *the dots were not adequately connected* (Milizia 2009).

The terminology adopted here – concgram, prototypical, canonical – is based on Cheng, Greaves and Warren's work, but these concepts date back to 1970 when, in the OSTI Report (Sinclair et al. 2004), Sinclair spoke of the canonical form that would be the prototype of a phrase and the canonical form, distilled by the computer, with all the possible variations (see also Milizia 2009; Milizia & Spinzi 2008). The COBUILD team attempted, with limited success in the 1980s, to automatically search for non-contiguous sequences of associated words. *ConcGram* 1.0 has been designed to perform primarily fully automatic searches, and no prior intervention of the user makes the search a true corpus-driven analysis (Tognini-Bonelli 2001).

The word *climate* is regarded here as the origin in Greaves' terminology, and *change* as the associated word, but *change* does not stand in a hierarchical position with *climate* at the centre of the attention, as happens in the node and collocate relationship. Rather than focusing on the node, *ConcGram* highlights in colour[6] all the associated words in each concordance line. For lack of space, the 469 instances of contiguous association between *climate* and *change* will not be displayed. Here we report only a few examples of both positional and constituency variants to show the strength of attraction between these words:

…out of this struggle than we can opt out of the **climate** **chang**ing around us. Inaction, pushing…
…clear from all the scientific evidence that the **climate** is **chang**ing, and I don't think there a…
…and Africa being important because if the **climate** carries on **chang**ing then the situation in Africa.
…on warming. We do not know how much the **climate** could, or will **change** in the future. We…
… reality is that most people now accept our **climate** is indeed subject to **change** as a…
…the Kyoto Protocol, this is going to affect the **climate** as it is today, but the **changes** to that will be very…
…to win the argument for the Euro in such a **climate**. When do you expect this will **change**, …
…of how, over time, as a result of the **chang**ing **climate**, countries will have to invest very…
…two crucial issues. One is the **chang**e in the **climate** and the actions that we need to take…
…organised crime, the **chang**ing of the **climate** and the environment, defence, foreign…
…rs arising from predicted **chang**es in **climate**. Here the predictive capability of the science

Figure 1. Climate change concgrams

These discontinuous associations would go unnoticed if we were to look only at adjacent sequences.

Turbulence with respect to the canonical form is minimal in the first examples – *the climate changing around us, the climate is changing, the climate carries*

6. The origin is highlighted in pink, and the associated words are red and light blue.

on changing, the climate could or will change; greater diversion is displayed in the following examples – *the changing climate, the change in the climate, the changing of the climate* – for the sheer reason that we were accustomed to finding *climate* to the left of *change*. Considerable turbulence is exhibited in these two instances – *our climate is indeed subject to change, the climate as it is today but the changes to that* – where all the intervening words dilute the collocation. Despite the intrusion, the attraction between *change* (which in the above cases has both a noun and a verbal function) and *climate* is still very strong, and an endocentric relationship still holds between the two words, i.e. they are combined to create a single semantic entity (Sinclair & Warren 2006).

The last line may perhaps arouse interest, in that it is the only one where *changes* appears in the plural form, but a manual analysis has revealed that it was uttered, in a joint press conference held on 9 June 2006 with Jacques Chirac, by the French President.[7]

4.1 An issue, a challenge, a threat

Guided by the Clusters facility provided in *WordSmith Tools* and taking into account the phrases shown in Table 6, by means of *ConcGram* a search was made for 5-word concgrams, focusing on *the issue of climate change* (constraints of space do not allow us to illustrate two very common clusters, i.e. *the challenge of climate change* and *the threat of climate change*).

Although we read in Table 6 that the five words in *the issue of climate change* seem to co-occur on 36 occasions, Figure 2 shows that these words are found together 72 times. The concordance lines are displayed below to best show the variation found within a concgram: here the layout appears clearly, with *issue* set as the centred word and *climate change* forming a diagonal, thus giving a clear idea of the proportions of the pattern (Sinclair & Warren 2006).

1 **climate change**. But we have in both those huge **issue**s facing us the possibility of making a
2 on **climate change** the other day, this is the **issue** that is now being driven right across
3 of **climate change**. You have helped put this **issue** at the top of the international agenda.
4 to **climate change** and there are many other **issue**s that are important in this summit. But
5 Africa, on **climate change**, which are obviously **issue**s for our G8 chairmanship as well, and in

7. French, unlike Italian which seems to prefer the plural form, is indeed in favour of both: *changement climatique* and *changements climatiques*. Thus, *changes in climate* as spoken by the French President is the English translation of the plural form, which seems to be as common as the singular. Conversely, in the Italian corpus the canonical form is *cambiamenti climatici*, with a few occurrences of *cambiamenti climatico-ambientali*.

6 problems of **climate change**. And on a domestic **issue** that has picked up while you have been
7 global poverty, **climate change**, all of these are **issues** where Britain and Australia have much
8 introduced the **climate change** levy which was an **issue** for a lot of business, but actually what
9 such as **climate change** and environmental **issues**. I am very happy to have had this
10 today. Africa and **climate change** are two such **issues** that cry out for such an approach. We
11 didn't. Clearly the **climate change** thing is an **issue** which has arrived since Brandt and that
12 risks associated with **climate change** or related **issues** of security of energy supply, we need
13 Minister. Obviously **climate change** is a major **issue** for my generation and the next, therefore
14 time that although **climate change** is a major **issue**, although I think increasingly people
15 So on Africa, on **climate change** and on the **issue** of the Middle East, I think we have made
16 we need to lead the way. **Climate change** is an **issue** where charity very much begins at home.
17 Minister's commitment to the **climate change** **issue**, will he get his civil servants to look
18 Minister Blair has given on the **climate change** **issues**, including being prepared to put time
19 priority that we attach to the **climate change** **issue**. It is why we introduced the **climate cha**
20 of vehicle building and **climate change** **issues** when it is delivered by those sorts of
21 and you can't resolve the **climate change** **issue** without the involvement of the United
22 a much better focus now on that **climate change** **issue** than there was before, and indeed there
23 real progress has been made; but on the other **issue**, **climate change**, so far we have no
24 also saw eye to eye that on the Iranian nuclear **issue**, **climate change**, development assistance
25 be random and savage. So just take these three **issues**: **climate change**; Africa and world trade
26 treaty. The second point we discussed was the **issue** of **climate change** and energy. Last year
27 country leads the world both in terms of the **issue** of **climate change** and also of course me
28 this country has been at the forefront on the **issue** of **climate change** and will continue to
29 cause people know today that we have both the **issue** of **climate change** and also the concerns
30 bring America into the consensus on tackling the **issue** of **climate change**, we will never ensure
31 even in the past couple of years, on the **issue** of **climate change** that it is not only
32 day, I think that it is very important that the **issue** of **climate change** should become a major
33 particularly on the G8 process as well and the **issue** of **climate change**. And I think here that
34 and he will speak directly to you on the **issue** of **climate change**. We warmly thank him
35 step forward in the way that we deal with the **issue** of **climate change** and the environment
36 about security and cost of supply, and also the **issue** of **climate change**. Now I think in the
37 our deliberations: the **issue** of Africa and the **issue** of **climate change**. In respect of Africa
38 in there. Look first of all just on the **issue** of **climate change** and what we can do
39 especially in Africa, and then in respect of the **issue** of **climate change** where we had a very
40 including with America, on trying to tackle the **issue** of **climate change**. Interviewer But these
41 five countries that had come to discuss the **issue** of **climate change**, we all came together
42 whether it is on Aids, or on Africa, or on this **issue** of **climate change**, he is still providing
43 need this multilateral system to address these **issues** of **climate change** and so on, as well as
44 has done a tremendous amount in addressing **issues** of **climate change**. To remain competitive
45 that we saw exhibited yesterday. The second **issue** was **climate change**. Now here let me be
46 we get a good agreement at the G8 in July on the **issues** of **climate change** and Africa, which are

47	thank you very much for linking the two	**issue**s of **climate change** and poverty. You talk
48	importance of tackling common environmental	**issue**s such as **climate change**, which again if
49	from words turned into deeds. So take these	**issue**s: Africa, **climate change**, world trade.
50	not world trade, not the ability to tackle	**issue**s to do with **climate change**, none of
51	European grid, research and development on	**issue**s to do with **climate change** and of course
52	to do with development but of course on the	**issue**s to do with **climate change** as well, and
53	better on the Security Council. We will discuss	**issue**s to do with **climate change** where the
54	say than any other international leader on this	**issue**. He has put **climate change** at the heart
55	this G8 achieve, particularly on the difficult	**issue**s of trade and **climate change**? Sir Bob
56	say than any other international leader on this	**issue**. He has put **climate change** at the heart
57	this G8 achieve, particularly on the difficult	**issue**s of trade and **climate change**? Sir Bob
58	have to be. It is true that each of the three	**issue**s – world trade, **climate change**, Africa –
59	it was very nicely done. And I think that the	**issue**s incidentally of **climate change** and
60	have realised that. What does Europe mean for	**issue**s like, as you said, **climate change**,
61	to supply more than 30% of their water. These	**issue**s, and the challenge of **climate change**,
62	to supply more than 30% of their water. These	**issue**s, and the challenge of **climate change**,
63	have agreed to work very closely together on the	**issue** of the environment and **climate change**
64	simple answer. But I think when you look at the	**issue**s to do for example with **climate change**
65	right at the forefront of our deliberations: the	**issue** of Africa and the **issue** of **climate change**
66	the technologies necessary to deal with this	**issue**. And as we discuss a global **climate change**
67	in terms of the long term future there is no	**issue** that is more important than **climate change**
68	effectively and then get on with many of the	**issue**s to do with the economy, and **climate change**
69	it. Tomorrow we will obviously be talking about	**issue**s to do with the environment and **climate**
70	point in the right hon. Gentleman raising these	**issue**s while he remains opposed to the **climate**
71	**ange** because they have indicated they wish the	**issue** of **Climate Change** to be a major part of
72	summit where we will be discussing two crucial	**issue**s and one is the **change** in the **climate** and

Figure 2. 5-word concgrams with *issue* and *climate change*

As Figure 2 shows, the strength of attraction between *climate* and *change* persists but in different ways: *the issue of climate change* (lines 26–44) is easily identified as the prototypical configuration, mainly because it is the most prominent in terms of frequency. The canonical form serves as the benchmark against which all of the other configurations are compared: the greater the deviation from the base form the greater the turbulence (Cheng et al. 2009). Since English relies both on left and right constructions, *the climate change issue* (lines 17–22) can safely be defined as stable as the base form, adding zero turbulence to the form which we have defined prototypical.

The first instances of positional variation are provided in the first lines, e.g. *climate change is an issue, the climate change levy was an issue, the climate change thing is an issue, climate change is a major issue*. Minimal divergence appears in

lines 50–53, *the issue to do with climate change*. The same is true, of course, for the plural form, e.g. *the issues of climate change* (lines 46–47). Looking at the plural form of the noun, it becomes easy to identify the top priority issues which are at the heart of Tony Blair's agenda: climate change, Africa and world trade, as several lines show. Africa and climate change appear in the same neighbourhood on several occasions, and this was not surprising given that Africa was also picked up as key, both as an individual word (position 22), and in the cluster list (Table 6).

In some concordance lines, mainly those where *climate change* is displayed to the right of *issue*, so many words intrude that the phrase would undoubtedly be overlooked with a search engine able to handle only adjacent n-grams: *what does Europe mean for issues like, as you said, climate change* (line 60), *the issue to do for example with climate change* (line 64), *two crucial issues and one is the change in the climate* (line 72). Crucially, these forms are increasingly turbulent relative to the canonical form as the number of intervening words increases.

Line 67 – *in terms of the long term future there is no issue that is more important than climate change* – adds further support to the assertion that keywords predictably relate to the major ongoing topics of debate (Partington 2003: 24).

In a press conference with Tony Blair and Governor Arnold Schwarzenegger at a BP plant in Los Angeles, held on 31 July 2006, in the attempt to work together on reducing greenhouse gas emissions and promoting new clean fuel technologies, and in the hope of solving climate change and delivering a clean energy future, Steve Howard, CEO of the Climate Group, said:

> Prime Minister Tony Blair has done more I would say than any other international leader on this issue. He has put climate change at the heart of the G8 agenda, invited in the five major developing countries to build a strong international accord. He set a 60% reduction target for the UK, put climate change at the heart of the EU agenda and really is an exceptional leader on this subject.

5. Conclusions

This paper has attempted to describe how words significantly prefer each other's company whether in adjacent pairings or in discontinuous phrasal frameworks and are conventionally found to attract each other for factors that go beyond grammatical norms (cf. Renouf & Banerjee 2007b).

Specifically, the focus has been first on the word *climate*, which was found to be key by referencing one million words uttered by Tony Blair in the years 2005–2007 against five million words uttered by Bush over the same period, and then on the cluster *climate change*, in that *change* was also picked up as key

and was always found in company with *climate*. The keywords list yielded by comparing the speeches of Blair and Bush is a list of items which appear significantly more frequently in the corpus under study than they do in the larger reference corpus. The procedure used for identifying keywords in the present research was the one devised for use in *WordSmith Tools*, and is based on simple verbatim repetition, without any attempt to identify or match up the semantics or pragmatics (Scott 2009).

Climate – and even more so *climate change* – have in fact proved to be good indicators of keyness, clearly reflecting Blair's leading themes of the years under investigation. It has thus been found that this theme was topping his agenda, together with Africa and European matters (the United Kingdom held the Presidency of the Council of the European Union from 1 July until 31 December 2005).

Sinclair (Sinclair & Warren 2006), in this regard, has spoken of "aboutgrams" indicating what the text is really about and, in the case of spoken discourse, what the priorities of Tony Blair's government were in the years under investigation.

This work, part of a larger-scale project which analyses n-grams and concgrams in British, American and Italian political language, has also tried to illustrate that phraseology plays a prominent role in discourse, and phrases are usually much better at revealing the "ofness" of the text (and the context) than individual words. Thus, *WordSmith Tools*[8] and *ConcGram* were used to uncover all the phrases, both contiguous and non-contiguous, created around *climate change*, obtaining patterns of non-adjacent associations that would otherwise go unnoticed when some words – one, two or more – intervene to dilute the collocation. Hence, phrases that typically, or occasionally, occur in non-contiguous sequences risk going undiscovered. Indeed Cheng, Warren & Greaves (2006) have maintained that the majority of concgrams occur in non-contiguous sequences. More recently, Cheng et al. (2009) has gone even further by saying that "some MSUs[9] only realize the meaning uniquely when they are not contiguous".

Such searches, which have been crucial in highlighting all the possible words associations, contributing to the elucidation of the phenomenon of phraseology, have proved to be an invaluable aid to uncover more of what Sinclair has termed the "idiom principle".

---

8. The new version of *WordSmith Tools* 5.0, includes, among other things, the possibility to handle non-contiguous sequences. This tool is provided in Utilities, WSConcGram.

9. Sinclair (2007) was talking of Meaning Shift Unit (MSU) to refer to the lexical item (1996) (in Warren 2007).

References

Berber-Sardinha, T. 2004. *Linguistica de Corpus*. Barueri, Sao Paulo: Manole.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Cheng, W., Greaves, C. & Warren, M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics,* 11(4): 411–433.

Cheng, W., Greaves, C., Sinclair, J. & Warren, M. 2009. Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics,* 30(2): 236–252.

Duguid, A. 2004. Men at work: how those at Number 10 construct their working identity. In *Discourse, Ideology and Specialised Communication*, Garzone, G. & S. Sarangi (eds.), 453–481. Bern: Peter Lang.

Erman, B. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12(1): 25–53.

Fletcher, W. H. 2002–2007. *KfNgram*. Annapolis MD: USNA.

Greaves, C. 2009. *ConcGram* © 1.0. *A Phraseological Search Engine*. Amsterdam: John Benjamins.

Milizia, D. 2006. Classifying phraseology in a spoken corpus of political discourse. *ESP Across Cultures* 3: 41–65.

Milizia, D. 2009. Migration of n-grams and concgrams in political speeches. In *Forms of Migration – Migration of Forms*: *Atti del XXIII Convegno Nazionale AIA*: Bari: Progedit: 496–514.

Milizia, D. & Spinzi, C. 2008. The terroridiom principle between written and spoken discourse. *International Journal of Corpus Linguistics* 13(3): 322–350.

Partington, A. 2003. *The Linguistics of Political Argument*. London: Routledge.

Phillips, M. 1989. Lexical structure of text. In *Discourse Analysis Monographs* 12. Birmingham: University of Birmingham.

Renouf, A. J. & Banerjee, J. 2007a. Lexical repulsion between sense-related pairs. *International Journal of Corpus Linguistics* 12(3): 415–443.

Renouf, A. J. & Banerjee, J. 2007b. The search for repulsion: A new corpus analytical approach. In *Towards Multimedia in Corpus Studies*, Vol. 2, T. Nevalainen, I. Taavitsainen, M. Korhonen & P. Pahta (eds). <www.helsinki.fi/varieng/journal/volumes/02/renouf_banerjee/>.

Scott, M. 2001. Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In *Small Corpus Studies and ELT. Theory and Practice* [Studies in Corpus Linguistics 5], M. Ghadessy, A. Henry & R. L. Roseberry (eds), 47–67. Amsterdam: John Benjamins.

Scott, M. 2008. *WordSmith Tools* 5.0. Liverpool: Lexical Analysis Software.

Scott, M. 2009. In search of a bad reference corpus. In *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*, D. Archer (ed.), 79–92. Oxford: Ashgate.

Scott, M. & Tribble, C. 2006. *Textual Patterns. Keywords and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Sinclair, J. 1987. The nature of the evidence. In *Looking Up: An Account of the COBUILD Project in Lexical Computing*, J. Sinclair (ed.). London: HarperCollins.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: OUP.

Sinclair, J. 1996. The search for units of meaning. *Textus* 9(1): 75–106.

Sinclair, J. 1998. The lexical item. In *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 171], E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.

Sinclair, J. 2003. *Reading Concordances*. London: Longman.
Sinclair, J. 2007. Collocation reviewed. Ms, Tuscan Word Centre, Italy.
Sinclair, J., Jones, S., Daley, R. & Krishnamurthy, R. 2004. *English Collocation Studies. The OSTI Report*. London: Continuum.
Sinclair, J. & Warren, M. 2006. Manuscript "JMS comments on PUHK paper about 'different people'". Personal communication with the authors.
Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.
Warren, M. & Greaves, C. 2007. Concgramming: A corpus-driven approach to learning the phraseology of English. *ReCALL Journal* 19(3): 287–306.
Wilks, Y. 2005. REVEAL: The notion of anomalous texts in a very large corpus. Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 31 June–3 July.

Key words and key phrases in a corpus of travel writing

From Early modern English literature to contemporary "blooks"

Andrea Gerbig
University of Heidelberg, Germany

WordSmith Tools offers three procedures (key words, key-key words and associates) to select words from texts that are noteworthy in a statistical sense. All three procedures are applied in the present paper in order to make visible, statistically, the most noticeable propositional and stylistic changes in the representation of a specific field of interest over the course of five centuries. Some of these quantitative results (identified through key-key words) are followed up in more detail and their development and role throughout the centuries is discussed, for example in grammaticalization processes. These findings are supplemented further by an analysis of the most frequent phrases and n-grams, again diachronically. Together, they show the development of linguistic habits in society.

1. Introduction

Investigating "key words" has a long tradition in linguistics and related social sciences. I will very briefly trace the tradition from intuitive approaches, such as by J. R. Firth and Raymond Williams to the statistically based one developed by Mike Scott. This concept of 'keyness' seems to be of increasing interest in linguistics, worked at with different methods and aims in mind. Often, a (more or less explicit) constructivist view is taken, assuming that the specific use of the chosen words can give us insights into social and cultural preoccupations, attitudes and preferences. In the present context, a key word analysis is used to, among other goals, show the changing roles and implications of travelling in society over a considerable time span.

Individual key words, however, are usually only the centre of units of meaning in language. The concept of a 'unit of meaning' that goes beyond the individual

word has meanwhile been discussed extensively. Some of the leading work on the topic can be found in work by Sinclair (e.g. 1998, 1999), or Sinclair & Mauranen (2006), Stubbs (e.g. 2007) or Teubert & Čermacova (2007), to name just a few sources. Here, after having identified the key words in the texts I am investigating, I will look at multi-word units, or extended lexical units, but also briefly at the most frequent 5-grams, i.e. recurring 5-word combinations.

It will be interesting to see what information about texts and collections of text can be gleaned from different quantitative and qualitative techniques of investigation. Relative frequency of words in comparison to occurrence in other texts gives an indication of noteworthy propositional content, maybe also about deviant structural preferences. Propositional noteworthiness is always a good starting point to follow up its role in a socio-cultural context. To do so, we obviously have to go beyond the individual key word and look at phrases to perceive semantic convention and its probable pragmatic intention.

Absolute frequency foregrounds the structural aspects of texts. Function words are the most frequent ones in the language, and they are also the most frequent players in phrase structures. The frequency and distribution of key words and, to a certain extent also phrases, characterises texts. We need information from large background corpora, probably as well as from other, specialized corpora, to find out about important linguistic habits, in both general and text-specific terms. I will demonstrate these mechanisms in my diachronic travel corpus and will discuss the usefulness of the various relevant analytic procedures.

2. Key words: Implications and other studies

What makes a word a "key word"? Because of their habitual use in certain contexts along with the construction or transmission of contents, views and attitudes, these words gain a particular value in a socio-cultural community. It is not the individual key word that is of interest, but the whole unit of meaning with its typical collocational patterns and accompanying evaluation. Such units always occur as part of a particular text. Key words, as possible centres of units of meaning, can give information about the propositional and stylistic orientation of that text. Individual texts are tokens of text-types which themselves are higher structural orders in our conventions of communication. Discourse, as the analysable trace of our behaviour in the social world, shows our repeated, partly conventionalized ways of using language. The individual texts in my corpus are instances of a text-type. With each change in the texts, the diachronic development of the text-type becomes observable. This is of course a bi-directional process. As soon as writers are conscious that they are contributing to a text-type, for example a tradition of travel writing,

this influences their stylistic choices. Text-types then become intuitively definable. Over time, text-types also branch out into variants and sub-types. This process will be shown in the description of my sub-corpus of 21st century travel texts.

The idea of key words as indicators for central concepts and topics in a socio-cultural community has a long tradition. Firth (1935: 40–5) suggested investigating the "distribution of sociologically important words, what one might call focal or pivotal words in all their derivatives and compounds in sociologically significant contexts" (45). He had already referred to linguistic routines and habitually recurring collocations which he assumed to be linguistic representations of cultural concepts. Williams' *Keywords* (1976/83) is seen as a milestone (in linguistics as well as in cultural studies) because of its encompassing and detailed discussion of a choice of lexical fields. Hoggart (1957: 27–8) (independently of Firth), referred to repeatedly used expressions and phrases concerned with the "basic features of life". Such expressions are used by a society with commonly recognized meanings and evaluations.

There is related work for German, for example by Strauß, Haß and Harras (1989), who also intuitively select a range of words that they believe encapsulate and express major political, social and cultural themes in Germany at a certain period. The majority of key word studies on German concentrate on the assumed power of the choice of expressions in a political context. Wimmer (1982) investigates a "political language culture" (*politische Sprachkultur*) and Klein (1989) writes about "political semantics" (*politische Semantik*). Criteria for choosing the key words are their "specialness", "political and social explosiveness", "extraordinary pragmatic richness" and "appellative function". Klein (1989: 29ff) expands his analyses to entire lexical fields. He assumes that political parties want to fill such "lexical nets" with their political views and evaluations in a manipulative way. Critically investigating language use in political contexts was a historically and intellectually important exercise for Germans.

Bergsdorf (1985) explains the special role of key words with their alleged 'vagueness'. Politicians, for example, use this vagueness to reach as many voters as possible, who will construct their own interpretations building on the keywords. Together, this would make for a consensus amongst the people who would then feel positively towards the political parties and vote accordingly. Key words therefore can be thought to have an integrating function.

There are principled differences in finding key words. Williams (1983: 15), (like all the German writers mentioned above) uses a hermeneutic approach where key words are "significant, binding words in certain activities and their interpretation" and they are "significant, indicative words in certain forms of thought". This definition stresses individual choice. According to your personal political, or more general, "worldview", the decisions about "significance" will obviously

vary. Williams' evaluations and contextualisations of, for example, his key words *Alienation*, *Bourgeois* and *Capitalism* or *Ecology*, *Industry* and *Media* would very likely be approached differently today, because circumstances of life and political and social views have changed. Furthermore, a concept such as 'Alienation' (as well as the use of the word as such) probably has little meaning for a large part of today's language users.

A statistical method of selecting words according to their relative frequency is independent of personal views and preferences. The programme *WordSmith Tools* by Scott (1997a, b, 2000), also discussed in Scott & Tribble (2006), counts the occurrence of words in a text that one is interested in analysing and compares it with the frequency of these words in a reference corpus. Those words in the analysed text "that are most unusually frequent" in comparison with the reference corpus are key words. They indicate topic and (if not nouns) also style.

I will use the term "key words" here in the sense of Scott's approach, with the suite of programs *WordSmith Tools*. Following the criterion of significant relative frequency, I will further, based on intuitively felt 'importance' or 'expressiveness' categorise the words into lexical / semantic fields. This is an empirical way of determining aspects of the function of words in texts.

This procedure can successfully be applied to various tasks that focus on stylistic, socio-cultural, political, or other noteworthiness. In literary texts it can draw attention to otherwise difficult to detect stylistic moves of the author (see for example Stubbs (2005) on Joseph Conrad's *Heart of Darkness*, or Müller-Wood & Gerbig (2006) on Graham Greene's *Brighton Rock*). In work related to the paradigms of Critical Discourse Analysis, key words often serve as the starting point for analysing in more depth the representation of participants, situations and events in text types with a high impact (cf. Gerbig 2003).

3. The corpus

The corpus is a collection of travel writing from the 16th to the 21st century. Each century covers approximately 500.000 words, for the most part consisting of complete texts. The texts are as evenly distributed between the beginning, middle and end of each century as possible.

– C16: Contains writings by John Leland, Richard Torkington, Richard Hakluyt and Sir Walter Raleigh: We can read about descriptions of places in England, a pilgrimage to Jerusalem, and tales of travels and conquests of north-eastern Europe and adjacent countries, Madeira and the Canaries, ancient Asia and Africa, to the discovery of Guiana.

- C17: There are books by Robert Coverte, John Taylor, John Chardin, William Dampier, Celia Fiennes, John Fryer: They concern trade reports, a pilgrimage, and journeys through England, New Holland, East India and Persia.
- C18: Has work by Daniel Defoe, Henry Fielding, Samuel Johnson, James Cook, Hester Piozzi, Tobias Smollett, Lawrence Sterne: They write about journeys to east England, Scotland, Lisbon, France, Italy and Germany, towards the South Pole and round the world. They characterize their travel reports as *journal*, *observations*, *reflections*, and even *a sentimental journey*.
- C19: The books by Robert Louis Stevenson, John Franklin, Charles Dickens, Alexander Kinglake and Richard F. Burton tell about a pilgrimage through the south of France, a journey to the polar sea, through parts of the USA, to the near and middle East, and to east Africa.
- C20: The more recent contributions are by William Hudson, Norman Douglas, Bruce Chatwin, Bill Bryson and Beth Fowler. This is an interesting century because the means of travelling have changed so rapidly and dramatically in the time span from 1909–2000 that the above contributions cover.
- C21: The subcorpus of the 21st century consists of texts which are all published on the internet, on a well-structured and well-edited platform, not to be confused with the typical interactive weblogs, which are mostly of a highly colloquial style. The reports are of a coherent format, more like short-stories. The areas travelled to are Africa, Australasia, Asia, the Middle East and the Caribbean. The texts are travellers' reports and stories about their daily, pleasant and unpleasant experiences and adventures on the trips (they are all native speakers of English). Because the texts all appear on a platform maintained by a chain of shops providing the travellers with all necessary equipment, they are written for like-minded people, to prepare them for what to expect on their tours. (This chain of shops does not appear to interfere with the content of people's contributions, i.e. there are no marketing / advertising objectives noticeable although, obviously, a mere return to the website can be seen as a successful marketing move).

4. Method and findings

4.1 Key words in the travel corpus

The data were first accessed by checking the key words.

Scott & Tribble (2006) suggest a background corpus for comparison at least five times the size of the investigated text. As comparative data for the travel corpus (TC) C21 to C19 a self-compiled three-million word background corpus of

mixed written and spoken sources of contemporary, everyday British, Australian and American English was used. C18 to C16 were then compared with the Early Modern English section of the Helsinki Corpus (HC). The time frame for that part of the HC is 1500 to 1710 with a total of 551,000 words. The background corpus is here only slightly bigger than the investigated texts, but it is difficult to obtain other appropriate data from that time span.

The 100 most frequent key words derived for each century were then further intuitively grouped into sets of semantically related words. This grouping procedure bears interesting problems. It is very often not possible to assign the individual key word to a clear category. For example *wind*, out of its co-text, might be grouped along dimensions of 'weather' or 'sea travel'. This often makes it necessary to check the wider co-text of the concordance lines, again underlining Firth's legendary dictum that "you shall know a word by the company it keeps". The unit of meaning clearly is larger than a single word. One could of course also argue that 'weather' and 'sea travel' are not separate categories. The object of study is not self-evident but presents itself in collaboration with the researcher's intuitive grasp of the textual data.

In order of frequency, the semantically grouped key words in each century of the TC are the following:

C21 – comparison with current data

pronouns:	*we, my, I our, us, me* (top of list – PPs indicate a more spoken style)
travellers' life:	*bus, trip, beach, hotel, truck, driver, walk/ing/ed, tour, tourist/s, taxi, ride, backpackers, guide, food, hostel, restaurant, tent, airport, boat, locals, travel, flight, visit, arrived, border, breakfast, visa, bags, climb, stay, stop, toilet, guesthouse, photos, dinner*
countries/:	*town, city, road, island, Sydney, Australia, Bangkok, Thai/land,*
places:	*Korea/n Belize, mountain/s, rock, village, lake, river, coast, streets*

C20 – comparison with current data

location:	*Taiwan/ese, village, place, Ital/y/ian, town, earth, Calabria/n, Patagonia, streets, Naples, Rio, Mandarin, Chinese, Buenos Aires, valley, mountain/s, Ionian, river, Albanian, city, Chilean, Hellenic, road*
nature:	*sea, hill/s, tree/s, birds, stone/s, sheep, creatures, beast, landscape, rock*
colours:	*green, black, red, white, blue*
weather:	*sun/shine, rain*

C19 – comparison with current data

descriptive/:	*river, lake, Fort, hills, hunters, canoes, encamped/ encamp(e)ment,*
adventures:	*camels, (rein)deer, journey, tent/s, north, coast, shore, wood, sea, pines, voyage, desert, willows, trees, traveller/s, hill*
cultures:	*Indian/s, Somal/i, tribe, Berberah, Esquimaux, slave, Arab/s, Bedouins, Crees, Chipewyan*

weather: *ice, snow, wind, weather*
time: *seconds, morning, noon, night, day*

C18 – comparison with HC-EModE) among the 100 most frequent key words (spelling adapted):

sea travel: *island/s, board, latitude, longitude, sea, shore, captain, sail, bay, cape, boat*
weather: *wind, weather, ice, gale, snow*
destinations: *country, city, Italy, Paris*
people?: *natives, inhabitants, chief, people, public*

A comparison of the C18 data with the background corpus of current English (C20) stresses even more the frequency of key words relating to sea travel:

> *latitude, longitude, sea, island, shore, W(est), E(ast), captain, boat, bay, N(orth), voyage, cape, board, isle/s, canoes*

This shows how the background corpus brings out characteristic differences between the text under investigation and some "norm" where, obviously, today's norms are different from 18th century norms.

The key words procedure highlights the development in means of travelling. Although the travels in C16, 17 and 18 are already world-wide, this is not very prominently reflected in the key words. In C18, we see only reference to Italy and Paris, in C17 to Persia and in C16 to Russia, England, Persia and Spain. In the more recent three centuries investigated, references to countries are more varied and more frequent. Travelling by ship necessarily restricted the places that could be visited.

C17 – comparison with HC-EModE, among the 100 most frequent key words (spelling adapted):

travel/directions: *mile/s, country, north, west, south, distance – Persia/-n/-s*
description of countryside: *town, stone, hill/s, trees, house/s, garden/s, rocks, woods, wall/s, birds*
(few) sea travel: *island/s, sea, wind/s, land, bay, shore*

C16 – comparison with HC-EModE, among the 100 most frequent key words (spelling adapted):

pronouns: *they, their, our, we, us, themselves, them, you*
(sea) travel: *island/s, river, ship/s, sea, land, wind, captain, voyage, mariners, sailed, coast, journey*
: *north, east, west, northwest, southwest, south, southeast, northeast*
trading: *merchants, ware/s, goods, commodities, trade, delivered, company*
description/trajectory: *from, onto, upon, at, about, before, towards, against*

about destinations:	*country, Russe/s Russia, England, English, Persia, Moscow, Tartars, Spaniards*
rulers:	*emperor/s, majest/-y/-ies, ambassador*

These key words show the main topic of travelling throughout all centuries. Many of the more frequently occurring key words also appear well distributed as key words across the majority of texts. Such "consistency" (Scott & Tribble 2006: 29) can be seen as an indicator for genre specificity, considering that these words are unremarkable in terms of frequency in the background corpora.

Information about directions and position prevail in the earlier three centuries. In the later three centuries there is instead more information about the destinations travelled to as well as, obviously, that the means of travelling have changed.

Those key words that I have selected and grouped into semantically related sets are also indicative of changes in conventions of writing, or rather, of positioning oneself within the story. The most recent data, for example, are full of people's references to their own activities. Often, they are telling stories about themselves within exotic or strange settings. Their accounts of foreign countries are hardly informative about the places visited – which today is more or less taken as common knowledge anyway – they rather describe the writers' encounters with foreign influences while they themselves are the centre of attention. Although the investigated weblogs are not interactive, it seems to be a different text type than the other data, as such a result indicates. The newly created word *blooks*, something like blogs as books, indicates a fundamental change.

4.2 Key-key words and associates

Key-key words are those words that are found to be key words in a large number of individual texts in each travel sub-corpus. Key-key words show lexical choices which are typical for the genre. Because these key-key words occur so widely distributed and frequently throughout all the texts in the subcorpus, they are not specific to individual authors or peculiar choices of topic but rather document shared concerns and functions.

Associates are key words that are found to co-occur significantly with a key-key word, they "are *the set of words which are co-key with a given KW-node across a range of texts*. That is, if we take a KKW which is found in a number of texts and then determine all the set of co-key items in all those texts, we get associates" (Scott & Tribble 2006: 85). This gives a profile of the contexts of words and by extension also of the texts.

Key-key words

The most frequent key-key words in C21 (compared with two million words of mixed, every-day language use, also written) are the following (all occurring 5 times in each text):

– *guy, guys, locals, tourist, tourists*
– *tour, trip, road, walk, walking, ride, driver*
– *us, we, my*
– *water, beach*
– *up, around, back, off, to*
… *few, day, a next*

As far as possible, the key-key words were ordered semantically. The next frequent stratum is (all occurring 4 times):

– *hotel, village, town, city, center, shops, stay, styaed, night*
– *bus, taxi, boat, guide, hike*
– *breakfast, food, chicken, banana, lunch*
– *I, me, our, everyone*
– *bag, bags, backpackers, backpack, tourism, travelers*
– *along, after, out, headed*
– *hour, hours, minutes, awhile*
… *huge, didn't, stopped, decided, toilet*

From these lists, a predominance of nouns is visible. The topics covered seem to be typical for travellers or rather, backpackers. There is concern with the daily needs of moving around, staying overnight, eating and meeting people. Those people, however, seem to be characterized by a fairly restricted group of words, as the very general use of *locals*, *tourists*, *backpackers*, *travelers* and *guys* shows. Also, they are referred to with focus on their function as *driver* or *guide*.

The most frequent key-key words in C20 (compared with three million words of mixed, every-day language use, also written) are the following (there are fewer words occurring 5 and 4 times in each text than in C21, the majority of those given here occur 3 times):

– *hills, hill, place, spot, stones, sea, mountain, mountains, valley, land, landscape, trees, earth, sun, sunshine, earthquake*
– *green, red, black, little, remote, dim, charm*
– *city, village, villages, town*
– *me, my, myself, its, they, their, them, his, some, every*
– *distant, among, amid, hither*
– *beast, creatures, folks*
… *old, a, once, discovered, like, with, into, face, dead, and, ride, was, fashion, yet, had*

The key-key words of the C20 subcorpus still show typical coverage of travel topics, however, the difference to the C21 subcorpus becomes clear immediately. There is more concern with details of the things seen and places travelled to. The language seems much more descriptive. Compared to the C21 subcorpus, there is no informal use such as *guy/s*. There is further no reference to fellow travellers or tourists. Overall, coverage seems to be more varied, as the higher number of less frequently occurring key-key words than in C21 indicates. For the following centuries, only the top key-key words will be given. This is just to check if the general tendencies change.

The first 30 key-key words of C19 (compared with contemporary data, as above), occurring five, four and three times in each text are:

– *sun, trees, river, hills, hill, stones, land, mountains*
– *scarcely, steep, stony, distant*
– *with, upon, of, by, from, and*
– *fire, night, water*
– *traveller, neighbourhood, men*
– *their, his, we*
… *slave, saddle, portion*

C18 – comparison with the Helsinki Corpus part of Early Modern English (7,6,5 occurrences per text)

– *few, little, less, every, a, the*
– *with in, up, on, upon, at, from, near*
– *wind, island, inhabitants, sea, country*
– *we, their, its, own, myself*
– *situation, different, continued, general, attention, public, obliged*
– *seems, however, could, than*
– *having, has, have*
… *been, which, who, very, seen*

C17 – HC EModE (6,5,4)

– *water, sea, mile*
– *horses, birds*
– *side, on, near, at, top, between, along, from, about*
– *large, thick, little, very, some, black, great, round, abundance*
– *their, they, we, each, many*
– *the, a*
– *sorts, sort*
… *being, are, seen*

C16 – HC EModE (8,7,6,5)

- *of, upon, under, into, with*
- *south, west, east, north, towards*
- *we, our, their, themselves, us*
- *sea, sailed, ships, coast, land*
- *people, majesty*
- *have, having*

… *the, great, called, also, part, certaine, leagues, which*

The percentage of nouns among the key-key words declines drastically in comparison to C19, 20 and 21. In the earlier three centuries, there is apparently much more concern about location or orientation (e.g. *side, near, top, along, under*, etc). And, to underline a finding from the key words, travelling by ship, with its accompanying focus on the sea, the wind, land, coast and cardinal directions, is ubiquitous. The focus is on reaching destinations, which makes the actual travelling process itself very newsworthy. There is further indication of descriptive language use, which makes C21 the exceptional data collection among all travel writing of the corpus.

Associates

In an earlier study of the C21 subcorpus (Gerbig 2008), I found significant attitudinal differences in the references to travelling people as either *tourists*, *travellers* or *backpackers*. This difference became visible through collocational choices and their accompanying semantic prosodies. The "associates-procedure" does not pick up on the attitudinal differences, but rather shows the common ground covered by those three groups of people, concerned with basically the same activity, namely travelling. Key words, key-key words and associates are often nouns. Scott & Tribble (2006: 70ff) show that around 50% of all key words of 1000 texts from the BNC, spoken and written, are nouns. Counted together with proper nouns, this percentage rises to nearly 70. Attitude around a noun (tourist, traveller, backpacker) is rather unlikely to be expressed through other nouns as they cover the more topical side of the text.

In the data below, I marked the associates that are not shared by all three nodes in italics. Backpackers apparently are not interested in tours, as that implies coming back to a destination. And tourists seem to be interested more in beaches, which epitomize the tourists' wish for relaxation. Interestingly, there is no further reference to travellers and backpackers themselves as associates. Only tourist/s appear as associates to themselves and to backpackers. Backpackers seem to be slightly more concerned with time than tourists or travellers.

C21 associates of *tourists*, first 20
All occurring five times as associates of the key-key word *tourists*, in all texts in the subcorpus.

– ride, road, trip, driver, *beach*, tour, walking
– guy, guys
– off, to, *around*, next
– tourists, *tourists*
– *us*, my
… few, day

C21 associates of *travel(l)ers,* first 20
All occurring four times as associates of the key-key word *travel(l)ers*, in all texts in the subcorpus.

– ride, trip, driver, tour, walking, *walk*
– guy, guys
– my, *me*, *I*, us
– next, *out*, off, up
… few, locals, day, *water*

C21 associates of *backpackers*, first 20
All occurring four times as associates of the key-key word *backpackers*, in all texts in the subcorpus.

– ride, road, trip, driver, *guide, town*
– guy, guys
– *night*, day, *hours*
– locals, tourists
– my, *everyone*
– next, off, up
… few, *food*,

Key words are a result of relative frequency, specific to a particular text or text collection. Their occurrence provides for lexical cohesion in the texts. They are important starting points for a socio-cultural approach to text analysis. Key words essentially give topical information, as nouns tend to show up in the procedure more frequently than other word classes. Although, as we have seen, this is not true for the data from C16/17/18. There seems to be a more general stylistic change over time, at least in the present genre.

As the key word procedure relies on the recurrence of exactly the same word forms, we are likely to miss themes and topics that are expressed variably by using synonyms or expressions that are semantically related.

In the present data, the key-key word procedure further narrows down the topical focus. By concentrating on only those elements that are shared by all texts, we can reliably perceive lexical specifics of the genre. These key-key words given occur in 100% of the texts, distributed across the century. Apparently, they indicate important propositional and functional aspects of the genre.

Key words are the nodes of important extended lexical units bound to the respective contents being represented. Individual words alone do not make meaning, however. Naturally, after all the word crunching, we want to look at longer contexts, at full units of meaning.

4.3 Key words and their contexts – Extended lexical units

I picked one intuitively interesting key-key word from C17 and looked at its contextual realisation through the centuries covered in the TC. *Abundance* is a key-key word only in C17. For today's language users, the choice of *abundance* appears to be a rather marked way to indicate the idea of 'many', 'a lot' or 'much'. The variation in quantity and structure of the unit of meaning around *abundance* from C17 to C21 shows a change in evaluation.

Based on work started by Sinclair (e.g. 1998, 1999), I will look at concordance data of *abundance* as the node of a unit of meaning by specifying the following parameters:

- Collocation: this is the undirected co-occurrence of the node with neighbouring words / phrases.
- Colligation: this is the co-occurrence of the node with grammatical categories, such as for example prepositions or abstract nouns.
- Semantic preference: those collocates that share semantic meaning are summarized into sets.
- Discourse prosody: it shows the intention of the writer/speaker to use the unit in the first place, i.e. it shows the users' evaluation of what is at hand.

Lexical, syntactic, semantic and pragmatic elements are thus combined in this model of extended lexical units (for a full discussion of the concept of ELUs see Stubbs, this volume). Usually, for each unit of meaning, a canonical form is recognizable, with a fixed lexical core and mostly some systematic variation. Below, I indicate such variation by separating different, recurring element in rounded brackets (…), and optional elements in the structure in squared brackets […].

C 17 – key-key word *abundance* (87 occurrences)
The main structure of this ELU in TC17 is: (70 x)

> *abundance of* [optional classification / adjective] (concrete) noun <

Random selection from 70 occurrences:

1. I passed through **abundance of** villages almost at the End of Every mile,
2. Some green-turtle, a pretty many sharks, and **abundance of** water-snakes
3. all sorts of Turkey Leather: **Abundance of** it is made in Persia, and is exported
4. parlours, drawing-rooms and good stairs, there are **abundance of** Pictures,
5. There are very good monuments and **abundance of** niches in the walls where Statues
6. are chiefly sharks. There are **abundance of** them in this particular sound,
7. blinking creatures (here being also **abundance of** the same kind of flesh-flies
8. and in the bays by the waterside are **abundance of** coconut-trees.
9. but plenty of buffaloes in the woods, and **abundance of** fish in the sea
10. because there is not **abundance of** watery Matter to compose it.
11. serve you there with Coffee, very quick, and with **abundance of** Respect
12. intermixed with **abundance of** waste rocky land, unfit for cultivation
13. neither put it in their Meat nor Drink. but they use **abundance of** it in several
14. unwholsome where it grows because there breed **abundance of** Insects in that muddy

The node is never preceded by an article, as is almost always the case in today's data. There is only one abstract noun in all these examples following abundance of, i.e. Respect in Example 11. The concrete nouns are from a range of semantic fields, often animals, always physically 'graspable' things. Those things or animals are roughly in equal distribution either 'good', such as *coconut-trees, fish, Pictures,* or 'bad', such as *water-snakes, sharks, waste rocky land*, or 'neutral', such as *villages, watery matter, Turkey Leather*. Coconut-trees and fish seem to be inherently good or welcome because they supply food. Pictures (line 4) or niches (line 5) seem to be evaluated positively because they are mentioned alongside other, explicitly positively evaluated related objects, such as *good stairs* or *very good monuments*.

14 x *in abundance*
A discernible structure for this ELU is:

> concrete noun [verb phrase] *in [great/er] abundance* <

Again, the 'graspable' concrete nouns here are from a variety of semantic fields, mostly natural matter, plants and animals. They are often positively evaluated, some neutral and only two are negative.

1. there is Ice **in abundance**
2. limes **in abundance**, pomegranates, pomecitrons, plantains, banan

3. Their chief fruits are (besides plantains **in abundance**) oranges, lemons,
4. common oysters, growing upon rocks **in** great **abundance** and very flat;
5. the Turks eat more nourishing Meat, and **in** greater **abundance**:
6. There is Iron **in Abundance**, but it is not so smooth and tractable as
7. Perfumes consumed **in abundance**; and the Women being thus raised

In all occurrences, *abundance* can obviously be just a vague indication of number, size or scope. It appears to indicate not just 'plenty' of the things mentioned before but it carries an additional evaluation of slight surprise on the part of the writer, implying that he or she is impressed with the – not necessarily expected – number / size / scope of the nouns described.

3 x *abundance*
This bare use without preceding proposition or following *of* does not occur anymore in today's data.

1. sea and rivers have plenty of fish; we saw **abundance**, though we caught but few
2. but those are coarse and low priced; but **abundance** are there vended;
3. As to their jewels, the Men wear **abundance** upon their Fingers,

In approximately half a million words in this C17 subcorpus of the travel corpus, there is a total of 87 occurrences of the node *abundance* (i.e. 0,16 per 1000 running words). In comparison, in two million words of contemporary, mixed written data, there are only 8 occurrences (i.e. 0,004 per 1000 running words). In these contemporary data, *abundance* occurs in the structures: 5x *in abundance*, 2x *abundance* of preceded by an article, and one saying: *the good old days of abundance,* which is inherently a positive thing.

In the TC C21 subcorpus, there are just 2 occurrences of *in abundance.* Interestingly, in the first example, the node is preceded by a series of verbs. Evaluations appear to be positive to neutral.

1. else to do on a dive boat except eat, sleep and dive and we did all **in** great **abundance**.
2. Except surfers and backpackers, which you can find **in abundance**, there are black locals

C20 (7 occurrences)
In C20, there are 2 occurrences of *abundance of*, always preceded by an article. The concrete nouns following carry either a more positive or neutral evaluation. The concrete, inanimate nouns preceding the structure >*in [greater/er] abundance*< (5x) carry a positive to neutral evaluation.

1. a lovely young woman in a blue robe with *an* **abundance of** loose golden-red hair
2. It's clear from *the* **abundance of** statues outside the town hall that Manchester

C19 (37 occurrences)
In C19, again most occurrences are in the structure:

20 x > *(the) abundance of* [adjective] (concrete) noun phrase <

The phrase *abundance of* is only 5 times preceded by an article. The majority of the nouns are concrete. There are however, three more abstract nouns, i.e.: *amusement, entertainment, stately reception.*

There are further 15 occurrences of the structure

15 x > (concrete noun phrase) [verb phrase] *in [great/er / sufficient] abundance* <

These nouns are mostly food, animals, crop and other concrete items of provision. Altogether, in both structures there are just three negatively evaluated occurrences, the majority is positively evaluated, mostly very much so.

The two bare uses of *abundance* can be seen in the examples below:

1. cultivated in considerable quantities around the city: an abundance is grown in the lands of the Gallas.
2. lives are alternate changes from the extremity of want to abundance.

C18 (34 occurrences)
The node *abundance* occurs here as:

- 18 x *abundance of*, only once preceded by an article. The discourse prosody is 9 x positive, 8 x negative, 1 x neutral.
- 14 x *in [great] abundance*. 5 x positive discourse prosody, 2 x negative, 7 x neutral
- 2 x bare use of *abundance*, 2 x neutral discourse prosody

Of all the nouns, only four are of an abstract kind; they are *spirit*, *revelling*, *trouble* and *time*. The other nouns again are concrete, physical things, often animals, vegetation, houses, objects of art and people.

Here are a few selected examples:

1. Raasay has wild fowl **in abundance**, but neither deer, hares, nor rabbits.
2. but I never received any satisfaction, and have lost **abundance of** time.
3. when Nature is pouring her **abundance** into everyone's lap and every eye is lifted

C16 (68 occurrences)
In the 16th century we have 48 x the structure

> (adjective) *abundance of* [(concrete) noun] <

Of these, 24 occurrences are positively and 16 negatively evaluated, 8 have a neutral discourse prosody. On eight occasions, an article is used before *abundance of*. The concrete nouns are plants, animals, food, grain, valuables such as gold and silk, and threats such as snow, hail or ashes.

There are 16 occurrences of the structure

> [concrete noun phrase] *in great abundance* <

Of these, 10 are with positive evaluation and three each with negative and neutral evaluation. The nouns again refer to commodities and food, such as animals and grains, but there are also two occurrences of graves and pictures, and also threats such as wind and rain.

There are four bare uses of *abundance*, one with a positive and three with a negative evaluation.

Here are three illustrative occurrences:

1. Upon this sight, and for the **abundance of** gold which he saw in the city, the
2. rains came down in terrible showers, and gusts **in** great **abundance**; and
3. they permit us to take of their things, such whereof they have **abundance** in their

The data presented here can of course only give an indication as to the tendencies of use around the node *abundance*, relying on merely half a million words per century. Such a tendency, however, seems clear. In summary, the table below shows the sharp decline of the use of the node *abundance* from the earlier to the later centuries. There is also a clear development in distribution from a preference of the structure *abundance of* to *in abundance* from C16 to C21. In total, negative evaluations decrease only slightly from C16 to C18, but after that, from C19 to C21, they disappear. *Abundance* has an inherently descriptive aspect; therefore it is overall more frequent in travel literature than in other text types that are not as much concerned with describing situations. ELUs are evaluative units capturing cognitive schemata which in turn give evidence of culturally shared knowledge.

The following table summarises the statistics around the node *abundance* in the three structures, as explained above, with evaluation. The last column summarizes the total values of evaluation per century for all three structures together.

The obvious benefit of the key word procedure is that it shows in a principled way lexical differences to larger collections of mixed text (as I used it here). This indication of text-related, or even text-type related special-ness can be the starting point for more detailed investigations of individual items found. A complete analysis of all key words would probably show, beyond issues of content, interesting socio-cultural and group-specific particularities. The present examples were, however, selected intuitively. You have to start somewhere to check the potential of a method first.

Table 1. Summary for *abundance*

abundance of			***in abundance***			**bare use of *abundance***			**total evaluation**
C16									
70,5%			23,5%			6%			pos 52%
pos	neg	neut	pos	neg	neut	pos	neg	neut	neg 32%
50%	33,5%	16,5	62%	19%	19%	25%	75%	0%	neut 16%
C17									
80,5%			16,5%			3,5%			pos 40%
pos	neg	neut	pos	neg	neut	pos	neg	neut	neg 29%
34%	33%	33%	64%	14%	22%	66,5%	0%	33,5%	neut 31%
C18									
53%			41%			6%			pos 42%
pos	neg	neut	pos	neg	neut	pos	neg	neut	neg 29%
50%	44,5%	5,5%	36%	14%	50%	0%	0%	100%	neut 29%
C19									
54%			40,5%			5,5%			pos 70%
pos	neg	neut	pos	neg	neut	Pos	neg	neut	neg 8%
70%	10%	20%	73%	7%	20%	50%	0%	50%	neut 22%
C20									
29%			71%			0%			pos 57%
pos	neg	neut	pos	neg	neut	pos	neg	neut	neg 0%
50%	0%	50%	60%	0%	30%	0%	0%	0%	neut 43%
C21									
0%			100%			0%			pos 50%
Pos	neg	neut	pos	neg	neut	pos	neg	neut	neg 0%
0%	0%	0%	50%	0%	50%	0%	0%	0%	neut 50%

Along this line, I happened to note the following use via the node word *would*. I checked the word lists of the subcorpora for absolute frequencies and in the TC C18 there was a sudden increase to 20%. TC C16, C17 and HC EModE show a percentage of only 12 and 13. TC C20 and TC C21 rise further up to 24%. However, *would* is marked as a key word only in TC C18. It is not a key-key word though which shows that idiosyncrasies of one or a few texts can influence values meant to apply to larger data collections.

From the examples below we can find implicit norms of travellers' behaviour in the context of a phrase around *would*. In C18, those who travelled and wrote about it were in a different position in society – privileged and knowledgeable to a degree only few could share. And the travellers obviously wanted others to profit from their experience. Hence the frequency of expressions such as:

(description of state [often abstract N] (, / ;) *I would advise* (N person/s) (action / behaviour)

↓ often negative ↓ travellers ↓ exert caution / don't do / do

C 18 – *I would advise*

1. If these objections were in some measure removed, **I would advise** valetudinarians, who come
2. In such an emergency, **I would advise** the traveller to put up with the four, and he will find
3. turn all your cloths topsy turvy. And here, once for all, **I would advise** every traveller who
4. A man in good health may put up with any thing; but **I would advise** every valetudinarian who
5. as the post-master, whose house **I would advise** all travellers to avoid.
6. consuls at Genoa and Leghorn, a precaution which **I would advise** all travellers to take,

If one starts looking at the absolute frequency of the words within a text or text collection, it becomes clear that the more frequent words are highly frequent because they occur in frequent phrases / n-grams. This is again text specific for those that occur frequently for reasons of topic, but also general for those that take on general functions in a variety of language uses. So, altogether we are able to quickly generate information about individual texts or text collections which would otherwise be very difficult or impossible to attain. We can see differences between text types, in particular if we expand the notion of key words into key phrases.

4.4 Key phrases

I will only very briefly touch upon the potential of combining the analysis of key words with those of phrasal structures that were derived in quantitative terms. If we assume that frequency, relative and absolute, gives an indication of various aspects of noteworthiness, we can naturally accept a concept of key phrases. In another investigation of the travel corpus (Gerbig & Hallan in preparation) we look at the distribution and frequency of n-grams and part-of-speech-grams in more detail. There is interesting diachronic change, partly due to language-internal grammaticalization processes, partly to language-related socio-cultural development. For this study, we use software written by Bill Fletcher (http://pie.usna.edu). Stubbs and Barth (2003) and Stubbs (2004) have shown that the distribution of n-grams can indicate genre specificity, giving information about collections / classes of texts. Recurrent n-grams are not necessarily linguistic units themselves, sometimes just being fragments of

phrases. N-grams are frequent because they capture basic text-structuring functions, indicating temporal, spatial and causal information. As Stubbs (2004) has shown, the most frequent 5-gram in the BNC is *at/by the end of the.*

I will show only the most frequent 5-word phrase-frames over the centuries in the TC. A phrase-frame looks for recurring structures that vary in one slot.

C21

in the middle of * — 64 occurrences, 4 variants
* = (*the* 38, *nowhere* 13, *a* 10, *frikkin* 3)

*in the * of the* — 60 occurrences, 6 variants
(*middle* 38, *back* 6, *shade* 5, *center* 4, *centre* 4, *heart* 3)

*at the * of the* — 48 occurrences, 7 variants
(*end* 16, *front* 7, *top* 7, *foot* 6, *side* 5, *edge* 4, *base* 3)

C20

*at the * of the* — 39 occurrences, 6 variants
(*back* 11, *end* 10, *foot* 8, edge 4, *top* 3, *door* 3)

The most frequent phrases in C21 and C20 are in the structure

>preposition-article-noun-*of*-article<

The noun slot is mostly filled with fairly general spatial terms. This conforms largely to the findings from the BNC, as above.

C 19
Most of the frames fill the main content slots with *degrees*, *minutes*, *seconds*. The others are:

*on the * of the* — 108 occurrences, 12 variants
(*morning* 35, *banks* 26, *part* 8, *borders* 8, *evening* 6, *sides* 5,
surface 5, *summit* 3, *side* 3, *margin* 3, *night* 3, *outside* 3)

C18
Most of the frames fill the main content slot with *latitude*. The others are:

*in the * of the* — 136 occurrences, 15 variants
(*morning* 59, *afternoon* 11, *course* 8, *middle* 8, *evening* 7, *beginning* 6, *bottom* 6,
neighbourhood 6, *garden* 5, *skirts* 4, *midst* 4, *rest* 3, *cool* 3, *entrance* 3, *mouth* 3)

In these two centuries, after the substantial number of topic-related, more or less fixed phrases, many of the phrase frames show the same structure as above in C21 and 20, i.e.

>preposition-article-noun-*of*-article<

The realization of the noun slot changes in interesting ways, however. The spatial references are more concrete, temporal ones increase and there are some metaphorical uses.

C17

*in the middle of ** 37 occurrences, 4 variants
(*the* 23, *ye* 8, *a* 3, *it* 3)

*the * side of the* 31 occurrences, 5 variants
(*north* 9, *west*8, *south* 6, *other* 5, *east* 3)

*on the * side of* 29 occurrences, 5 variants
(*north* 9, *west* 6, *other* 5, *east* 5, *south* 4)

C16

*in the * of the* 65 occurrences, 14 variants
(*midst* 11, *name* 9, *middle* 7, *time* 6, *absence* 5, *spring* 4, *bottom* 4, *presence* 4, *place* 3, middes 3, *dominions* 3, *middest* 3, *mouth* 3)

*at the * of the* 33 occurrences, 8 variants
(*mouth* 7, *discretion* 5, *time* 5, *end* 4, *sign* 3, *beginning* 3, *charge* 3, *request* 3)

C17 is very much concerned with cardinal directions and both C17 and C16 with *the middle* of something which can be spatial or temporal. In C16, we can also see several more fixed phrases with nodes such as e.g. *discretion, charge* or *request*. General aspects of the nouns as today in C21 and C20 were clearly not in place then.

The most frequent phrase-frames are concerned with structural aspects and therefore only rarely overlap with the most frequent key words which are concerned with propositional aspects. Both procedures taken together structure central concerns of the texts. A future aim will be to combine more comprehensively systematic lexico-grammatical and phrasal analyses of texts to be able to better model text-types. Linguistic conventions thus recognized can more easily be interpreted along potential socio-cultural representations.

References

Bergsdorf, W. 1985. Über die Schwierigkeit des politischen Sprechens in der Demokratie. In *Sprachkultur* [Jahrbuch 1984 des Instituts für deutsche Sprache], R. Wimmer (ed.), 184–195. Düsseldorf: Claassen.

Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society*, 36–72.

Gerbig, A. 2003. *Korpus und Kultur: Korpuslinguistische Analysen zu Repräsentationen deutscher und britischer Politik in den Printmedien.* University of Trier.

Gerbig, A. 2008. Travelogues in time and space: A diachronic and intercultural genre study. In *Language, People, Numbers: Corpus Linguistics and Society,* A. Gerbig & O. Mason (eds), 157–176*,* Amsterdam: Rodopi.

Gerbig, A. & Hallan, N. In preparation. Cultural representation and grammaticalization: A diachronic study of phrasal structures.

Hoggart, R. 1957. *The Uses of Literacy*. London: Chatto & Windus.

Klein, J. (ed). 1989. *Politische Semantik. Bedeutungsanalytische und sprachkritische Beiträge zur politischen Sprachverwendung*. Opladen: Westdeutscher Verlag.

Müller-Wood, A. & Gerbig, A. 2006. A literary-linguistic reading of Graham Greene's *Brighton Rock*: Interdisciplinarity in practice. In *How Globalization Affects the Teaching of English: Studying Culture Through Text,* A. Gerbig & A. Müller-Wood (eds), 231–50. Lampeter: Mellen.

Scott, M. 1997a. PC analysis of key words – and key key words. *System* 25(2): 233–245.

Scott, M. 1997b. The right word in the right place: Key word associates in two languages. *AAA – Arbeiten aus Anglistik und Amerikanistik* 22(2): 235–248.

Scott, M. 2000. Reverberations of an Echo. In *PALC'99: Practical Applications in Language Corpora,* B. Lewandowska-Tomaszczyk & P. Melia (eds), 49–65. Frankfurt: Peter Lang.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Sinclair, J. M. 1998. The lexical item. In *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 271], E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.

Sinclair, J. M. 1999. A way with common words. In *Out of Corpora,* H. Hasselgård & S. Oksefjell (eds), 157–79. Amsterdam: Rodopi.

Sinclair, J. M. & Mauranen, A. 2006. *Linear Unit Grammar: Integrating Speech and Writing* [Studies in Corpus Linguistics 25]. Amsterdam: John Benjamins.

Strauß, G., Haß, U. & Harras, G. 1989. Brisante Wörter: *Von Agitation bis Zeitgeist. Ein Lexikon zum öffentlichen Sprachgebrauch*. Berlin: de Gruyter.

Stubbs, M. 2004. On very frequent phrases in English. <http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/icame2004.htm>.

Stubbs, M. 2005. Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature* 14(1): 5–24.

Stubbs, M. 2007. Quantitative data on multi-word sequences in English: The case of the word "world". In *Text, Discourse and Corpora,* M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds), 163–185. London: Continuum.

Stubbs, M. & Barth, I. 2003. Using recurrent phrases as text-type discriminators. *Functions of Language* 10(1): 61–104.

Teubert, W. & Čermacova, A. 2007. *Corpus Linguistics: A Short Introduction*. London: Continuum.

Williams, R. 1983. *Keywords*, 2nd edn. London: Fontana.

Wimmer, R. 1982. Überlegungen zu den Aufgaben und Methoden einer linguistisch begründeten Sprachkritik. In *Holzfeuer im hölzernen Ofen. Aufsätze zur politischen Sprachkritik,* H. Heringer (ed.), 290–313. Tübingen: Narr.

History v. marketing

Keywords as a clue to disciplinary epistemology

Donatella Malavasi and Davide Mazzi
University of Modena and Reggio Emilia, Italy

We present the findings of an introductory study of keywords across two disciplines, i.e. history and marketing, focusing on terms that bring insights into disciplinary epistemology. The study relies on two comparable corpora (2.5 million words each) comprised of history and marketing research articles respectively. We comparatively investigate the collocational patterns of keywords, focussing in particular on one of their most frequent collocates i.e. reporting verbs. Our quantitative and qualitative keyword analysis points to differing collocations for history and marketing, but it validates the assumption that keywords may serve as effective clues to the epistemology of a discipline with a view to its agentive subjects, objects and research procedures defining the construction of knowledge in specific contexts.

1. Introduction

The emergence of a genre-analytic perspective in the 1990s has prominently affected research in academic discourse. According to the ESP approach (cf. Swales 1990, 2004; Bhatia 1993, 2004), growing attention has been paid to text similarities in terms of move structure, communicative purposes, intended audiences, contents and formal properties. Overall, a repertoire of more or less conventionalised written academic genres – viz. the research article, the abstract, the book review and the review article (Swales 1990, 2004; Myers 1992; Bhatia 1993; Motta-Roth 1998; Bondi 1999; Hyland 2000; Stotesbury 2003; Diani 2004) – has been largely accounted for in an attempt to throw light on text, lexico-grammatical and rhetorical regularities.

Nonetheless, the attention paid by genre research to text similarities has incipiently switched towards a closer examination of lexico-grammatical features in terms of keywords. The expression "keyword" has been investigated from a variety of perspectives. For instance, Williams (1976, 1985) concentrates on "cultural

keywords" which constitute a "dictionary" that allows for an exploration of the ideological framework of a specific culture or society. Equally important are the cognitively-oriented studies by Wierzbicka (1992, 1997, 2002, 2003, 2006), who argues that the lexicon of a language is the key to the history, culture and society that produce it, as well as to the cognitive structures of its speakers. From a corpus-based standpoint, Stubbs (2001) has shown the importance of the context of keywords with an emphasis on concordances, collocational patterns and patterns of semantic preference (cf. Sinclair 1991, 1996, 2003).

Drawing an analogy between local culture and disciplinary culture, we will look at how different disciplines offer a representation of their own research activity. Terms which refer to the agentive subjects of the discipline, its focus of interest and its epistemology will thus be considered as keywords for our analysis. Starting from these, the aim of our study is to examine the reporting verbs which collocate with a sample of five keywords within two disciplines, i.e. history and marketing.

The comparative analysis devised in the paper reflects the burgeoning interest in cross-disciplinary studies (cf. Hyland 2000; Dahl 2004; Bamford & Bondi 2005; Hyland & Bondi 2006a) and relies on the assumption that "scholarly discourse is not a single uniform and monolithic entity, differentiated merely by specialist topics and vocabularies" (Hyland & Bondi 2006b: 7). On the contrary, academic writing is strongly influenced by a discipline's own epistemic conventions, ways of constructing, formulating, negotiating, and disseminating knowledge. Accordingly, cross-disciplinary studies have highlighted the "empirical" nature of history which relies on factual reasoning and discussion of events and trends (Bondi 2005; Bondi & Mazzi 2007), and the more theoretical, model-testing, and speculative essence of economics. The prerogatives of the latter also contrast with the model-developing and empirical peculiarities of business/marketing (Hyland 1999, 2000, 2006; Hemais 2001; Bondi 2006).

Set against this theoretical background, our study focuses on two disciplines which are both characterised by chiefly empirical concerns at the outset – the analysis of market trends in marketing, the study of sources and documents in history. However, each discipline may also be said to display a number of traits which differentiate it from the other.

In marketing, on the one hand, the rhetorical articulation of research reports mirror-images the IMRD (Introduction-Methods-Results-Discussion) structure discussed by Swales (1990 and 2004) as prototypical of the genre: as a result, the transition from methodological premises to the presentation of findings is much more neatly perceptible at the level of argumentative and at once textual organisation. In history, on the other hand, it is sometimes hard to draw such clear-cut

boundaries between the various sections of articles: in fact, a peculiarity of the disciplinary transmission of knowledge lies in the constant interplay of narration and argumentation. This is enlightened in a number of studies on the multiplicity of perspectives involved in narration – e.g. the tension between between *histoire* and *discours* (Benveniste 1971), or (in narratological terms) between time of the story and time of discourse (Chatman 1978) – as well as on the argumentative nature of historical discourse, whereby historians strive to make sure that the determination of the importance of every element in historical narratives is safe from arbitrariness (Perelman 1979: 262).

These considerations lead us to hypothesise and then analyse the distinctive language choices which reflect the divergent reasoning patterns of the two disciplines.

On the basis of the existing literature, marketing and history have been separately examined in their peculiar epistemological/methodological configurations but scant comparative research has pointed to their differences and similarities. In this respect, results will not only point to collocational differences between the two disciplines – e.g. the higher frequency of cognition verbs in history as opposed to the overall prevalence of research verbs in marketing. Data will also provide evidence of a divergent attitude to knowledge, with history more narrowly focused on the speculative inquiry of documentary sources, and marketing laying emphasis on the experimental verification, as it were, of empirical hypotheses.

2. Materials and methods

The data for this study consist of two corpora of research articles from the two academic disciplines of business/marketing and history. The marketing corpus consists of 322 texts (2,468,157 words), whereas the history corpus is comprised of 306 texts (2,416,834 tokens): the greatest factor constraining the composition of the two corpora lay in the relatively equal distribution of tokens across disciplinary sub-divisions, in order to construct a comparable composite of each. The amount of data collected – 5 million words altogether – is hardly representative of the two research fields involved: however, the size of the two corpora renders them a manageable resource for a more detailed qualitative study of keywords.

Research articles included in the two corpora were selected as full texts, whereby only footnotes, tables and bibliography have been removed. The articles were drawn from the years 1999 and 2000 from a range of specialised publications: in this respect, even though journals were partly identified through exogenous criteria such as availability in electronic form, recourse was made to

disciplinary experts who suggested a set of reliable publications[1] to choose from. Finally, the journals have been also selected in an attempt to account for the variety of the sub-disciplines which constitute the two macro-fields of enquiry – e.g. management and administrative sciences for marketing, medieval and labour studies for history.

As for methodologies, a frequency wordlist was created for the two corpora by means of the linguistic software package *WordSmith Tools 3.0* (Scott 1998). Secondly, wordlists were compared through WordSmith's keyword function, in order to obtain two keyword lists (history v. marketing and marketing v. history respectively). Thirdly, both keyword lists were processed for the purpose of deriving a sample of five elements per discipline.

Obviously, there is no pre-determined set of key-terms one should take into account in designing the analysis proposed in this paper. In the light of the exploratory nature of the study, however, the investigation was restricted to five items in order to test the explanatory value of the methodological cornerstone advanced here, i.e. the relationship between keywords and epistemological features.

At a preliminary level, keywords were selected by virtue of their hypothesised capability of providing significant evidence as to the subjects, objects and procedures characterising the disciplines. More precisely, our focus was first of all on forms of self-representation, and then on words that, although not immediately related to the object of the disciplines, could be associated with their procedures and epistemology, i.e. the cognitive tools of history and marketing respectively. Categories such as 'objects of the discipline' and 'disciplinary procedures' are inherently conventional, because they might occasionally overlap and they may be difficult to investigate uniformly across disciplines. Nonetheless, their classificatory value and separation, however disputed, were retained for simplicity, since history and marketing would be problematically comparable by nature anyway (cf. Section 1).

Each keyword was concordanced in order to analyse its collocational and inter-collocational patterns. For this stage of the analysis, reference was made

1. The marketing corpus (HEM-Marketing) gathers 322 RAs published in the following journals: *Academy of Management Journal* (AoMJ), *Administrative Science Quarterly* (ASQ), *Business and Society Review* (BaSR), *Business Strategy Review* (BSR), *Journal of Marketing Research* (JoMR), *Journal of World Business* (JoWB), and *Marketing Science* (MS). The history corpus (HEM-History) includes 306 RAs taken from the specialised journals that follow: *Labour History Review* (LHR), *Historical Research* (HR), *Gender & History* (GH), *Journal of European Ideas* (JEI), *Journal of Medieval History* (JMH), *Journal of Interdisciplinary History* (JIH), *Journal of Social History* (JSH), *Studies in History* (SH), *American Quarterly* (AQ), *American Historical Review* (AHR).

to Sinclair's (2004) terminology, whereby collocation is defined as the simple "co-occurrence of words" (2004: 141); colligation as the "co-occurrence of grammatical phenomena" (2004: 142); semantic preference as "the restriction of regular co-occurrence to items which share a semantic feature" (2004: 142); and semantic prosody as a "subtle element of attitudinal, often pragmatic meaning" (2004: 145) words derive from the wider patterns they occur within.

The quantitative and qualitative study of selected keywords in context – particularly through their frequent co-occurrence with different kinds of reporting verbs – was aimed at verifying to what extent the five sample terms enjoy the status of keywords as defined in Section 1. Following Thompson & Ye (1991) and Thomas & Hawes (1994), *verba dicendi* were studied in their three-fold categorisation: research verbs, which may be recognised in statements of findings (e.g. *observe*, *discover*) or procedures (*analyse*, *calculate*); cognition verbs related to mental processes (*believe, conceptualise*); and discourse verbs which imply verbal expression (*ascribe, hypothesize*). In the light of these considerations, analysis was meant to test the intercollocation of collocates and the overall heuristic potential of that concept of keywords in terms of its effectiveness in disclosing deeply rooted disciplinary practices.

3. Results

By applying the criteria formulated in Section 2 to our corpora, we browsed through the two keyword lists resulting from a comparison of history with marketing and vice-versa. A focus on the first 100 keywords of each list drew our attention to five terms per discipline reported in Table 1 below with related frequency and ordinal rank:

Table 1. History (v. Marketing) and Marketing (v. History) keywords with related frequency and rank

History	Frequency	Frequency per 100,000 words	Rank	Marketing	Frequency	Frequency per 100,000 words	Rank
He	9,049	374	2	We	13,260	537	2
Science	1,189	49	20	Results	3,391	137	18
Historians	1,021	42	28	Effects	3,149	128	19
Texts	687	28	47	Research	3,846	156	20
Society	1,322	54	61	Data	3,261	132	24

In this section, we will illustrate the findings of the concordance-based analysis carried out for each of the terms included in Table 1. In particular, 3.1 is devoted to terms that prove to be indicators of self-representation within the disciplines; 3.2 deals with terms whose usage serves to shed light on disciplinary objects; and 3.3 is centred on the analysis of terms which contribute to revealing epistemological procedures informing research in history and marketing.

3.1 Self-representation across disciplines: *We* v. *historians*

In history, *historians* (1,021) is the only form among those in Table 1 that appears to denote the main actors of historical discourse. It is often pre-modified by elements qualifying it quantitatively – e.g. the indefinites *some* (48 occurrences) and *many* (47) – geographically – e.g. *American* (19) – or sub-disciplinarily, as it were – cf. *feminist* (35). But most interestingly, *historians* co-occurs with a high number of reporting verbs, the most frequent of which are *argue* (12), *emphasise* (9), *suggest* (9), *agree* (7), *stress* (6), *question* (6) and *recognise* (5).

A more detailed study of textual instances where *historians* occurs within the collocational patterns described shows that the term is mainly used by RA authors in order to map the reference research field. This tends to be achieved, by laying emphasis on works by discourse community peers, however mentioned in a generalised, indefinite manner as in (1) and (2) below:[2]

(1) These [unspecified before, ndr] historians argue that child-savers positively changed the position of the child within the family and society… (JSH)

(2) Although some historians emphasize that the lowest common denominator is very low […], anthropologists regard European witchcraft as a regional variant… (JIH)

By contrast, marketing researchers more often refer to themselves directly as *we* (13,260). The conspicuous use of the first person plural pronoun gives evidence of the preferred tendency of marketing authors to enter the text rather personally. *We* predominantly combines with a wide repertoire of research verbs, which either describe the findings obtained (e.g. *find, observe, demonstrate* or *establish*), or which denote more pervasively the research procedures followed (viz. *use, examine, test* and *study*, to name but a few). The most frequent verbs which collocate with the pronoun are *use* (622 entries), *find* (498), *examine* (248), and *test* (189). Here is an example:

---

2. In all numbered examples of the paper, emphases, underlinings as well as commentaries in brackets are ours.

(3) We investigate some covariates in our meta-analytic model and find that cross-price effects are greater when there are fewer brands competing in the category. We also find some evidence indicating that cross-effects are greater among brands in nonfood household product categories than among brands in food products. (MS)

Writers depict themselves as pro-active researchers when referring to themselves by the so-called *agentive self* (Dyer & Keller-Cohen 2000: 294). In addition to research verbs, *we* also collocates, even if to a lesser extent, with cognitive or mental verbs such as *expect* (315) exemplified in excerpt (4), *believe* (230) and *assume* (216), as well as discourse verbs like *propose* (115) occurring in instance (5), *focus on* (112), or *present* (110).

(4) We expect that the type and magnitude of the service failure will influence customers' evaluations of a service failure/ recovery encounter because the failure context serves as a reference point from which customers judge the fairness of the encounter. (JoMR)

(5) Thus, we propose that social cognitive theory can help explain and provide insights as to how cognitive processes can/help cope with the hostile environment in entrepreneurial development. (JoWB)

On the basis of the collocations highlighted before, marketing RA authors mainly depict themselves as *researchers*, whereas historians tend to take on the rhetorical role of *arguers* (cf. Fløttum, Kinn & Dahl 2006).

3.2 Objects of the disciplines

The capability of keywords to provide insights as to the objects of historical discourse is testified by the two forms *science* (1,189) and *society* (1,322). *Science*, to begin with, is interestingly preceded by nouns such as *history* and *philosophy*, which contribute to defining a more specific research framework within history as a whole. This trend is exemplified by the pattern *history of science* (91 occurrences).

The most noteworthy aspect disclosed by the concordances of *science*, however, is its presence within wider patterns that point to a common core of historical discourse: the attempt to define and conceptualise what science is, or even better what it was, across centuries, what paradigms it followed and how it changed from one era to another. The most frequent patterns are reported below:

1. …*distinction between science and non-science*… (25 occurrences)
2. 'Concept' + *of* + *science* + *as*… (14)
3. 'Separate' + *science* + *from* + 'non-science' (6)
4. …*definition* + *of* + *science* + *may* (4)

If the first and the fourth string occur invariably as illustrated above in all of their 25 and 4 occurrences respectively, some explanation should be given of the second and the third patterns. Words in single quotes (‘) indicate the semantic element shared by the lexical elements that collocate with *science*. Thus, for instance, a common semantic element of ‘Concept’ appears to underlie lemmas like *understand*, *concept* and *view*, whereas verbal forms such as *free*, *separate* and *distinguish* share a semantic element of ‘Separate’. In (6) and (7), the second and third pattern schematised above are illustrated:

(6) …his attempts to embrace market ideology give way to a definition of science as a discourse of truth that is removed from and above the world of commerce. (AQ)

(7) Pagel […] proposed a deeply historicist approach, rejecting the positivist protocol of separating science from magic or the occult… (SH)

Moving from *science* to *society* as an object of historical research, corpus data reveal that the latter tends to be pre-modified by elements that clarify its profile geographically – e.g. *American society* (31 occurrences) – temporally – cf. *contemporary society* (10) – or culturally/ideologically – see *western* (11), *democratic* (11), *colonial* (9), *communist* (8) *society*. Excluding relatively peculiar occurrences hardly relevant for our research purposes – e.g. *Royal Society* (23) referring to a specific institution rather than society as a sociological pole of historical attraction – two further remarks are worth making.

On the one hand, the construct *society as* * significantly implies evaluative statements (Hunston & Thompson 2000) in 19 of its 39 attested occurrences (55.8 per cent). In these cases, evaluation is mainly expressed in the form of attributions (Sinclair 1986; Hunston 2004) (73.7 per cent). Only minimally is evaluation to be ascribed to averrals and therefore related to the author’s voice (26.4 per cent). In (8) below, an evaluative use of *society as* * embedded in an attributive statement is shown:

(8) Bogdanov represents human society as an integral part of nature, an energetic social process itself subject to self-adjusting natural processes… (SH)

On the other hand, *society* often precedes cognition reporting verbs, i.e. *demand*, *emphasise*, *expect*, *perceive*, *place priority on…*, *think*, *value*. The co-occurrence of *society* with this type of reporting verbs concerns statements in which historians aim at expressing the “voice” of society, as it were, within a specific historical and cultural context. As a result, these are statements where the peculiarities of society are evaluated by historians who attribute society itself an own point of view. This kind of anthropology-driven trend of historical discourse is well described by Tosh (1989), who reconciles it with the approach of historians such as Marc Bloch and his pupil Lucien Febvre.

Essentially, the collocational patterns of *society* lead us to observe that a salient trait of historical discussions is represented by a retrospective evaluation of the distinctive properties of society taken altogether, or *a* society considered within more circumscribed spatio-temporal boundaries. The use of *society* illustrated in the last paragraph is documented in (9) below:

(9) She was forced to do what the family and society perceived as fairly menial work. (GH)

In marketing, effects turn out to be the preferred objects accounted for in the research articles under study. By means of passive constructions (i.e. *effects* + *are/were/have been* + research verbs), marketing authors thematize the *effects* (3,149 entries), and the implications that research has for specific business variables, consumers or, more in general, the wider community. However, *effects* occurs more frequently as the direct object of a wide repertoire of procedural and findings verbs, such as *examine, test, find, show* etc. The collocates of this keyword highlight that the implications of experiments, case studies and tests, which all concern market-related trends and phenomena, are meticulously examined and detailed in the papers under investigation. The centrality of effects in marketing research is exemplified by the following two instances:

(10) To do this, we reexamined the strongest effects found on CEO salary and bonus, those of 1992 PMT focus and exposure and TQM implementation. (ASQ)

(11) First, advertising effects are measured at the quarterly level, while pricing/promotion effects are weekly. (MS)

The essence of *effects* can be better grasped if some colligational patterns are delved into. First of all, the dilemma "the effects of what on what" can be unravelled through the inspection of the co-occurrence of the search word with two prepositions: *of* (1,183 entries) and *on* (280). A host of diversified variables (e.g. *product modification, consumer motivation, pricing policy, promotions, multinationality, strategy* in ex. 12 etc.) is observed and estimated by marketing researchers in the implications they produce on products and firms (e.g. *distribution, reputation, profit, profitability, performance* cf. instance 12 etc.), people (e.g. *person's desire, community support, consumers' welfare, reactions, choice* etc.), and, more extensively, on countries (e.g. *the United States, the country's export performance*).

(12) In the case of firms whose foreign market focus was primarily other developing markets, we found stronger effects of a differentiation strategy on performance. (AoMJ)

Secondly, effects, as the focus of marketing studies, are not merely presented as the outcome of mathematical reasoning procedures or empirical analyses; they are also repeatedly assessed by RA writers in positive v. negative terms – i.e. in the Value dimension of Evaluation (see Hunston & Thompson 2000). The dichotomy between 'good' v. 'bad' effects manifests itself in the see-sawing between evaluatively positive adjectives such as *significant, strong, important*, *consistent,* and their antonyms *negative, insignificant, deleterious, contradictory*.

(13) These studies show a variety of contradictory effects, including a negative relationship between a firm's experience and local equity ownership levels (Davidson 1980; Gatignon & Anderson 1988; Johanson & Vahlne 1977), a positive relationship between the two (Davidson & McFetridge 1985; Stopford & Wells 1972), and curvilinear effects of experience on ownership (Erramilli 1991). (AoMJ)

3.3 Disciplinary procedures

HEM-History data highlight that both *texts* (687) and *he* (9,049) point to central features of history research procedures. *Texts* is frequently preceded by elements that share a semantic element of, say, 'document typology/source'. This can be noted through collocates like *written* (21 occurrences), *literary* (10), *medieval* (9), *collected from…* (7), and *authentic* (5), which relevantly show historians' care and rigour in specifying the nature of documentary materials they avail themselves of in their research.

In addition, *texts* collocates with reporting verbs, with a preference for research verbs bearing on results – i.e. *reveal* (3), *indicate* (2), *convey*, *depict*, *illustrate*, *present*, *show*, *specify* – and discourse verbs – notably *narrate* (2), *focus* (3), *affirm*, *deliver*, *say*, *speak*, *tell*, *testify*. Reporting verbs act here as linguistic tools through which historians attribute an authoritative voice to a text, in order to validate their hypotheses and provide their argument with strength and scientific worth. Example (14) shows how textual usage reflects this function:

(14) Instead such texts reveal that the milieus of Francia and eighth-century Northumbria bore strikingly similar characteristics. (JMH)

It is highly significant that the authoritativeness surrounding texts as the primary documentary source of historical evidence is confirmed by the pattern *texts as* *, which collocates with positively evaluative lexical elements in 8/19 occurrences (42.1 per cent). This confers an overall positive semantic prosody to *texts*: as a result, the analysis of this term signals the importance of text in history research procedures, where its key-role is to increase the solidity and scientific credibility of the scholar's argument. In (15) below, the pattern *texts as* * is shown in use:

(15) Boas fashioned these texts as authoritative remnants of a distanced, bounded and disappearing world of tradition. (AQ)

To put it in Febvre's words, documents are for historians what flowers are for bees, namely raw materials from which to derive the skilfully crafted end-product that matters most for them: historical truth – however construed and controversial it may be (Perelman 1979; Antiseri 2005) – and honey respectively. And indeed, there appears to be a correspondence between the textual usage retrieved for *texts* and their status of intermediaries between the basic historical operation of "letting the audience know what happened" and the crucial stage of "getting the audience to believe" the narrative shaped by historians. In this respect, Lozano (1991) elegantly argues that the passage from the former to the latter is only warranted by argumentation based on supporting documents which rigorously certify the acceptability of historical interpretations.

The position of *he* as an epistemological signal of historical research may sound more puzzling. In fact, corpus evidence underscores a link between *he* and *texts*. The in-depth study of 100-occurrence random sample suggests that *he* principally refers to historical characters (e.g. *Kant*, *Churchill* etc.) (92 per cent). Only in 6 per cent of those occurrences has *he* the name of another historian, i.e. a member of the discourse community, as a co-referring item; in 2 per cent, finally, the referent of *he* is a non-identifiable person, as in the hypothetical pattern *Let us assume that…*, which can be associated with any individual.

Moreover, the most recurrent collocates of *he* are reporting verbs that denote discourse acts – *write* (203), *say* (195), *claim* (74), *tell* (61), *state* (60), *conclude*, *report* (36), *explain*, *propose* (33), *add*, *declare* (26). The particularly high frequency of *write* substantiates an interpretation according to which historians very often use the pronoun *he* in order to cede the floor to one of their reference texts, a documentary source that speaks straight through the author's own voice evoked by *he* as a signal of explicit attribution. Hence the link between the pronoun and *texts* announced earlier on, with the effect that historians quote both texts and their authors in order to support their arguments in an equally convincing way, albeit with a slight stylistic variation. In (16) and (17), the key-function of *he* emphasised here is further clarified:

(16) [George Orwell] He wrote at the time that it "would be absurd to imagine that Britain is on the verge of violent revolution, or even that the masses have been definitely converted to Socialism. (LHR)

(17) [Cardinal Jean Lemoine] He says that St Bernard alleged that people in his day came to Rome more out of *ambitio*, than *devotio*. (JMH)

The specificity of history can be even more neatly perceived by way of comparison with marketing. The empirical research-based gnoseology of the latter is signalled by the emphasis placed on the keywords *research* (3,846), *results* (3,391) and *data* (3,261). The high co-occurrence of these search words with discourse verbs (e.g. *suggest, support, focus on, confirm*) suggests that evidence and truth in marketing derive from empirical sources, exploratory studies and their upshots. In particular, scientific and data-based *research,* which collocates predominantly with the tentative verb *suggest* (cf. ex. 18), turns out to be the source of evidence of the discipline and the cornerstone of subsequent developments and 'discoveries' in this field of enquiry.

(18) However, compared with Study l, the attention duration in Study 2 was substantially shorter, and recent research suggests that when consumers spend less time on commercial stimuli, attention to the textual information suffers most (Pieters & Warlop 1999). (JoMR)

Existing studies form the basis of following investigations and experiments: researchers rely on *previous research* (in example 19) in order to carry out their own analysis, which is presented as either confirming or contradicting prior findings (cf. ex. 19). The interrelation existing between past, present and future studies is evidenced by the co-occurrence of *research* with temporal adjectives such as *previous* (181), *prior* (178), and *past* (29); *recent* (52), *current* (31), and *present* (26); *future* (268), *further* (167) and *additional* (42).

(19) Whereas previous research has measured average response times within action and reaction dyads in a given year (cf. Smith et al. 1991), we, in contrast, accounted for cases in which a firm carried out two or more successive moves before the alternate firm (challenger or leader) undertook an action (cf. Young, Smith, & Grimm 1997). (AoMJ)

Research is evaluated not simply on a temporal scale, it is also classified according to its sub-disciplinary nature and judged in terms of how 'good' or 'bad' it is. On the whole, we find that *research* is labelled as *empirical* (86) v. *theoretical* (6), *scientific/experimental* (4 and 6) v. *interpretative* (5), *quantitative* (4) v. *qualitative* (17). *Research* is also categorised (even if less pervasively) with respect to its essence and the field of enquiry it is related to. Different typologies of *research* are hinted at by attributes such as *cross-cultural, longitudinal, historical, psychological…* Finally, *research* is assessed by RA authors, who enter the text rather explicitly and who evaluate it as positive and negative, i.e. as *important* (ex. 21), *interesting, original* etc.

(20) Recent empirical research has suggested that there may be a positive link between the overall quality of management in a firm and its social performance, defined in terms of stakeholder relationships.(n 6) (BASR)

(21) Important research is currently being conducted investigating cultural characteristics, such as universalism-particularism, that were first identified by the anthropologists Kluckhohn & Strodtbeck (1961; cited in Maznevski & DiStefano [1998]). (AoMJ)

The prominent co-occurrence of *research* with *empirical* unravels one key prerogative of marketing epistemology, i.e. its experimental and exploratory nature. This is further confirmed by the substantiality of *data* and *results* in the discipline.

Data represent the hub around which marketing empirical reasoning revolves. *Data* can be defined as the key source of evidence in the discipline as testified to by the high frequency of occurrence of the lemma throughout the research articles under study. Researchers rely on data in order to validate theories, challenge models, generate hypotheses and support intuitions. Therefore, in some cases *data* is combined with verbs expressing tentativeness (such as *suggest* and *may* in ex. 22), whereas in some others it collocates with more certainty-related argumentative verbs, like *support* in excerpt (23).

(22) This suggests that panel data may be able to provide key insights into the effectiveness of marketing activity despite the sampling issues. (MS)

(23) The data support all the main hypotheses about viewing preferences presented previously. (JoMR)

In addition to *data, results* is also a keyword of the distinctive epistemology and gnoseology of marketing. The collocates of *results* clarify their nature as the outcome of experiments and as the source of evidence which enables researchers to confirm theories, contradict assumptions, or *suggest* new models. Accordingly, the most frequent verbs co-occurring with *results* are discourse verbs such as *suggest* (140), *support* (82), *provide evidence/support* (32), *present* (33), and *confirm* (13).

Apart from *verba dicendi, results* is pervasively pre-modified by adjectives which describe the upshots of studies in neutral terms as *empirical* (52) rather than *theoretical* (5), or evaluatively as *significant* (7), *positive* (6), *key* (6), *inconclusive* (6), *inconsistent* (5), *conflicting* (5), *detailed* (4), or *interesting* (4).

(24) Empirical results have confirmed that firm performance is mainly determined at the business-unit level and not at the corporate level. (AoMJ)

4. Conclusions

The results presented in the paper afford valuable insights into the cross-disciplinary study of keywords. In the first place, data confirm the centrality of keywords in the characterisation of disciplines: *historians* and *we* acted as useful indicators of self-representation in history and marketing respectively; *science* and *society* projected us into the objects of history, whereas *effects* proved a significant tool to access the objects of marketing studies; similarly, *texts* and *he* (for history), *research*, *results* and *data* (for marketing) cast light on some of the peculiarities of the respective research procedures.

Secondly, the concordance-based study of the collocational patterns of each word showed a considerable inter-collocability between selected keywords and reporting verbs. In particular, there seems to be a preference for cognition verbs in history – especially with regard to disciplinary subjects and objects – whereas in marketing, research verbs prevail, above all in the textual projection of disciplinary procedures.

Finally, analysis led us to note a number of distinctive traits of the two disciplines considered. On the one hand, for instance, the statistical incidence of *historians* showed how important it is for the historical discourse community to map the research territory as they construct a scientifically viable narrative, whereas the prominence of *we* pointed to the central role that scholars undertaking the investigation of market phenomena and tendencies ascribe to themselves as proactive scholars in the field. On the other hand, the analysis of the term *effects* provides good evidence that marketing essentially concentrates on the impact of research on business variables and people in the short, middle and long term, whereas history is more interested in drawing up a retrospective meditation about macro-entities such as *science* and *society* across centuries.

Obviously enough, the keywords of history and marketing are not limited to the 5-element sample collected and analysed in the paper. Nonetheless, the corpus-based findings documented here reasonably suggest that keywords may serve as effective interpretive clues to specific disciplinary peculiarities. It can be hoped that further cross-disciplinary research – e.g. comparing marketing with economics or history with art history, and centred on a wider range of keywords with respect to that adopted in this paper – will provide more substance to our view of keywords as the yardstick by which to evaluate not only cultural aspects (cf. Wierzbicka 2006), but also disciplinary epistemology at large.

References

Antiseri, D. 2005. *Introduzione alla Metodologia della Ricerca*. Soveria Mannelli: Rubbettino.

Bamford, J. & Bondi, M. (eds). 2005. *Dialogue within Discourse Communities: Metadiscursive Perspectives on Academic Genres.* Tübingen: Niemeyer.

Benveniste, E. 1971. Le relazioni di tempo nel verbo francese. In *Problemi di linguistica generale*, E. Benveniste, 283–297. Milano: Il Saggiatore.

Bhatia, V. K. 1993. *Analysing Genre. Language Use in Professional Settings.* London: Longman.

Bhatia, V. K. 2004. *Worlds of Written Discourse. A Genre-Based View.* London: Continuum

Bondi, M. 1999. *English across Genres: Language Variation in the Discourse of Economics.* Modena: Il Fiorino.

Bondi, M. 2005. Metadiscursive practices in academic discourse: Variation across genres and disciplines. In *Dialogue within Discourse Communities: Metadiscursive Perspectives on Academic Genres,* J. Bamford & M. Bondi (eds), 3–29. Tübingen: Niemeyer

Bondi, M. 2006. 'A case in point': Signals of narrative development in business and economics. In *Academic Discourse across Disciplines*, K. Hyland & M. Bondi (eds), 49–74. Bern: Peter Lang.

Bondi, M. & Mazzi, D. 2007. The future in history: Projecting expectations in historical discourse. In *Linguistica, linguaggi specialistici, didattica delle lingue. Studi in onore di Leo Schena,* G. Garzone & R. Salvi (eds), 85–93. Roma: CISU.

Chatman, S. 1978. *Story and Discourse*. Ithaca NY: Cornell University Press.

Dahl, T. 2004. Textual metadiscourse in research articles: A marker of national culture or of academic culture? *Journal of Pragmatics* 36: 1807–1825.

Diani, G. 2004. A genre-based approach to analysing academic review articles. In *Academic Discourse, Genre and Small Corpora*, M. Bondi, L. Gavioli & M. Silver (eds), 105–126. Roma: Officina.

Dyer, J. & Keller-Cohen, D. 2000. The discursive construction of professional self through narratives of personal experience. *Discourse Studies* 2(3): 283–304.

Fløttum, K., Kinn, T. & Dahl, T. 2006. „We now report on ..." versus „Let us now see how ...": Author roles and interaction with readers in research articles. In *Academic Discourse across Disciplines*, K. Hyland & M. Bondi (eds), 203–224. Bern: Peter Lang.

Hemais, B. 2001. The discourse of research and practice in marketing journals. *English for Specific Purposes* 20(1): 39–59.

Hunston, S. & Thompson, G. (eds). 2000. *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: OUP.

Hunston, S. 2004. 'It has rightly been pointed out that…': Attribution, consensus and conflict in academic discourse. In *Academic Discourse, Genre and Small Corpora*, M. Bondi, L. Gavioli & M. Silver (eds), 15–33. Roma: Officina.

Hyland, K. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3): 341–367.

Hyland, K. 2000. *Disciplinary Discourses. Social Interaction in Academic Writing*. Harlow: Longman.

Hyland, K. 2006. Disciplinary differences: Language variation in academic discourses. In *Academic Discourse across Disciplines*, K. Hyland & M. Bondi (eds), 17–45. Bern: Peter Lang.

Hyland, K. & Bondi, M. (eds). 2006a. *Academic Discourse across Disciplines*. Bern: Peter Lang.

Hyland, K. & Bondi, M. 2006b. Introduction. In *Academic Discourse across Disciplines,* K. Hyland & M. Bondi (eds), 7–13. Bern: Peter Lang.
Lozano, J. 1991. *Il Discorso Storico*. Palermo: Sellerio.
Motta-Roth, D. 1998. Discourse analysis and academic book reviews: A study of text and disciplinary cultures. In *Genre Studies in English for Academic Purposes*, I. Fortanet, S. Posteguillo, J. C. Palmer & J. F. Coll (eds), 29–58. Castelló: Universitat Jaume I.
Myers, G. 1992. Textbooks and the sociology of scientific knowledge. *English for Specific Purposes* 11(1): 3–17.
Perelman, C. 1979. *Il Campo dell'Argomentazione. Nuova Retorica e Scienze Umane*. Parma: Pratiche.
Scott, M. 1998. *WordSmith Tools,* Version 3.0. Oxford: OUP.
Sinclair, J. 1986. Fictional Worlds. In *Talking about Text*, M. Coulthard (ed.), 43–60. Birmingham: University of Birmingham.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
Sinclair, J. 1996. The search for units of meaning. *Textus* 9(1): 75–106.
Sinclair, J. 2003. *Reading Concordances*. London: Longman.
Sinclair, J. 2004. *Trust the Text. Language, Corpus and Discourse*. London: Routledge.
Stotesbury, H. 2003. Evaluation in research article abstracts in the narrative and hard sciences. *English for Academic Purposes* 2: 327–341.
Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Swales, J. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: CUP.
Swales, J. 2004. *Research Genres: Explorations and applications.* Cambridge: CUP.
Thomas, S. & Hawes, T. P. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes* 13(2): 129–148.
Thompson, G. & Ye, Y. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics* 12(4): 365–382.
Tosh, J. 1989. *Introduzione alla Ricerca Storica*. Scandicci: La Nuova Italia.
Wierzbicka, A. 1992. *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations*. Oxford: OUP.
Wierzbicka, A. 1997. *Understanding Cultures through their Key Words*. Oxford: OUP.
Wierzbicka, A. 2002. 'Right' and 'wrong'. From philosophy to everyday discourse. *Discourse Studies* 4(2): 225–252.
Wierzbicka, A. 2003. *Cross-cultural Pragmatics. The Semantics of Human Interaction*, 2nd edn. Berlin: de Gruyter.
Wierzbicka, A. 2006. *English. Meaning and Culture*. Oxford: OUP.
Williams, R. 1976. *Keywords: A Vocabulary of Culture and Society*. London: Fontana Press.
Williams, R. 1985. *Keywords: A Vocabulary of Culture and Society*, rev. edn. Oxford: OUP.

Metaphorical keyness in specialised corpora

Gill Philip
University of Bologna, Italy

What is more important in text: the topical content, or the manner in which topical content is presented? While statistically-generated key words tell us about a text's content, the inter-relation between these words and the message of the text can be difficult to ascertain. One method of doing so is to observe the inter-relation of key words with evaluative language: in this case, metaphor. Metaphors are notoriously difficult to locate in corpora, but this paper sets out a method for their semi-automatic identification, and demonstrates how their interaction with keywords is both systematic and pervasive. Studying the interaction of key words and metaphors brings to light attitudes which lurk beneath the surface of text.

1. Introduction

The function of a key is to open locks; and locks are used to ensure the safekeeping of those things we consider valuable, precious, or important. The identification of key words in a text provides the reader with the opportunity to gain direct, unfettered access to its content, circumventing the need to pick the lock (by reading the full text in detail, then analysing its contents) or peer through the keyhole (by making generalisations based on its most striking features). The identification of key words in a specialised corpus reveals textual information on a larger scale, highlighting the matters which the overall discourse is concerned with. Yet while key words tell us *what* is important, they do not tell us *why*.

It is not necessary for a word to appear at the top of the word-frequency list for it to qualify as key, just as not all high-frequency vocabulary necessarily has key word status (see Appendix 1). However, it must recur: a discourse cannot hinge upon hapax legomena. For this reason, the calculation of key words is dependent on frequency measures and repetition, yet these matters are not entirely unproblematic. In particular, a language with very few inflected forms has more recurrent forms than a fully inflected one, which in turn has fewer forms

than agglutinative or infixing languages. While each word form attracts its own distinctive patterning, the dispersion of closely-related meanings over variant forms of a lemma may affect frequency measures and statistical calculations[1]. A further matter, and one which is the central concern of this paper, is that there are some kinds of repetition which do not involve replication or reiteration of the same word forms. This is true of semantic relations in particular, and though semantic annotation tools can aid the recognition of related senses (Rayson 2005), they are limited both in terms of availability for languages other than English, and in the types of relation that they can uncover[2]. The grouping of semantically-related words into lexical sets implies recurrence, but not of a sort which can be measured in a straightforward way. Lexical sets can form around a key word, thus acquiring the a status of *key-by-association*, and these add richness to the lexis regarding a text or discourse topic. Many other lexical sets may also occur, but their presence is less easily noticed. These sets lurk in low-frequency vocabulary; they do not seem to be relevant to the aboutness of the discourse, indeed they may seem out of place. These are the groupings that suggest metaphorical activity.

At first glance, the relevance of metaphor to notions of keyness might appear minimal, as keyness is concerned with subject matter, and, as a general rule, metaphors are not informative nor are they central to the transmission of content[3]. What metaphors are used for, however, is the expression of "affect and attitude along with ideational content" (Cameron & Deignan 2006: 676), meaning that metaphors play an important evaluative role in a text or discourse. This is as true of metaphors which are deliberately used for their rhetorical function as it is for language which is less actively metaphorical and not used with deliberate rhetorical intent – conventional and delexical expressions. Both these types of metaphor interact with high-frequency lexis and key words, and they do so in different,

1. In this study, the language of the data is Italian; thus nouns, verbs, and adjectives are inflected. There are also variant forms for articles and determiners, and fused forms for certain pronominal constructions and for article + preposition.

2. If the lexis is related taxonomically, semantic annotation poses few problems; however conventional notions of semantic relatedness are less effective in the treatment of metaphor and other figurative language, where relations tend to be based on attributes and characteristics (Glucksberg & Keysar 1993).

3. Exceptions to this rule include dead metaphors and metaphorically-motivated terminology whose meaning is considered "basic" (or "literal") in the language as a whole, or in a given discourse; and explicative metaphors and analogies whose function is to shed light on new information by comparison with something more familiar.

but complementary ways; however it is the latter that form the main focus of this study. Conventional, delexical metaphors recur almost invisibly over many texts. The reader's attention is rarely drawn to them and yet, at an almost subliminal level, they contribute to the meaning of the high-frequency lexis and key words. If a better understanding of the *whys* of key words is sought, the analysis of recurrent metaphors can be fruitful. This paper discusses the location of metaphors in a corpus of specialised language, and the identification of "metaphor themes" (Black 1993; see 2.1) which run through the discourse. The more pervasive of these – the *key metaphors* – are then studied to determine their evaluative power over the statistically-generated key words in the same data set.

2. Metaphor and textual meaning

2.1 Terminology note

The study of metaphor assumes the knowledge of some core terminology. This paper looks at single instances of metaphors – *linguistic metaphors* (after Steen 1994) in the corpus data, and groupings of these linguistic metaphors into semantically related areas, known in the cognitive linguistics literature as *conceptual metaphors* (Lakoff & Johnson 1980); in this paper I adopt the theory-neutral term *metaphor theme* (after Black 1993).

In all metaphors, one entity, idea, or action is described, defined or explained through another. The relationship between these parts is described for linguistic metaphors in terms of *vehicle*, *target*, and *ground* (after Richards 1936), while for metaphor themes it is described in terms of *source* and *target* domains (after Lakoff & Johnson 1980). The lexical items (in the case of linguistic metaphors) or semantic domains (for metaphor themes) must be distinct, and metaphoricity is created because of the incongruity of their use together. In the linguistic metaphor, *il mondo è […] la nostra ancora di salvezza*[4] [the world is our safety anchor], there is incongruity between *mondo* [world] (the metaphor target), and *ancora di salvezza* [safety anchor] (the metaphor vehicle), and the ways in which the world can be conceptualised in terms of anchors forms the motivation (ground) for the metaphor. In a conceptual metaphor or metaphor theme, the metaphorical activity is no longer tied to specific words, but rather occurs in the abstract: the target domain (realised through a variety of possible words) is expressed in terms of the source

4. Unless otherwise stated, all examples are from the ComInt corpus (see 3.1).

domain (again realised through a variety of possible words)[5]. Metaphor themes are identified in text by grouping together apparently related linguistic metaphors, which may or may not feature the same vocabulary. Figure 1a shows the ECONOMIC PRODUCTIVITY IS A HEALTHY BODY metaphor[6], illustrated by *paziente convalescente ma robusto* (convalescent but strong patient), *diagnosi* (diagnosis) *sano* (healthy) and *stato di salute* (condition of health), while INTERNATIONAL TRADE IS WAR is illustrated in Figure 1b through the lexical items *agguerrita* (ready for war, fierce) *conquista* (conquer), *vincenti* (winning) and *battaglia* (battle).

2.2 Metaphors in text

While literary metaphors are highly visible in text because of their novelty, the same cannot be said of non-literary metaphors. Most metaphors are in fact naturalised lexical items, forming part of the conventional vocabulary and phraseology of the language, and their metaphorical nature tends to pass unnoticed, both in production and in reception. In common with all lexical items, conventional metaphors attract distinct collocational patternings which delineate their meanings (Deignan 1999), and the extended unit of meaning (Sinclair 1996) of a metaphor is no different to any other in that it has semantic preferences and, crucially, semantic prosodies (Philip in press a). These are often invisible but, following a pragmatic approach to metaphor (after Charteris-Black 2004), this invisibility appears to be central to the rhetorical and persuasive use of metaphor in non-literary texts. While Louw (2000) is justified in arguing that the use of conventional metaphor is not indicative of active metaphorical conceptualising, the fact remains that conventional, delexical forms transmit evaluative meaning, irrespective of whether they are consciously produced (or understood) as figures of speech. The repeated use of particular metaphors or metaphor themes, especially when these are used with reference to a limited range of subjects, conveys a sense of the underlying attitudes regarding the subject in hand. While the metaphors are indeed virtually invisible, their presence can be perceived subliminally.

---

5. In a comparable account of metaphor, "above the actual metaphor as a speech act, in our linguistic competence there is an image field as a visual structure." (Weinrich 1967: 283; this translation cited in Jäkel 1999: 18). In this account, the source domain is labelled the image donor field, and the target domain is the image recipient field (Weinrich 1958: 284). However, as the degree to which metaphors are visualised remains subject to individual variation, metaphor theme is considered the most satisfactory term for use in the present study.

6. Metaphor themes are conventionally notated in small capitals. In the examples, the target domain is underlined, and the source is in bold face.

(i) Importanti centri studi hanno fotografato l'Italia economica come un **paziente convalescente ma robusto**.
'Important study centres have pictured economic Italy as a convalescent, but strong, patient.'
(ii) La **diagnosi** che mi sento di condividere è che il sistema produttivo italiano è un sistema **sano**…
'The diagnosis that I feel able to share with you is that the Italian productive system is a healthy one.'
(iii) Il buono **stato di salute** del cinema italiano ha trovato risposte non solo nei paesi da sempre interessati…
The good health of Italian cinema has been noted not only in those countries which have always expressed an interest…'

Figure 1a. ECONOMIC PRODUCTIVITY IS A HEALTHY BODY

(i) Il che significa che, in presenza di una concorrenza **agguerrita**, noi siamo **soccombenti** per motivi di costi…
'Which means that, in the presence of fierce competition, we are the losing party because of prices.'
(ii) Il Made in Italy **conquista** nuovi mercati, Russia e Cina, ma preoccupa flessione in Usa e Giappone.
'The Made in Italy [brand] is conquering new markets, Russia and China, but is suffering a downturn in the USA and Japan.'
(iii) Sui mercati mondiali siamo **vincenti** solo se riusciamo a trasformare la nostra migliore tradizione artigiana…
'We can be winners in international markets only if we manage to transform the best of our tradition of workmanship…'
(iv) …nella **battaglia** per affermare l'Italia sui mercati internazionali.
'…in the battle to assert Italy's place in international markets.'

Figure 1b. (INTERNATIONAL) TRADE IS WAR

Non-literary metaphors in text rarely occur in splendid isolation (Example 1), because they are a core part of the vocabulary of any language. Even in combination, when metaphors come in *clusters* (Cameron & Stelma 2004), their figurative nature tends not to be registered consciously. This is because the lexis is being used in formulaic ways which give the reader no reason to stop and evaluate whether any hidden or extended meanings are inferred (Example 2). The matter is somewhat different when the clustering involves lexis drawn from a single lexical set/conceptual domain which is clearly not congruous with that of the discourse topic (as in Example 3). In these instances the lexis, which is normally interpreted with its delexical or textual meaning, undergoes double processing: the words'

proximity and incongruity with the discourse trigger their reinterpretation as simultaneously literal and figurative, with the result that their metaphoricity is accentuated (Philip in press a). However these clusters, or conceits, are relatively rare (only a dozen instances located in the corpus studied), and their use seems deliberate in all instances.

(1) *È un processo delicato ma importante, tuttora in fase di* ***gestazione****…*
'It is a delicate but important process, still in gestation…'

(2) *…e mi auguro che le scelte che faremo non siano* ***soffocate*** *da* ***miopi*** *atteggiamenti politici di* ***corto respiro****.*
'…and I hope that the choices we make will not be suffocated by short-sighted, short-lived politicising.'

(3) *Il mondo non è il posto dove rischiamo di* ***naufragare****, ma la nostra* ***ancora di salvezza*** *contro i rischi di* ***impaludamento*** *che corriamo se restiamo nei nostri piccoli mercati locali.*
'The world not is the place where we risk getting shipwrecked, but our safety anchor against the risk we run of getting stuck in a marsh if we remain in our small local markets.'

Less deliberate, less noticeable, but perhaps more influential in the long run are metaphor themes. Conceits exert considerable influence within single texts, yet have little or no lasting influence over the discourse as a whole. Instances of metaphor themes, on the other hand, are relatively inconsequential in themselves but rise in importance with each subsequent reappearance in the discourse. The recurrence of particular themes over many texts produced or delivered by the same individual can reveal information concerning that individual's stance on the subject matter. Aspects of the broader socio-political climate in which the texts were produced may also surface, especially if the texts belong to a fixed timeframe. If the author or deliverer of the texts is a public figure, such as a politician, the relevance of stance and evaluation is magnified: thanks to the media's habit of reporting politicians' words verbatim, the phraseology used by one becomes the phraseology used by many, as the underlying implications of the politician's linguistic choices resonate beyond the original text into the language at large. Being conventional and unremarkable, these metaphors operate in silence, yet they help to shape the opinions of millions.

3. Metaphors and corpora

If metaphors have the potential to persuade as well as to evaluate, their presence merits attention. However, due to the fact that instances of metaphor themes are spread over many texts, their identification is problematic as well as time-consuming and laborious. Even bringing the texts together as a corpus of specialised language falls short of successfully addressing the problem, because query software is designed for word searches, not meaning searches.

Metaphor scholars working with corpora (Charteris-Black 2004, 2005; Partington 2003; Koller & Semino 2009) have tended to overcome this problem by engaging in partial or total manual searching of the corpus texts, then using computer search tools at the second stage of analysis. The first stage involves close, word-by-word reading of the text, assessing whether each lexical item is used metaphorically or otherwise[7], in accordance with the established metaphor identification procedure (Pragglejaz Group 2007). The second stage involves calling up concordances of all instances of the identified words and expressions (following the type of analysis proposed by Deignan 1999, 2005: Chapter 4), identification of metaphorical senses, then analysis of these with or without reference back to the extended cotext or full text in which the metaphors occur. The methods used in the present study take a different approach because the aim is not to locate all metaphors present in the corpus but to identify potential metaphorical themes. The data and methods are set out below.

3.1 Corpus composition

The corpus described in this study is one of six corpora compiled for a study of metaphor use by women Ministers in Italy (Philip 2009). The data was compiled from the speeches, communiqués and press releases produced by Emma Bonino, in her dual role as Minister for International Trade and Commerce, and Minister for European Policy (represented by the ComInt and PolEur corpora respectively). The corpus data covers the first year (June 2006 – May 2007) of the then centre-left government and the data was freely available for download from the Ministerial homepage. The corpus contains approximately 140,000 words of running text divided into four subcorpora, the details of which appear in Table 1.

7. This is determined as follows: "If the lexical unit has a more basic current-contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it." (Pragglejaz Group 2007: 3).

Table 1. Corpus composition

		PolEur		ComInt	
	overall	**sp.***	**p.r.**	**sp.**	**p.r.**
no. of texts	203	13	59	55	76
mean text length	688	2180	221	1291	358
tokens in text	139,605	28,344	13,057	71,024	27,180
tokens in word list*	136,788	27,832	12,704	70,155	26,097
types	11,865	4,727	2,704	8,948	3,795

* sp. – speeches and speech notes; p.r. – press releases

3.2 Locating metaphorical lexis

The methodology set out here was devised in order to identify metaphors in the corpus data without reading through and annotating the texts beforehand. Although the manual analysis of a small (<150,000 words) corpus is feasible, it is time-consuming and prey to inconsistencies, omission and misclassifications due to human error. A need therefore arises for a semi-automatic procedure for identifying metaphors, which eliminates the need to analyse a core of the data beforehand, and leads directly into concordance analysis. The procedure described was carried out using WordSmith Tools version 4 (Scott 2004), but can be carried out with any concordance package which generates word frequency and key word lists.

The most important feature of metaphor, as noted in 2, is that it requires the presence of contextually incongruous lexis, and this central fact about metaphor informs the method used in this study. If metaphorical lexis is not central to the topic of the discourse, it will not occur near the top of a word-frequency list, nor will it feature amongst the statistically-generated key words, but will stand in contrast to it. With this knowledge as a starting point, a word frequency list was created for the ComInt corpus, and a key word list was generated using the word frequencies of the entire political corpus data (430,000 words) as the control. The key words were thus identified with reference to the political discourse of the time-frame in which the data was collected, not to the language overall. This made it possible to identify the topic-related lexical areas of the specialised corpus, while eliminating lexis common to political language in general, such as 'Minister', 'country', 'government'; these appear high up on the frequency list, but are not key (see Appendix 1). The key words were lemmatised where appropriate, then grouped into semantic sets in order to establish the criterion of incongruity, necessary for metaphor identification. The groups identified were: markets and

sectors; business and industry; the economy; import and export, foreign countries and internationalisation; legislation; and development (see Appendix 2).

Some metaphor themes are known to be typical of certain discourses. In business and economics, these include orientation metaphors (ups and downs), metaphors of growth, relationships, and water, amongst others. In the key word list, two such metaphors occurred, namely *crescita* [growth] and *flussi* [flow]. Both also occurred as high-frequency lexis, suggesting that they are not metaphors but rather metaphorically-motivated terminology. As metaphors are created by contrasting basic (literal, terminological, salient) meanings with an incongruous domain, it is highly unlikely that a word form will be used both figuratively and literally within the same text or discourse. No near-synonyms were found for *crescita* and *flussi*, confirming the hypothesis that these metaphors have terminological status. The absence of related forms makes it impossible to ascertain whether terminological status is limited to these lemmas or whether it also extends to the semantic domain to which each belongs. Both *crescita* and *flussi* were excluded from consideration as possible metaphors themes in this corpus.

While high-frequency content words sum up the aboutness of the data, it was hypothesised that low-frequency content words (LFCWs) would provide a source of metaphorically-used lexis. There are two obvious problems with this observation. The first is that LFCWs make up a very large proportion of any corpus – in ComInt, hapax legomena alone account for over 40% of all word forms (see Table 2). The second problem is that not all low-frequency lexis is metaphorical.

Table 2. Details of lowest-frequency word forms (bottom 15% of tokens)

no. *	rank range	types		tokens	
		total	%	total	%
1	11,865–6805	5060	42.65	5060	3.70
2	6804–3679	3125	26.35	6250	4.55
3	3678–2854	753	6.35	2259	1.65
4	2853–2340	513	4.30	2052	1.50
5	2339–1998	341	2.90	1705	1.25
6	1997–1716	281	2.40	1686	1.25
7	1715–1506	209	1.77	1463	1.05
	overall	**10,202**	**86.66**	**20,475**	**14.95**

* number of occurrences of each type

The analysis of the ComInt data started by (manually) lemmatising the top of the word frequency list in order to eliminate low-frequency inflected forms which, when grouped together, constituted medium- to high-frequency lemmas. This

process concerned the top 500 types and their co-inflected forms, this slice of the data being the limit beyond which no key-words were located. The remaining lexis was sorted into very broadly-defined semantic groups, starting at the bottom of the list with the hapax legomena then proceeding upwards, with an initial cut-off point established at 3 occurrences (this is equal to approximately 75% of the running words in the corpus). This cut-off point allowed for the identification of semantic groups which were sufficiently distinct from those represented by the key words so as to be classed as potential metaphorical source domains (a semantic group was defined as such if it contained five or more lemmas; fewer than five proved to be insufficient grounds for classification). Lemmas which could not be grouped with others, such as those listed in Table 3, were considered as candidates for linguistic metaphors, but not metaphor themes.

Table 3. Selection of ungrouped metaphor candidates

baricentro [centre of gravity]	cabina [cabin]
catene [chains]	clima [climate]
fetta [slice]	guai [trouble]
labbra [lips]	lupo [wolf]
naufragare [shipwreck]	ombelico [navel]
orizzonti [horizons]	riva [(river) bank]
scale [stairs]	seno [breast]

The higher the frequency of a metaphor candidate, the more consistency there was in its patterning in the corpus overall. The crystallisation of patternings around the node word was observed at >5, and at 7 occurrences was already marked in some cases. Strong collocational and phraseological preferences affect the polysemous potential of a word, limiting the likelihood that it will be used both literally and figuratively in the same discourse. As different meanings imply different patterns, the emergence of dominant patternings in a text or discourse makes it less likely that other patterns – and hence, other senses – will occur. This hypothesis is confirmed by the data in the corpus, where evidence is found of the crystallisation of collocational patternings, visible with upwards of ten concordance lines for the same word form, and occasionally with even fewer.

Figure 2 shows the concordance lines for *penetrazione* (penetration), where it can be seen that there is a preferred collocate *commerciale* (commercial), and a less striking but nonetheless visible preference for *penetrazione* to concern markets and sectors (*mercati*, *settori*), abroad rather than at home (*esteri*, *internazionali*). Metaphor candidates occurring in the middle-frequency bands (below the key word threshold, and, in this corpus, above ten occurrences), demonstrate greater cotextual stability than the lower-frequency candidates, and as a result begin to

a punto un piano strategico di	penetrazione	commerciale dal 2008 al 2010 , indivi
delle azioni di sostegno alla	penetrazione	commerciale del sistema Italia. Siamo
capacità di esportazione e di	penetrazione	commerciale dei nostri imprenditori e
no strumento strategico per la	penetrazione	commerciale delle nostre imprese. I m
utti , per la maggior parte di	penetrazione	commerciale finanziati attraverso la
re operazioni più complesse di	penetrazione	commerciale. Mi auguro che questi nuo
ia attività di promozione e di	penetrazione	commerciale. Per l ' anno 2007 il Min
so di internazionalizzazione e	penetrazione	commerciale. Il Ministero del Commerc
e sue possibilità di ulteriore	penetrazione	commerciale su mercati maturi ma anch
ed iniziativa che rafforzi la	penetrazione	delle imprese editoriali italiane nei
lo di forte protagonismo nella	penetrazione	dei mercati esteri. Nella situazione
la meccanica strumentale , una	penetrazione	nel settore dei servizi e in quelli a
nazionali , di accompagnare la	penetrazione	sui mercati internazionali con adegua

Figure 2. Crystallisation of patterning around 'penetrazione': All occurrences

consolidate themselves as domain-specific vocabulary or indeed terminology (Philip in press b), but this aspect is beyond the scope of the present study[8].

While it is true that metaphorical activity is determined by phraseology, and that word counts fail to distinguish between different word senses, the method described above is designed to aid the identification of metaphor candidates, not as solid proof of metaphorical activity. The words identified as potentially metaphorical were concordanced, and on the basis of the cotextual evidence, literal uses were discarded from the data set.

4. Metaphor themes and key metaphor themes

The analysis described in 3.2 yielded a large set of word forms that constitute a broad category of war and violence, plus smaller groups representing hunting, risk, submission and suffering, health, birth, death, and emotion. These areas constitute the source domains identified, but they do not become metaphor themes until the target domains have been ascertained. It is not enough to say that BUSINESS IS WAR, or RISK, or A LIVING ORGANISM: these themes are too general and of limited value in text analysis. The metaphorical targets are identified by concordancing each of the word forms separately, which makes it possible to arrive at more detailed metaphor themes which specify, for instance, that ECONOMIC PRODUCTIVITY IS A HEALTHY BODY (c.f. Figure 1a) or that INTERNATIONAL TRADE IS WAR.

8. For detailed treatment and evaluation of the methodology used in this paper, see Philip (in press b).

War metaphors are the most frequent in this corpus, and they are the one theme that seems to be common to all political discourse (Philip 2009). Identifying the metaphor targets through corpus analysis makes it possible to penetrate the mass of apparently common metaphors and home in on domain-specific themes and their uses. By looking in detail at the use of the source domain lexis, it was established that the lexical set comprising battles, fights and skirmishes was specifically linked to foreign trade, with particular reference to the main emerging economies, China and India. However, the different aspects of this state of affairs are expressed by specific word choices, meaning that the source domain terms are not interchangeable: the *battaglia* [battle] is for Italy to maintain its competitive advantage in world markets; the luxury *Made in Italy* brand is engaged in a *lotta* [fight, struggle] against the influx of imitation and counterfeit goods. Consistent with the reference to foreign trade, this metaphor also includes invasion of foreign territories, and once again the source-domain words are not interchangeable: *invadere* [invade] and *penetrare* [penetrate] are effectively synonymous, but while Italy's expansion into foreign markets (especially China and India) is expressed as *penetrazione* [penetration], the expansion of those same nations into Europe (and Italy in particular) is described as *invasione* [invasion] against which the reputation of *Made in Italy* products must be defended (Philip 2009: 105).

It is not the ubiquity of the war metaphor that makes it a key metaphor. Key metaphors are key not because they are the most frequently used within a text or discourse, but because they interact in significant ways with the key words. An apparently pervasive source domain may be found to disintegrate into several relatively infrequent metaphor themes, reducing its status. A less frequent source domain which is mapped onto just one target may, on the other hand, prove to play a more significant evaluative role in the discourse. The case of the war metaphors in this corpus is different again: as a source domain, war is by far the richest, comprising four times the number of word forms (112) assigned to the next-most frequent grouping, life and death (34). The subdivision of this domain into specific metaphor themes yields INTERNATIONAL TRADE IS WAR, which can be subdivided into TRADE IS AGGRESSIVE BEHAVIOUR, and TRADE IS MILITARY WARFARE, with EMERGING ECONOMIES ARE A THREAT, IMPROVEMENT IS A CHALLENGE, and EXPANSION IS INVASION. The concordances in Figure 3, featuring the lemmas *conquistare* [to conquer], *vincere* [win], *battaglia* [battle] and *invadere* [to invade], give a taste of these metaphor themes.

The metaphor themes identified all interact with the keywords in ways which can be considered significant. Some of the source domain lexis forms stable collocations with particular key words, such as *strumenti di difesa* [defensive tools] with *commerciale*, *conquistare* [to conquer] with *mercati*, and *sfida* [challenge] with *globalizzazione* [globalisation], and in so doing contribute to the discourse

DELL 'IMPORT Il made in Italy	conquista nuovi mercati, Russia e Cina, ma pr
argini di crescita passano per la	conquista dei mercati internazionali. Che cosa
la distribuzione. La necessità di	conquistare mercati stranieri può offrire nuove
produttive per poi lanciarsi alla	conquista dei mercati esteri. Altri paesi hanno
ese italiane che hanno tentato di	conquistare questa fascia alta di mercato
opa è stata una scommessa davvero	vincente anche dal punto di vista degli scambi
liana. Sui mercati mondiali siamo	vincenti solo se riusciamo a trasformare la nos
namente dimostrano quanto sia più	vincente una strategia positiva di adattamento
isulta a volte indispensabile per	vincere l'agguerrita concorrenza; sostegno che
l nostro territorio. Uniti si può	vincere perché siamo in grado di promuovere e v
utori stranieri: divisi possiamo	vincere qualche battaglia, ma alla lunga non s
portunità di lavorare insieme per	vincere la sfida dell'internazionalizzazione.
concorrenza e trovare formule per	vincerla è necessaria un'azione sinergica che c
: divisi possiamo vincere qualche	battaglia, ma alla lunga non sapremo valorizza
ngo. Diritti e libertà: di tante	battaglie per l'affermazione dei diritti indivi
o di servizi alle imprese, nella	battaglia per affermare l'Italia sui mercati
conclusa una delle più importanti	battaglie per la supremazia mondiale sul mercat
sono stati visti come pericolo di	invasione dai cui difendersi". Cina e India
el tessile, perché è vero che ci	ha invaso dopo la fine delle quote ma è altrett
uote ma è altrettanto vero che ci	ha invaso con un' azione predatoria puntando s

Figure 3. The emergence of metaphor themes from concordances

norms. However, although conventional and partially delexical, the metaphorical meaning still permeates the discourse:

> Through multiple on-line events, certain linguistic forms evolve to become the preferred ways of expressing metaphorical ideas across discourse communities. The language and the conceptual content stabilise, together and co-adaptively, into a particular restricted set of forms and ideas that become part of the resources of language and thinking available in the discourse community. (Cameron & Deignan 2006: 680)

The institutionalised co-selection of a war metaphor with a key-word makes it extremely difficult for the subject to be mentioned other than in terms of warfare, and this colours perceptions of international trade and commerce in general.

Not all source domain lexis forms such visible collocational relationships with the key words. In this case, the significance of the key metaphors is derived not from the fact that they are significant collocates of the keywords by *upward collocation* (Sinclair 1991: 116), but because the keywords are significant collocates of the source domain lexis, by *downward collocation* (ibid.). Thus a concordance of *Cina* [China] or *India* does not find war-related lexis to collocate particularly frequently, but individual concordances of each of the war-related lexis repeatedly

feature the target domain group including *Cina*, *India*, *Asia*, and (*Estremo*) *Oriente* [(Far) East). Examples 4–6 illustrate this phenomenon.

(4) *…ha fatto emergere quel Paese così determinante per il futuro dell'economia mondiale che si chiama Cina. Un Paese che non dobbiamo* ***temere****…*
…allowed that country which will decide the future of the world economy to emerge: China. A country which we must not fear…

(5) *La paura che le nostre imprese uscissero* ***soccombenti*** *dalla* ***sfida*** *della globalizzazione, soprattutto di fronte* ***all'aggressività*** *commerciale dell'Estremo Oriente…*
The fear that our businesses will end up as the loser in the globalisation challenge, especially when faced with the commercial aggressiveness of the Far East…

(6) *Ma la Cina rappresenta il vero futuro del tessile, perché è vero che ci ha* ***invaso*** *dopo la fine delle quote ma è altrettanto vero che ci ha* ***invaso*** *con* ***un'azione predatoria*** *puntando sulla bassa qualità e sul basso prezzo.*
'But China is the real future of the textile [industry], because it is true that it has invaded us since quotas ended but it is just as true that it has invaded us with predatory behaviour aiming at low quality and low prices.'

This kind of co-selection is important because it is hidden, and because it is pervasive. If most instances of the lemma *invadere* have China, India, Asia, or the Far East as collocates, then the meaning of *invadere* comes to be associated with those nations, as the "use and re-use of metaphors leads to the conventionalization of attitudinal judgements attached to them" (Cameron & Deignan 2006: 676). The relationship between the target and the source evades identification because downward collocation, while providing semantic information about a word (Sinclair 1991: 116), concerns low-frequency lexis which is rarely the focus of any linguistic enquiry, with or without the aid of corpora. The war metaphor is key, but covertly so.

5. Conclusions

The main difference between the keyness represented by statistically-generated key words and that represented by the metaphors discussed above, is that the former is overt while the latter is hidden from view. Overt keyness applies to those words that are frequent enough and prominent enough to attract attention. It sums up the aboutness of the text or discourse, insofar as its topics and themes are concerned. Covert keyness, on the other hand, can be hidden from statisti-

cal measures of significance because it describes repetition which operates above the level of the word. The abstract nature of covert keyness can make it difficult to pinpoint, but once located it sheds light on the underlying attitudinal stance expressed in the texts. Thus overt keyness tells us *what* is key, and covert keyness tell us the reasons *why*.

Keywords are neutral, but key metaphors are not. The war metaphors discussed in 4 are not limited to the Minister's speeches, but are found in the language at large, feeding the perception that emerging markets pose a threat to Italy, that they must be fought, and that the national territory must be defended at all costs. The textual use and function of metaphor themes interact with the aboutness of the text, and they do so in ways that are perceptible, yet barely visible. Downward collocation is difficult to spot without the use of concordancing software, but once identified, it can provide substantiation for those perceptions. In this data, the key word list indicates that China (and latterly, India[9]) is important, but does not provide the reasons for the prominence given to this nation over any of the others in the corpus. By examining the metaphorical activity in the data, however, and by examining the interaction of the covert and overt keyness, two aspects of Italo-Chinese relations come to the surface: (i) that Chinese exports and current Chinese economic policy are perceived as a threat to the Italian economy, and (ii) that China is mentioned frequently because of the threat it poses. Thus within this discourse, the key word *Cina* is seen to be bound up with the key metaphor theme of war; and pragmatically, China is painted as an enemy and a threat.

The evaluative function of metaphor is well documented, but rarely refers to anything other than instances of linguistic metaphors in texts. It has been demonstrated here that metaphor themes also play an important evaluative role, both in text and within specialised discourse as a whole. It has also been demonstrated, however, that the source domain lexical items are not interchangeable. In common with all words, metaphors form units of meaning, and it is by examining the collocational profiles of each source domain lexical item that the metaphor's evaluative meaning is built up. Abstract metaphor themes make it possible to make generalisations about key lexis, but the details are found in collocation.

9. The appearance of India on the key word list is a more recent phenomenon; it was not present in the list generated during a preliminary study carried out on the data spanning June 2006 – January 2007.

References

Black, M. 1993. More about metaphor. In Ortony (ed.), 19–41.

Cameron, L. & Deignan, A. 2006. The emergence of metaphor in discourse. *Applied Linguistics* 27(4): 671–690.

Cameron, L. & Stelma, J. 2004. Metaphor clusters in discourse. *Journal of Applied Linguistics* 1(2): 107–136.

Charteris-Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan.

Charteris-Black, J. 2005. *Politicians and Rhetoric: The Persuasive Power of Metaphor*. Basingstoke: Palgrave-MacMillan.

Deignan, A. 2005. *Metaphor and Corpus Linguistics* [Converging Evidence in Language and Communication Research 6]. Amsterdam: John Benjamins.

Deignan, A. 1999. Linguistic metaphors and collocation in non-literary corpus data. *Metaphor and Symbol* 14(1): 19–36.

Glucksberg, S. & Keysar, B. 1993. How metaphors work. In Ortony (ed.), 401–424.

Jäkel, Ö. 1999. Kant, Blumenberg, Weinrich: Some forgotten contributions to the cognitive theory of metaphor. In *Metaphor in Cognitive Linguistics* [Current Issues in Linguistic Theory 175], R. W. Gibbs & G. J. Steen (eds), 9–27. Amsterdam: John Benjamins.

Koller, V. & Semino, E. 2009. Metaphor, politics and gender: A case study from Germany. In *Politics, Gender, and Conceptual Metaphors,* K. Ahrens (ed.), 9–35. Basingstoke: Palgrave Macmillan.

Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*. Chicago IL: The University of Chicago Press.

Louw, W. E. 2000. Some implications of progressive delexicalisation and semantic prosodies for Hallidayan metaphorical modes of expression and Lakoffian "Metaphors We Live By". Privately-distributed version of: Progressive delexicalization and semantic prosodies as early empirical indicators of the death of metaphors. Paper read at the 11th Euro-International Systemic Functional Workshop: Metaphor in systemic functional perspectives, University of Gent (Belgium), 14–17 July 1999.

Ortony, A. (ed.). 1993. *Metaphor and Thought,* 2nd & rev. edn. Cambridge: CUP.

Partington, A. 2003. *The Linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House*. London: Routledge.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22(1): 1–39.

Philip, G. 2009. 'Non una donna in politica, ma una donna politica': Women's political language in an Italian context. In *Politics, Gender, and Conceptual Metaphors,* K. Ahrens (ed.), 83–111. Basingstoke: Palgrave Macmillan.

Philip, G. In press a. *Colouring Meaning: Collocation and connotation in figurative language* [Studies in Corpus Linguistics]. Amsterdam: John Benjamins.

Philip, G. In press b. Locating metaphor candidates in specialised corpora using raw frequency and key-word lists. In *Metaphor in Use: Context, Culture, and Communication*, F. MacArthur, J. L. Oncins-Martínez, M. Sánchez-García & A. M. Piquer-Píriz. Amsterdam: John Benjamins.

Rayson, P. 2005. *Wmatrix: A Web-based Corpus Processing Environmen*t. Computing Department, Lancaster University. <http://www.comp.lancs.ac.uk/ucrel/wmatrix>.

Richards, I. A. 1936. Metaphor. In *The Philosophy of Rhetoric,* I. A. Richards (ed.). London: OUP.
Scott, M. 2004. *WordSmith Tools,* Version 4. Oxford: OUP.
Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
Sinclair, J. M. 1996. The search for units of meaning. *Textus* 9(1): 71–106.
Steen, G. 1994. *Understanding Metaphor in Literature*. London: Longman.
Weinrich, H. 1958. Münze und Wort: Untersuchungen an einem Bildfeld. In *Sprache in Texten*, H. Weinrich (ed. 1976), 276–290. Stuttgart: Klett.
Weinrich, H. 1967. Allgemeine Semantik der Metapher. In *Sprache in Texten*, H. Weinrich (ed. 1976), 317–327. Stuttgart: Klett.

Appendix 1

Top 50 content words. Keywords appear in bold.

1. **ITALIA** [Italy]
2. **IMPRESE** [businesses]
3. MINISTRO [minister]
4. PAESI [countries]
5. **COMMERCIO** [commerce]
6. INTERNAZIONALE [international]
7. PAESE [country]
8. **CRESCITA** [growth]
9. **MERCATI** [markets]
10. GOVERNO [government]
11. **SISTEMA** [system]
12. **CINA** [China]
13. POLITICHE [political]
14. **SVILUPPO** [development]
15. **COMMERCIALE** [commercial]
16. **MERCATO** [market]
17. **SETTORE** [sector]
18. **EUROPEA** [European]
19. PARTE [part]
20. COMMISSIONE [commission]
21. EUROPA [Europe]
22. **INTERNAZIONALIZZAZIONE** [internationalisation]
23. **ITALY** –
24. **ECONOMIA** [economy]
25. **EUROPEE** [European]
26. **PRODOTTI** [products]
27. POLITICA [politics]
28. **INDIA** [India]
29. **ESTERO** [foreign]
30. MINISTERO [ministry]
31. MONDO [world]

32. ANNI [years]
33. **ESPORTAZIONI** [exports]
34. PARTICOLARE [particular]
35. INVESTIMENTI [investments]
36. ITALIANO [Italian]
37. **MADE** –
38. ITALIANE [Italian]
39. EURO [Euro]
40. **ECONOMICO** [economic]
41. STATI [states]
42. ANNO [year]
43. **ECONOMICA** [economic]
44. **SETTORI** [sectors]
45. **UE** [E.U.]
46. PRESIDENTE [president]
47. MAGGIORE [greater]
48. RISPETTO [(with) respect (to)]
49. **EUROPEO** [European]
50. ITALIANA [Italian]

Appendix 2

Key words

1. AREA [area]
2. AZIENDE [companies]
3. CINA [China]
4. COMMERCIALE/I* [commercial]
5. COMMERCIO [commerce]
6. CONCORRENZA [competition]
7. CRESCITA [growth]
8. ECONOMIA [economy]
9. ECONOMICO/A [economic]
10. EMERGENTI [emerging]
11. ESPORTAZIONI [exports]
12. ESTERO/I [foreign]
13. EUROPEA/E [European]
14. GLOBALIZZAZIONE [globalisation]
15. IMPORTAZIONI [imports]
16. IMPRENDITORIALE [entrepreneurial]
17. IMPRESE [businesses]
18. INDIA [India]
19. INDUSTRIA [industry]
20. INDUSTRIALE/I [industrial]
21. INFRASTRUTTURE [infrastructure]
22. INFRAZIONE [infringement]

23. INTERNAZIONALE [international]
24. INTERNAZIONALIZZAZIONE [internationalisation]
25. INTERSCAMBIO [interchange]
26. INVESTIMENTI [investments]
27. ITALIA [Italy]
28. ITALY –
29. MADE –
30. MANIERA [manner/ way]
31. MEDIE [medium-sized]
32. MERCATO/I [market(s)]
33. MILIARDI [billion]
34. MONDIALE [world]
35. PICCOLE [small]
36. PROCEDURA/E [procedure(s)]
37. PRODOTTI [products]
38. PRODUTTIVO [productive]
39. PRODUZIONE [production]
40. RIPRESA [upturn]
41. SETTORE/I [sector(s)]
42. SISTEMA [system]
43. STRATEGIA [strategy]
44. SVILUPPO [development]
45. UE [E.U.]

* Alternative forms are provided only when they appear on the key-word list; in all cases, the gender and number of the listed word forms is to be considered significant.

SECTION III

Critical and educational perspectives

A contrastive analysis of keywords in newspaper articles on the "Kyoto Protocol"

Erica Bassi
University of Trento, Italy

In 1997, an international environmental agreement was negotiated in Kyoto. It became famous as the Kyoto Protocol. Using corpus-linguistics tools, this chapter analyses the treatment of the controversial "Kyoto Protocol" in two national newspapers: the Italian *La Repubblica* and the American *The New York Times*. As a preliminary study, I compute the keywords for the articles about Kyoto, and then group the keywords into semantic fields to study the main specific groups of meaning to which Kyoto is related. I then focus on the words denoting 'disaster' and 'alarm'. To conclude, I consider some positive keywords studying the concordance lines in order to describe their associated semantic prosody. The analysis will show how, despite some similarities in the lexical usage, the meaning conveyed by the two dailies is quite different.

1. Introduction

This study aims at illustrating a method for the contrastive analysis of keywords computed for corpora in different languages but on the same subject matter. It will also show how the comparison highlights diverse perceptions of the same theme.

The theme considered is the Kyoto Protocol. The Kyoto Protocol, an international environmental treaty signed in 1997, set mandatory rules aiming to reduce the production of gases held responsible for the greenhouse effect and global warming. The Protocol came into force in 2005 but was crippled by being rejected by the United States, which, clinging to its belief in the lack of irrefutable scientific evidence of human responsibility for global warming, refused to ratify the accord.

I will analyse some aspects of the media coverage of the Kyoto Protocol, focusing on the newspaper articles published between 1997 and 2006 in two influential national newspaper comparable in ideology and audience: the Italian daily *La Repubblica* and the American daily *The New York Times*.

2. Materials and preliminary methods

To carry out my study, I collected two "Kyoto-Corpora": *La Repubblica* Corpus (832 texts, 507,533 tokens) and *The New York Times* Corpus (657 texts, 647,355 tokens), gathering all the articles dealing with the international agreement published between 1997 and 2006. The sources of the texts are two online databases: Repubblica Archive and Times Select. The selection of texts is performed according to a key-theme. The Kyoto Protocol is the key-theme through which the articles are considered. Accordingly, I included all the articles dealing with the Kyoto Treaty, whether it was the main subject or was just mentioned in a text with a different main topic (I excluded the articles where Kyoto was just named in a list).

I used TreeTagger (Schmid 1996) for morphosyntactic annotation and an Access database to manage metadata categorizations (such as author, date of publication, page, section, main topic).

Keywords, concordances and collocations were calculated with *WordSmith Tools* 4.0 (Scott 2004). Positive keywords, in Scott's sense, are statistically salient words for the text or texts under analysis compared to a bigger reference corpus. Negative keywords are the statistically unusual words of the corpus in comparison with the average language of the reference corpus (cf. Scott & Tribble 2006).

The distinctive features of my corpora were determined comparing them to the general language used by the respective newspapers, using the following reference corpora:

1. *The New York Times* portion of the American National Corpus (ANC Consortium 2006); it contains 4,148 articles published in every section of the newspaper in 2002, the overall number of words is 3,625,687. Further information can be found on the web site: http://americannationalcorpus.org/SecondRelease/contents.html#nytimes
2. *La Repubblica* corpus, compiled at the University of Bologna, made up of 380 million words, contains the articles published between 1985 and 2000. For further information visit the web site: http://sslmit.unibo.it/repubblica (see also Baroni et al. 2004)

3. Defining a keyword in news discourse

Newspapers can be considered, as proposed by Marrone (2001) and Landowski (in Semprini 1990), as semiotic instruments of the rewriting of reality or semiotic subject embedded in society. The events they describe are irrecoverably lost,

inaccessible; the way they are represented is always loaded with interpretation. In this context words do not reveal reality but create a naturalized reality. According to Landowski, newspapers are a community in which the discourse defines the identity of the daily and, at the same time, favours the reader's identification. Therefore, considering newspapers as a discourse community, with which the audience identifies, their average lexicon shapes, describes and expresses what is accepted by their community. In Fairclough's words "the naturalized lexicon is seen as commonsensical and based on the nature of things and people rather than in interest of classes and other groupings" (Fairclough 2000: 35).

The readers recognise themselves in the newspaper's language and ideology, therefore, keywords are extremely important because they are marked words more liable to be noted and associated to the themes the articles are about. The newspaper's average lexicon is what is posed as normal, the words peculiar to a group of articles highlight the deviation from the norm. On the one hand, the theme is responsible for this shift, and on the other hand, the keywords indicate the perspective from which this theme is presented. I define the keywords of my newspaper corpus as the deviation from the norm in connection with the theme being analysed. Given these premises, I can state that the set of words that can be associated to Kyoto are fundamental to an understanding of the representation of the Kyoto theme.

4. Semantic fields

In the first stage of my analysis, I grouped the keywords into semantic fields. The semantic fields were elaborated manually, referring to a dictionary of synonyms and antonyms.

As a preliminary step, I analysed the negative keywords; Table 1 reports the main semantic fields among them.

In both newspapers there was a below-average presence of words in the semantic field of 'family' and 'entertainment'. This could be a sign of the disconnection of the theme "Kyoto" from concrete life. I can therefore formulate a hypothesis of the abstract coverage of the theme, which keeps it at a distance from everyday life.

Table 1. Negative keywords semantic fields

La Repubblica	*The New York Times*
Economics (borsa, titoli, tassi…)	Economics (investors, stock, financial…)
Family (moglie, fratello, figlio…)	Family (family, mother, parents…)
Entertainment (film, spettacolo, concerto…)	Entertainment (sport, film, music…)

As we will see later, we can find a class of words referring to economics, both in positive and negative keywords. This inconsistency between the presence of terms in the semantic field of economics both in the positive and negative keywords deserves further comment. We can understand this peculiarity by investigating the nature of economic reference in positive and negative keywords. In the positive keywords the lexicon refers to carbon credit exchange and industrial production; in the negative ones the terminology is mainly financial. The Protocol regulates *International Emissions Trading*, which controls the exchange of emission credits among the countries. Europe has its own *European Union Emission Credit Scheme*, but this is not extended to international markets. The lack of financial terminology in the Kyoto corpora mirrors the perception of environmental commerce as something external to the normal economic flux.

The following table shows the semantic fields found in the positive keywords of *La Repubblica* and *The New York Times*.

Table 2. Positive keywords semantic fields

La Repubblica	*The New York Times*
Economics	Economics
Effect on the Planet	Effect on the Planet
Power sources	Power sources
Environmentalist jargon	Environmentalist jargon
Negotiation	Negotiation
Politicians	Politicians
Parts of the Planet	Part of the Planet
	Politics
Change	Reduction
Science	Science
Chemical substances	Chemical substances
States	States
Transports	Transports

The labels denominating the semantic fields correspond to one another but, considering the keywords themselves, many differences come to light. The most meaningful semantic fields can now be discussed in detail: 'negotiation', 'change' and 'effect on the Planet' (see Table 3).

The language of negotiation in *The New York Times* presents a lexicon that, along with words referring to ratification and commitment, denotes antagonism: see words such as *talks*, *debate*, *negotiators*. In *La Repubblica* the lexicon connected to negotiation is more unequivocal; we find words such as *accordi*, *obiettivi* and

Table 3. Words in semantic fields

Negotiation		**Change**		**Effect on the planet**	
LR	***NYT***	***LR***	***NYT***	***LR***	***NYT***
Ratifica	talks	ridurre	reductions	inquinamento	warming
impegni	ratification	riduzione	reduce	riscaldamento	greenhouse
accordo	delegates	aumento	reducing	surriscaldamento	pollution
obiettivi	debate	crescita	cut	caldo	heat
accordi	negotiators	sviluppo**	reduction	calore	
obiettivo	ratified		limits	uragani	
ratificare	rejected			disastri	
ratificato*	ratify			alluvioni	
	issues			desertificazione	
	issue			siccità***	

* Ratification, dealings, accord, targets, accords, target, to ratify, ratified.

** To reduce, reduction, increase, growth, development.

*** Pollution, warming, overheating, warm, warmth, hurricane, disasters, flood, desertification, drought.

ratifica[1], which represent the activity of negotiation as plain and unproblematic. In *La Repubblica*, negotiation is shown as an activity whose aim is to reach an agreement, rather than discuss two different positions.

Some words among the keywords can be associated on the basis of their reference to 'change'. In *The New York Times* we notice prevalence of the concept of 'reduction'. The focus on privation and deficiency gives a negative meaning to the corpus. In *La Repubblica* the presence of verbs of 'reduction' is combined with other words indicating 'change', such as *crescita* and *sviluppo*[2]. Therefore, there is not the same emphasis on privation and limitation. The representation of the characteristic actions in the Kyoto Protocol's universe of discourse opens up a wider possibility of action.

Still considering the positive keywords, let us proceed to compare the lexicon used to name the effects of climate change on the Earth. *Greenhouse effect*, *pollution* and *warm* are found in both newspapers; in addition to them, in Italy, we notice the presence of words referring to environmental disasters (*uragani, disastri, alluvioni, desertificazione, siccità*[3]). The absence of words denoting 'disaster' in the

1. Accords, targets, ratification.

2. Increase and development.

3. Hurricane, disasters, floods, desertification, drought.

New York Times keywords shows that, in the American discourse on global warming, natural calamities do not have a predominant position in comparison to the average vocabulary of the paper.

To attempt to explain this divergence, I isolated the keywords in the sub-corpora of texts whose main topic is 'alarm'. A sub-corpus is a group of texts extracted from a corpus on the basis of a distinctive feature. The "alarm sub-corpora" were sorted in accordance with their topic (thought as a pragmatic category). I assigned a topic to each article on the basis of the "aboutness" (Phillips 1989; van Dijk 1999) suggested in the headline, it being a pointer of the perspective from which the article will be interpreted (Eco 1971).

The sub-corpora keywords were computed comparing the sub-corpus with the whole Kyoto-corpus. In the following table, I show the main Italian and American keywords in the alarm sub-corpora.

Table 4. Positive keywords in alarm sub-corpora

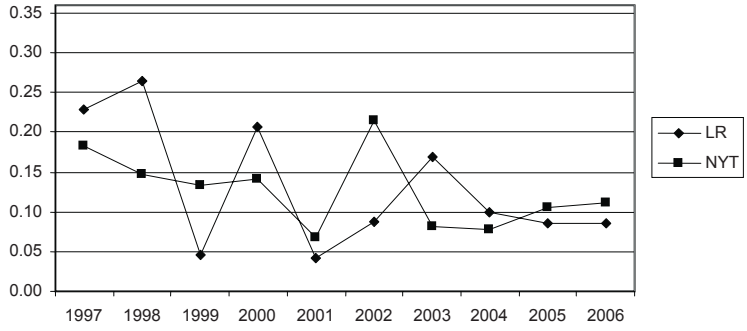
Keyword ALLARME – Repubblica
gradi, caldo, temperatura, climatico, estate, allarme, più, uragani, mare, scienziati, aumento, ghiacciai, terra, ghiaccio, acqua, climatici, cambiamento, secolo, clima, ghiacci, disastri, ultimi, fenomeni, temperature, aria, mari, nord, siccità, acque, mediterraneo, calda, rischio, estremi, eventi, danni, anni, riscaldamento, pioggia, serra.
Keyword ALARM – New York Times
ice, island, arctic, snow, birds, polar, temperatures, sea, bear, Chinese, weather, barrow, water, Asia, bears, ocean, temperature, degrees, China's, nest, warm, years, summer, year, seas, north, data, black, report, smoke, fish, George's, China, rise, die, feet, melt, cold, beach, warmer, Niño, colony, winter, climatic, pack, rain, century, village, Asia's, season, droughts, scientists, bird, surface, average, Inuit, sky, feathers, lagoon, peasants.

We notice that, in *La Repubblica*, we have explicit alarmist words, reference to general words indicating natural calamity, nomination of concrete problems, and words that point to warming. In the *New York Times* the alarm is launched mainly in reference to the *ice melting*, the *sea rising* and to the animals that will suffer the consequences of climate change: *birds* and *bear*. In the *New York Times* the problem is thus presented as remote and faraway. The limited reference to hurricanes and to high summer temperatures in cities and towns, the avoidance of the designation of events with disaster-like words and the lack of explicit alarmist vocabulary minimize the problem of global warming and relegate its consequences to outside the United States' territories.

To check to what extent explicit alarm words, general denomination of disasters and words describing the effects of climate change are connected, I have

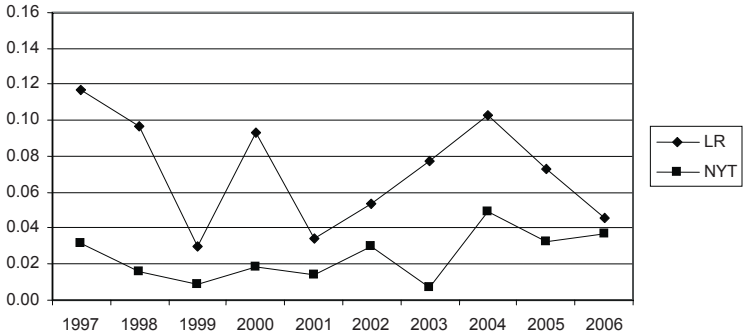
calculated the diachronic evolution of their absolute frequency in the Kyoto corpora. The frequency is normalized on the basis of the number of tokens per year. Words included in this calculation are taken from the *WordSmith Tools*-computed wordlist. The use of the wordlist allows us to tackle the subject from another point of view and rectify any imprecision engendered by the keyword calculation on inflected languages.

Graph 1 shows the diachronic evolution of words pointing to a specific natural event, such as *hurricane*, *floods*, *ice melting* and *global warming* etc. The graph highlights how, despite the fluctuations over the years, on average, the facts connected to climate change are named with the same frequency in the two newspapers.



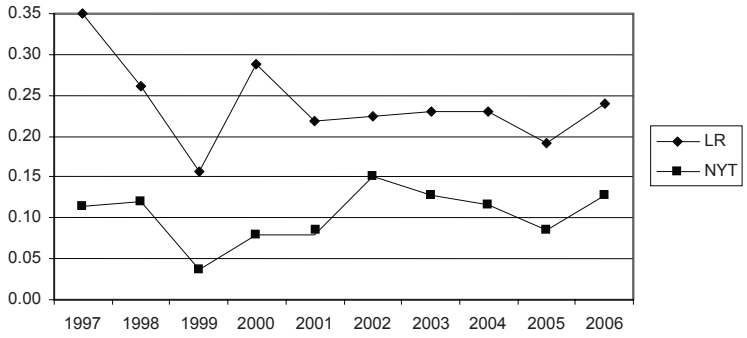
Graphic 1. Warm/hurricane/ice trend (1997–2006)

In the following graphs, I show the evolution of occurrences of words in the semantic field of 'disaster' and 'alarm'. If we consider the occurrence of words in the semantic field of 'disaster' we find confirmation of what we have noticed in the keywords of the alarm sub-corpora. We notice that, although the weight of the denomination of specific natural events due to climate change is about the same in the corpora, there is far less use of words explicitly denoting disaster in the *New York Times*:



Graphic 2. Disaster trend (1997–2006)

Similarly, the frequency of words denoting alarm is higher in *La Repubblica*:



Graphic 3. Alarm trend (1997–2006)

5. Keywords

Until now we have considered the semantic fields of the Kyoto corpora keywords and compared some groups of words between the newspapers. To conclude this section, I would like to concentrate on the study of some words taken among the positive keywords of the corpora. Working across two different languages entails a distance in lexical usage, the following keyword comparison will focus on words pointing to roughly corresponding concepts. The keywords I want to analyse are: *ambiente*[4], *environment*, *gas*, *gases* to see how they are represented and enacted in the discourse on Kyoto. This will imply studying the concordance lines (Tognini-Bonelli 2001; Sinclair 2003) in order to describe their semantic prosody (Sinclair 2004) and the evaluative language associated with them (cf. Hunston & Thompson 2000).

We can see some sample concordances of "ambiente" and "environment" in Table 5.

Environment appears in the first positive keywords as a noun in *La Repubblica* and has position 35 in the *New York Times*. The noun form is used in both newspapers as a modifier. *Environment* appears in the dailies as a very politicized word, often used as a noun modifier with words in the semantic field of politics. In other cases the environment is described as a passive entity to be defended, protected and regulated. As Halliday (1985) explains, the modifier qualifies the head of the noun phrase, restricting its reference to a category. The attention is thus shifted from the environment to what is in relation to it.

4. Environment.

Table 5. Concordances "ambiente"/"environment"

La Repubblica	
(1)	"scema" la mia proposta e giudica inutile l' iniziativa del **ministro dell'Ambiente**,
	vittima a suo avviso di "tentazioni bulgare" al punto che potrei
(2)	amo vivendo il discorso cambia: bisogna fare altre scelte. Come **ministero**
	dell'Ambiente abbiamo già cominciato investendo 325 miliardi in progetti pilota per i
(3)	ai cittadini. Nell'agenda di Dario Esposito, riconfermato **assessore all'Ambiente** nella
	seconda giunta Veltroni, sono questi i punti programmatici che do
(4)	razione (28 a carico dell' Italia) aperte da Margot Wallstrom, **Commissario**
	all'ambiente, per violazione delle direttive europee. «L'Italia è stata denunciata
The New York Times	
(1)	not meet their commitment is not clear. Mr. Mills and Mr. Becker said Canada's
	environment minister, Stephane Dion, was trying to amend the country's Environme
(2)	Senator James M. Inhofe, the Oklahoma Republican who is chairman of the Senate
	environment committee. The hearing on Wednesday and the coming Senate vote
(3)	orate the common rules into their national legislation. But she is the first **environment**
	commissioner to use the power of her office to impose fines for
(4)	do this or that if their bosses are not doing it," said Yoshihisa Fujita, the **environment**
	ministry official in charge of the campaign. "We targeted top exe

Table 6. Concordances "gas"

La Repubblica	
(1)	firmarono un accordo per diminuire in modo drastico l'emissione dei **gas responsabili**
	dell'effetto serra. Sostenuti dalle varie lobbies, Hagel
(2)	i governi di molti paesi si sono impegnati a ridurre le emissioni di CO2 e dei **gas**
	responsabili dell'effetto serra, ha lanciato "Cambio di clima (meno consumi
(3)	è di gran lunga il paese più inquinante per quanto riguarda le emissioni di **gas**
	responsabili dell' effetto serra. In caso di rifiuto degli accordi di Kyoto,
(4)	discorso sullo stato dell'unione per annunciare una inversione di rotta sui **gas che**
	provocano l'effetto-serra. Resta infatti il problema numero del nostro
(5)	approvare i provvedimenti assunti alla conferenza di Kyoto per la riduzione dei **gas che**
	provocano l'effetto serra".
The New York Times	
(1)	loser to final approval of the first international treaty to limit emissions of gases **linked**
	to global climate change. The council voted
(2)	how their current emissions contribute to concentrations of carbon and other gases
	linked by scientific organizations to a global warming trend. Some
(3)	would be obligated under the Kyoto accord to specific cuts in emissions of the gases
	believed to contribute to global warming. At least initially, developing
(4)	a specific reduction in the emission of carbon dioxide and the other greenhouse gases
	that **contribute** to global warming, agreeing only to ''substantial reduction
(5)	the Kyoto agreement on climate change. Steps are being taken to control gases that
	contribute to planetary warming, even though it could be

The gases named in the Kyoto Protocol are mainly greenhouse gases. The modifier *greenhouse* denotes a judgement on the gases because it holds that they are responsible for the greenhouse effect. The gases in *La Repubblica* are "emessi" and "prodotti", they are "controllati" and "tagliati"[5]. *La Repubblica* is explicit in its negative attitude towards gases, denominating them: *nocivi*, *inquinanti* and *tossici*[6]. The use of descriptive prepositions to claim knowledge on the gases and establish their negativeness is, in Fowler's words, an "unquestioned assertion" (Fowler 1991/2003: 127). As we can see in some randomly selected concordances in Table 6.

La Repubblica highlights the influence of the gases in the production of the greenhouse effect with the words *provocano* and *responsabili*[7]. The responsibility for climate change is assigned in the *New York Times* through verbs such as *contribute*, *believed* and *linked*. The American daily does not define the gases negatively but establishes the uncertainty of the attribution of responsibility by presenting it as an opinion and not as a fact. What is particularly noteworthy is that scientists are named to establish this uncertainty.

To test how scientists are represented in connection to gases in the Kyoto corpora, I take into consideration the concordance lines in which gases and scientists co-occur. A sample is shown below:

(1) have," the site says, using the scientific shorthand for carbon dioxide, the **gas some scientists say helps cause** global warming, "and that these high levels

(2) year trying to build a consensus for long-range policies to reduce greenhouse **gas emissions scientists have linked** to warming. The new round of negotiations

(3) national commitments to cut emissions of carbon dioxide and other heat-trapping **gases that scientists link** to global warming. "You are watching 163 nations

(4) measures is used to stem tailpipe and smokestack emissions of heat-trapping **gases that scientists say are contributing** to global warming. "When you're

(5) binding limits on the output of carbon dioxide and other so-called greenhouse **gases that scientists believe** are causing traumatic changes in the climate.

(6) advice and publish audited inventories of the companies' emissions of the **gases, which scientists say appear** to be contributing to a potentially harmful

(7) deadline in Kyoto, Japan, for talks on an agreement to cut emissions of **the gases that scientists warn may be** warming the planet. After delegates caucused

5. Emitted, produced, controlled, cut.

6. Harmful, polluting, toxic.

7. Cause, responsible.

We notice that the concordance lines reveal a unique pattern with different synonyms: the lemma *gas* is followed by a relative clause that specifies the scientists' view about the gases themselves. The opinion on global warming remains an opinion, even if it is reported by scientists. Various strategies are used to cast doubt on what the scientists say:

1. *some scientists say*: implies that there is not a shared opinion on climate change;
2. the verb *contribute* entails that gases are one of the problems among others; this is a way of reducing their responsibility;
3. the verbs *lin*k and *connect* imply that there is not a cause–effect relationship between gases and global warming, but a vague and indefinable connection;
4. the use of *believe*, a word in the semantic field of religion and faith, reaffirms that the certainty of the scientists' conclusions is not founded.

In this way the impression of authority that the word "scientist" could confer to the sentence is undermined. In the Kyoto matters their view is represented as an opinion to be trusted and not as a demonstrated or demonstrable fact. The same search on *La Repubblica* Kyoto corpus does not produce any results.

6. Conclusions

Despite the fact that the keywords of the two corpora shared the same semantic fields, the relevance and co-text of these keywords convey different meanings for the two newspapers. To conclude, I can state that the analysis reveals two problematic levels in the coverage of Kyoto in the dailies: the negotiation level and the environmental level. In the United States there is a major recognition that enacting the Protocol would lead to concrete consequences: the negotiation is therefore felt as problematic. In contrast, the lexical usage in *La Repubblica* reveals a positive tendency towards the agreement, which is supported by the unquestioned assertion of the cause–effect relationship between gas production, pollution and global warming; a connection that in the States is alluded to but never accepted as a fact. The *New York Times* reveals a major awareness of the difficulty of change, while *La Repubblica* reveals a major awareness of the serious negative consequences of climate change.

The two newspapers use different strategies to talk about a problem while keeping it at a distance. The Italian newspaper builds the myth of an easily saveable environment. However, this myth of salvation represented in the newspaper is unrealistic, as it does not argue for a concrete change in lifestyle but only an easy political agreement. In the US, the reluctance to recognise the problem of global warming is explained by a realistic perception of the difficult task of changing ingrained habits.

References

ANC Consortium. 2006. *The American National Corpus*. Distributed by Linguistic Data Consortium.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. & Mazzoleni, M. 2004. Introducing the "*La Repubblica*" Corpus: A large, Annotated, TEI(XML)-compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*.

Eco, U. 1971. *Le forme del contenuto*. Milano: Bompiani.

Fairclough, N. 2000. *New Labour, New Language?* London: Routledge.

Fowler, R. 1991/2003. *Language in the News*. London: Routledge.

Gruppo Editoriale L'Espresso. 2006. *Repubblica Archive* (online database) <http://www.repubblica.it/>.

Halliday, M. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hunston, S. & Thompson, G. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP.

Marrone, G. 2001. *Corpi sociali, processi comunicativi e semiotica del testo*. Torino: Einaudi.

Phillips, M. 1989. *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham.

Schmid, H. 1996. TreeTagger [Computer Software]. Institut fuer maschinelle Sprachverarbeitung (IMS) Universitaet Stuttgart.

Semprini, A. (ed.). 1990. *Lo sguardo semiotico: pubblicità, stampa, radio*. Milano: Franco Angeli.

Scott, M. 2004. *WordSmith Tools* 4.0. Oxford: OUP.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Sinclair, J. 2003. *Reading Concordances*. London: Longman.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

The New York Times Company. 2006. *Time Select* (online database). <http://www.nytimes.com>.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.

van Dijk, T. 1999. Opinions and ideology in the press. In *Approaches to Media Discourse,* A. Bell & P. Garrett (eds), 21–63. Oxford: Blackwell.

Keywords in Korean national consciousness

A corpus-based analysis of school textbooks

Soon Hee Fraysse-Kim
Nagoya University of Commerce and Business, Japan

This study identified keywords that trigger national consciousness (here termed "Weness") of Koreans through an analysis of school textbooks using corpus-based techniques. The corpus is taken from Korean language textbooks compiled by the authorities and used in elementary schools in four Korean communities: in South Korea, North Korea, Korean Residents in Japan, and Korean Residents in China. Although the members of these four Korean communities come from the same ethnic stock, the current socio-political characteristics of each group are clearly different. Nevertheless, "Weness" in the Korean mindset, often expressed by the word *wuli* (we, our), suggests an immediate sense of homogeneity common across politico-social borders. Assuming that school textbooks are a linguistic reflection of a prevailing ideology, I show how this feeling is crystallized in a few special keywords frequently used in textbooks, and point out how Korean national identity is internalized and reproduced by school education.

1. Introduction

The Korean peninsula lies between China and Japan and, historically, it has been within the political and cultural sphere of influence of these two countries. From 1910 to 1945, the Korean peninsula was under Japanese colonial rule. During this period a sizeable number of Korean peasants and workers left their homeland and mostly settled in Manchuria and Japan. After World War II, many of them stayed in China or Japan in pursuit of better economic opportunities. With Japan's surrender in 1945, Korea was liberated from thirty-six years of Japanese colonial rule. But soon after independence, in a time of post-colonial confusion, the Korean peninsula was caught in an ideological power game between the USA and the USSR. War broke out in 1950, and the peninsula has remained divided ever since in two socio-politically opposed states.

Thus, four Korean communities now coexist in the Korean peninsula and its two surrounding countries. In addition to the communities of South Korea (SK) and North Korea (NK), there are also the communities of Korean residents in China (KRC), and Korean residents in Japan (KRJ). The founders of the community of Korean residents in China, and of the community of Korean residents in Japan, were immigrants who left the Korean peninsula for economic reasons, or who were forced to leave during the Japanese colonial occupation. Today, Korean residents in Japan still live there as foreigners, even after three generations, while Koreans in China are regarded as one of the many distinct ethnic minorities in China. However, after almost a century, these two overseas Korean communities still seek some kind of education aimed at preserving their children's Korean ethnic identity through schooling.[1] North Korea is the last Stalinist-style socialist state in the world. South Korea, by contrast, is a westernized capitalist democracy. As Table 1 shows, the socio-political situations of these groups today are various.

Table1. Socio-political situation of the 4 communities

Community	**SK**	**NK**	**KRC**	**KRJ**
Ethnicity	Korean			
Location	Southern part of the Korean Peninsula	Northern part of the Korean Peninsula	North-eastern part of China	Japan
Regime	National citizenship		Ethnic minority	
Population status	Citizens			Permanent foreign residents
Socio-political system	Democratic capitalism	Totalitarian communism	Communism	Democratic capitalism
School system	Public education		Ethnic autonomy	Miscellaneous schools

1. 1.9 million ethnic Koreans live in Northeast China. The largest ethnic Korean community is located in *Yanbian* Korean autonomous prefecture where an autonomous ethnical school system, from elementary to university levels, operates. As of 2002, there is one university, eight high schools, twenty-five junior high schools and seventy-four elementary schools. In Japan, Korean residents whose population is estimated at around 0.6 million (mainly concentrated in the *Osaka* area), are divided into two groups by political affiliations between South Korea and North Korea. The ethnic school system is actually run by The General Federation of Korean Residents, a group affiliated with North Korea. Currently, there are forty elementary schools, twenty-eight junior high schools, eleven high schools and one university (Fraysse-Kim 2006).

In this paper I use, without distinction, the terms "We consciousness" or "national consciousness", or "Weness". By these terms I refer, in Foucault's terminology, to *la conscience du Même* in relation to *l'Autre* (Foucault 1969), as it is developed by ruling ideologies. The ideologisation of "Weness" takes place in the context of socialization during schooling, through which children become aware of their identity as members of society. Such conscientisation, according to Pêcheux, occurs through the "interpellation" of individuals: the process by which they become subjects of their discourse (Pêcheux 1975: 113). For instance in national language textbooks the process is accomplished by teaching children via the medium of words, the idea behind the visual or vocal symbol, especially of those "keywords" which are "significant, indicative words in certain forms of thought" (Williams 1976: 13) in the society in question. Those words have their meaning determined by the contexts of any discussion employing them in the discourse of textbooks. Thus what children learn is "the meanings of the discourse object" which has "nothing to do with reference to the discourse-external reality" (Teubert 2007: 68–69). Nevertheless the transparency of the meaning is universally and implicitly agreed to by all of "us" in the discourse of textbooks. In identifying keywords through these orthodox meanings, children become subject to the "correct use of language", correct in this case being the meaning sanctioned by the dominant ideology. In this sense it could be said that a "Weness" develops through an implicit consensus about the meaning of the keywords.

In this way school textbooks are "ideological materials" (Pêcheux 1975, 1981, 1983) where social-institutional practices are discursively accomplished. In North Korea, school textbooks are saturated with the official ideology *juche* which impels people to form one block dedicated to the "Dear Leader". In South Korea, following the accession to power of the longtime oppressed political leaders of the Democratic Party, the revised school textbooks strove to establish a new social order, or even make a clean break with the past, modifying and correcting the boundaries of "We" (from a defensive-anti communist "We" to a peaceful reunificationist "We"). In the textbooks of KRC, Chinese nationality is strongly emphasized, but as in the textbooks of KRJ, the discourse is mainly constructed in a conscious effort to maintain the collective ethnic identity. Thus in both overseas Korean communities, *minzok hakkyo* (ethnic schools) act as a centripetal force to keep their members pulling together.

Although Koreans of the four communities have a shared history, they share neither the present nor perhaps the future. But it seems that Koreans are easily able to attain some sense of homogeneity between the groups beyond the particularity of each community. Thus "Weness" in Koreans' minds, expressed by the word *wuli* (we, our), suggests not only a sense of fellowship towards the other members of the community, but also the boundaries that can be easily extended

to *kyore* (same tribe) and *donpo* (compatriot), hinting at an immediate sense of homogeneity common across the politico-social borders. It seems for Koreans that there are two boundaries of "Weness"; one is closed for the individual as a member of a collectivity and the other is open to all Koreans, in which Koreans feel each other to be homogeneous and compatriotic no matter how varied and different the geo-political backgrounds between them may be.

My assumption is as follows: even after a century of individual politico-social evolution, if between these four communities' members a certain feeling of homogeneity is still shared, among the factors which constitute each community's national consciousness, there must exist some elements in common. As mentioned above, in all four communities, the elementary school textbooks are compiled by their respective authorities and thus exist in only one version in each community. Consequently, these textbooks are the only media presented in school education and therefore the fact they constitute "official knowledge" (Apple 2000) seems incontrovertible. Moreover, since primary education is compulsory, we can assume that in any kind of community the knowledge purveyed in the textbooks of elementary school is singularly prevalent. Therefore, taking a close look at the content of school textbooks appears to be an excellent means to discover whether there exists such cultural transmission as would inspire a "We consciousness" among Koreans. The study therefore investigates the keywords which may convey this consciouness by clarifying 'what is central and typical in the language' (Sinclair 1991: 17) found in these textbooks, using corpus linguistics methodology.

"[E]vents which are frequent are significant" (Stubbs 2001: 29). If we take into account the fact that the frequent repetition of a certain pattern of lexical collocating is "tending towards the cliché" (Sinclair 1966: 412), the statistical analysis of an ideological discourse, such as that evident in school textbooks, proves to be very effective in filtering ideologically fixed expressions.

2. Data

For this paper I established an electronic corpus of approximately three hundred thousand words in Korean from the four communities' Korean language textbooks. The corpus is composed from four sets of textbooks in use in South Korea, North Korea, and by Korean residents in Japan and Korean residents in China (named sk, nk, krc and krj respectively). Each set of textbooks was currently in use in 2006 and covered the six primary school grades.

Table 2. Corpus

sub-corpora	total number of characters (unit: 100000)	total number of words
sk	3.49	114,312
nk	2.78	82,766
krc	2.27	66,839
krj	1.23	35,716

I will first observe and compare the most frequent words in and between each set of textbooks. I will then investigate the co-occurrence pattern of a word *wuli* (we, our) whose high frequency was almost uniform through the four textbooks. Then, I examine the usages of foreign country names in order to locate "other" in relation to *wuli*, the quintessential expression of "us" in Korean. Finally, I will investigate the use of the word *ilbon* (Japan) which shows a universal tendency in frequency of occurrence and a contextual environment across the four textbooks.

In the Korean language, a word is considered as a string of characters with a space on either side. But the application of this segmentation is different across the four communities, hence the statistical comparisons between the four textbooks in this paper are performed on the basis of relative values with the number of character occurrences as the standard. The results derived from quantitative analysis are shown by frequency lists and graphs obtained by the correspondence analysis[2].

3. Findings

3.1 Frequency

Table 3 presents the word order frequency list of the ten most frequently used nouns and pronouns in the four textbooks. The raw frequencies of occurrence of words are followed by the normalized figure of the number of occurrences per 100,000 characters.

2. Correspondence analysis is a statistical visualization method for displaying the associations between the levels of a two-way contingency table. Greenacre pointed out that "a graphical description is more easily assimilated and interpreted than a numerical one". In this paper I try to visualize the data tables in graphic form using this method in order to provide a global view of information with "possible explanations"(Greenacre 1984: 3).

Table 3. The 10 most frequent nouns and pronouns found in each textbook

	sk			nk			krc			krj		
	noun/ pronoun	Freq	Normal-ised	noun/ pronoun	Freq	Normal-ised	noun/ pronoun	Freq	Normal-ised	noun/ pronoun	Freq	Normal-ised
1	I	1572	450	we	797	287	I	786	346	I	452	367
2	we	926	265	great-marshal	670	241	we	534	235	we	304	247
3	you	507	145	I	606	218	person	402	177	person	209	170
4	person	458	131	father	573	206	mother	278	122	you	168	137
5	home	359	103	marshal	504	181	father	262	115	mother	140	114
6	teacher	341	98	mother	391	141	home	244	107	friend	139	113
7	child	325	93	child	354	127	you	233	103	tiger	127	103
8	thought	293	84	teacher	344	124	friend	189	83	school	126	102
9	friend	280	80	friend	283	102	teacher	158	70	teacher	121	98
10	father	279	80	person	269	97	heart	134	59	rabbit	119	98

Table 3 shows that in "I" use, there is some diversity between the four textbooks. Compared to the other textbooks, the sk "I" use occurs clearly more than in any other textbook. It suggests that SK society emphasizes individuality more than do the other communities. On the contrary, in nk "we" is dominant over "I". Moreover, expressions like "great-marshal", exclusively referring to North Korea's leaders, Kim father and son, occur more frequently than "I". Meanwhile, there is some quantitative similarity in the use of "we" across the four textbooks. Statistically speaking, the frequencies of "we" across the four textbooks are not significantly different ($\chi^2(3) = 5.954$, ns) though there is a significant deflection in the frequency in the use of "I" ($\chi^2(3) = 80.054$, $p < .01$).

3.2 The use of *wuli* (*we, our*)

The word *wuli* means 'we' when used as a pronoun and 'our' when used as an adjective while the singular form *na* means 'I' and *nae* means 'my'. In ideological discourse such as school textbooks, the function of the personal pronoun *we* is, in general, the "*inclusive we*" which seeks to bring the reader into a consensus, or the "*rhetorical we*" which is used in the collective sense of the community (Quirk et al. 1985). It is a solidarity-triggering expression *par excellence*, as the following SK textbook example shows:

> People of our country may like to use *wuli*. We Koreans prefer to say "our school" or "our mother" rather than "my school" or "my mother". *Wuli* is the word by which a speaker designates friends. *Wuli* is the word to express the sense of community.
>
> (from Korean national language textbook for 6th grade, translated from Korean by the author)

When *wuli* is used to mean 'our', as in "our X", the X is considered, especially in an ideological discourse, as something that "we" together should care for. In other words, where instead of "Korean traditional culture", "Korean army", "Hangul alphabet" or "Korean athlete", textbooks use the terms "our traditional culture", "our army", "our alphabet" or "our athlete" respectively, these can be seen as pure ideological expressions that reinforce the solidarity between members and encourage a sense of collective duty towards this common property that becomes a reflection of the supreme reason for the existence of the community. Hence it is interesting to identify the words frequently co-occurring with *wuli* (such as "our N") in order to understand what is considered important to community solidarity.

For that purpose the following procedures were undertaken. In each textbook I asked: (1), what nouns occur immediately after *wuli* (our) (2), how frequently does a particular collocation of "*wuli N*" occur (3), are there any other textbooks in which this particular collocation occurs? Then I selected, among all nouns which come after *wuli*, the nouns whose co-occurence with *wuli* is found more than once in a normalized count of the number of occurrences per 100,000 characters in each textbook. Then I calculated the MI score[3] of each pronoun-noun collocation in order to find the strength of co-occurrence between a particular noun and the pronoun *wuli*. An example of the data is shown in Table 4 with the list of some nouns and their MI score.

Table 4. Nouns collocating with *wuli* and their MI score

words	textbooks				words	textbooks			
	sk	nk	krc	krj		sk	nk	krc	krj
home	4	4	3	5	army	0	4	7	0
mother	3.15	1	0.43	1.34	brother	0	5.16	0	5.07
school	2.34	3	3.22	4	society	3.17	0	0	0
country	5.27	5.31	6	6	hometown	0	4	4	2.07
class	6.42	5.16	5.38	5.17	homeland	0	2.44	0	0
language	2.37	3	3	5	land	2.28	0	0	0
ancestor	7	0	6	5.29	party	0	5.31	0	0
living	3	0	3.3	4.37	village	3.24	4	4.12	0
traditional-culture	5	0	0	0	troop	0	4.33	4	0
friend	0	3	0	2.1	letter	2.2	2.44	4	5.39
mangyondai	0	0	0	6.29	team	0	0	0	6.42

The value 0 means that there is no co-occurrence or lower frequency than once in a normalized count in each textbook.

3. The index showing the strength of mutual connection.

For a global view of the data, I have drawn a representation (Figure 1 below) yielded by correspondence analyses of these data[4].

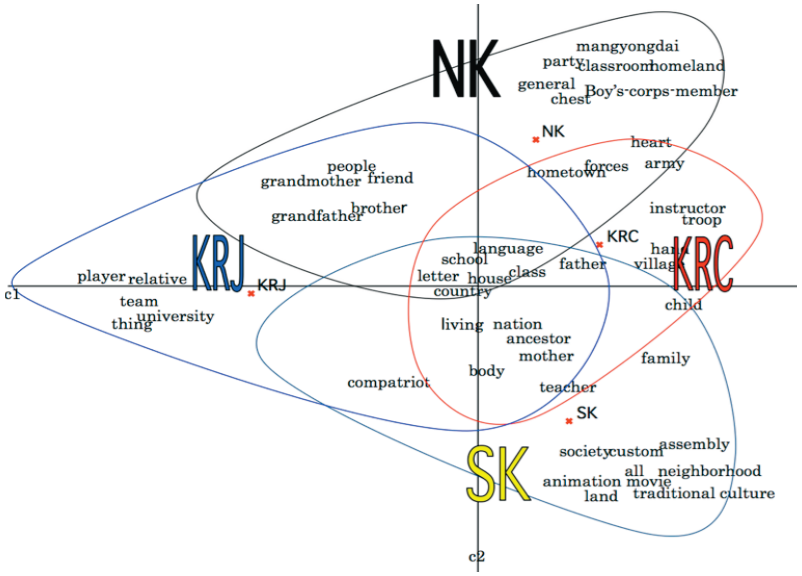


Figure 1. Nouns collocating with *wuli* in the four textbooks

In Figure 1, there are row points for the nouns and column points for the textbook marked by community with an asterisk dot. Using Lee's (1996) explanation, it can be understood that the distance between nouns is a measure of the similarity between the row score profiles. Each noun lies in the neighborhood of the textbook's name in which the noun's MI score is prominent. E.g. *mangyongdai* [5] and "traditional-culture" are far from each other because their profiles are different (co-occurrences like *wuli mangyongdai* appears only in nk textbooks while the expression "*wuli* traditional-culture" occurs only in sk textbooks). Also the word "ancestor", whose co-occurrences with *wuli* are observed in the three textbooks of sk, krc, krj simultaneously, lies between three column dots of krc, krj and sk, but nearer to sk than to the others. The reason can be inferred from the fact that the noun's MI score in sk is higher than in krc and krj.

I have circled a group of collocational nouns in each textbook marking the textbook name in big characters. Visualizing the boundaries of noun use helps to

4. I have made a slight alteration to the original graphic scattering of nouns which tend to be plotted at a same point when they have similar values, for the sake of better visibility.

5. Name of birthplace of Kim Sr. which is considered a holy place in North Korea.

identify what kind of nouns co-occur with *wuli* and in which textbooks, and also how those co-occurrences are shared between the four textbooks. For instance, expressions such as "our homeland", "our general" occur only in nk. Ones such as "our relative" occurs only in krj, "our village" in krc and "our neighborhood" in sk. In the same way, the boundary lines show that co-occurrences such as "our forces" or "our army" appear only in nk and krc, meanwhile "our people" or "our brother" appear only in nk and krj. One can infer that between North Korea and China there is a common interest in military affairs while North Korea and the Community of Korean Residents in Japan share a certain "compatriot-ship". Tracking these shared aspects between two or three particular textbooks would probably reveal much about the socio-politico-cultural affinities between those particular communities. In this paper, I focus only on the collocations which are common across the four textbooks.

In Figure 1, the six nouns co-occurring with *wuli* (our) in all textbooks are plotted at the center of the figure. They are "school", "house", "class", "alphabet", "language" and "country". Setting aside "school" and "class", collocations like "our <house, alphabet, language, country>" constitute a common collocational pattern throughout the four textbooks whereas the referents of "our country" and "our house (home)" differ by community and individual. Supposing a sense of co-owning bound to a house (home) or a country is limited to family members or community members, by contrast "alphabet" and "language", i.e. the Korean language, are co-owned by Koreans. It could be inferred that "our language" plays the foundational role of producing a homogeneous feeling among Koreans beyond the particularity of individual belonging.

3.3 The use of *ilbon* (Japan)

If we infer that *wuli mal* (our language) exerts a centripetal force in arousing a sense of "us" in Koreans, one relevant question is "who or what is the other?" since the idea of the other is a necessary complementary component in the construction of "us".

It is common practice in the school textbooks to refer to historic events, often in the form of tales of heroism repulsing the enemy's attack, as constituting "glorious past events". In this sense, looking closely at the use of these proper nouns denoting names of specific people and places is a relevant way to identify the "other". Investigation of the use of proper nouns, mainly names of specific people and places, revealed that between the four textbooks, there is only rare agreement on the model personalities or on the historical events considered as important. But most of the selected "heroes" in the four textbooks have the common

characteristic of anti-Japanese colonialism-resistance. For instance, even in North Korean textbooks, Kim Il Sung's greatness is exclusively based on his anti-Japanese resistance activities.

As with historical figures, there are only few foreign countries' names commonly mentioned. A rare exception is *ilbon* (Japan). Table 5 shows that *ilbon* is the foreign country most frequently occurring throughout the four textbooks. The frequencies may be small and the frequency gap between the four textbooks may be large, but they nonetheless show that Japan is the most-mentioned of the foreign countries in all the textbooks.

Table 5. Frequency of foreign country and region

Country & Region	**textbooks**			
	sk	**nk**	**krc**	**krj**
Japan	11	38	4	12
USA	4	30	1	3
Korea peninsula	5	29	7	28
South Korea	0	2	0	0
Manchuria	0	2	0	0
China	7	0	0	4
Egypt	1	0	4	2
England	6	0	3	0
Greek	0	0	1	0
Germany	1	0	1	0
Spain	0	0	1	0
France	3	0	1	0
Africa	10	0	0	0
South America	0	0	1	0
North Korea	7	0	0	0
Antarctic	5	0	0	0
India	5	0	0	0
Australia	3	0	0	0
Malaysia	2	0	0	0
Russia	1	0	0	0
USSR	0	0	1	0
Italy	1	0	0	0
Chile	1	0	0	0
South Africa	1	0	0	0
Mexico	1	0	0	0
Saudi Arabia	1	0	0	0

*The numbers are normalized figures of the number of occurrences per 100,000 characters.

Figure 2 plots the graphic performing correspondence analysis of the frequency data of Table 5, computing the score for the rows (foreign country or region names) and columns (textbooks). In Figure 2[6], each country or region name lies in the neighborhood of the community's name in which the country's or region's frequency score is prominent. The country or region names which occur only in particular textbooks are plotted near the community's name. The country or region names which occur in more than two textbooks at the same time have their names plotted between the community names. In this case, the distance between the country or region names and the community names is relative to the frequency. Thus even though "Japan" and "USA" commonly occur in the four textbooks, they are plotted near nk because in NK textbooks, as Table 5 shows, these two countries not only have exclusive figures, but they are also mentioned much more frequently than in any other textbooks. Figure 2 provides us with a visual understanding of the openness of each community towards the world. Unsurprisingly perhaps, the South Korean textbook is more international in its outlook than that of the isolationist North Korea.

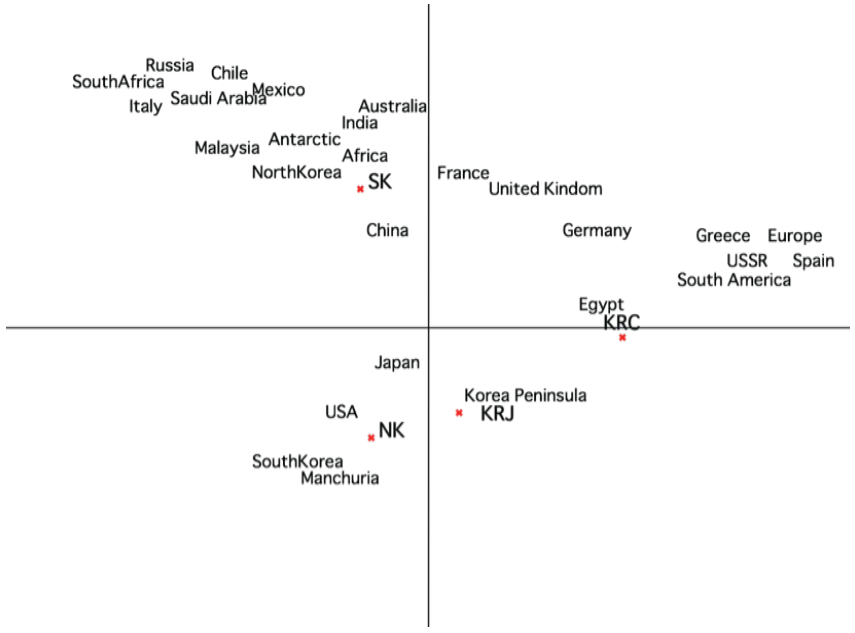


Figure 2. Foreign countries mentioned in the four textbooks

6. Here again, I made a slight alteration to the original graphic scattering of nouns for the sake of better visibility.

As for the two diasporic communities, the textbook of Korean residents in Japan is much poorer in foreign country references than its Chinese counterpart. Given the Japanese community's situation, as I mentioned above, in running an ethnic minority circle as foreigners in Japan, the very concept of "foreign country" would be opaque. Thus it is understandable that the textbooks have little room for discourse on foreign countries.

Further analysis shows that not only is Japan the center of concern throughout the four communities, but also that there is some kind of unanimous feeling in the discourse on this country. On the other hand, "USA" has a varied frequency order and ambivalent meanings across the communities.

In order to investigate the image of Japan in the textbooks, I analysed the word "in its own textual environment" (Sinclair 1991: 32) in terms of its collocational nouns. The following steps were taken:

1. All nouns that occur in the sentences where *ilbon* is mentioned in each community's textbook were collected.
2. Each noun was categorized according to:
   a. historical context: where in the textbooks *ilbon* is mentioned mainly in the background of 3 periods: the Japanese invasion of 1592[7], the colonial period (1910~1945) and the present.
   b. the implied sentiment towards *ilbon* which is conveyed by the story: this I have interpreted as revealing antipathy, neutrality or sympathy towards Japan.
3. The number of nouns of each category were calculated.

For example, suppose that in a particular community's textbooks there is the word "sword" among the words employed in a phrase where *ilbon* occurs. If the historical context of the story is the colonial period and the feeling towards Japan in the story is hostile, the word "sword" is classified as <Colonial, Antipathy> and so on.

Table 6 shows differences of occurrences of nouns used in the discourse on Japan in each community's textbooks (columns) and in each category.

7. The Korean peninsula has suffered sporadic attacks on its coasts by the Japanese. In 1592, Toyotomi Hideyoshi, having succeeded in unifying Japan, launched a series of invasions through Korea against the Ming Empire of China. The two most important campaigns took place in 1592 and 1597. In the course of the invasion, nearly the whole of the Korean peninsula became an arena of Japanese pillage and slaughter, and the animosity of the Korean people towards Japan remained alive long thereafter (Lee 1984).

Table 6. Quantity of nouns used with *ilbon* in each textbook and in each category

period and sentiment	sk	nk	krc	krj	total number	proportion
Colonial, Sympathy	0	0	0	0	0	0%
Colonial, Neutral	0	0	0	1	1	0%
Colonial, Antipathy	32	144	31	12	219	70%
Present, Sympathy	0	0	0	0	0	0%
Present, Neutral	4	0	0	55	59	19%
Present, Antipathy	3	15	0	0	18	6%
Invasion of 16C, Sympathy	0	0	0	0	0	0%
Invasion of 16C, Neutral	0	0	0	0	0	0%
Invasion of 16C, Antipathy	0	6	10	0	16	5%
total number	39	165	41	68	313	
proportion	12%	53%	13%	22%		1

*The numbers are normalized figures of the number of occurrences per 100,000 characters.

Although nine categories could be expected by combination of the elements mentioned above, only five of the categories shown in Table 6 have values. The table shows that sympathy or neutral feelings are almost never expressed in the stories set during the Japanese colonial period, as in stories set during the invasion of the 16th century.

The final two rows of Table 6 show the total number of nouns in each community's textbook and their proportions based on the totality of the four textbooks. They show that the occurrence of nouns used in sentences relating to Japan differs between textbooks, and that in nk textbook there are many more nouns used than in any other textbooks (53% of the overall corpus). It would appear that Japan and the United States of America dominate North Korea's foreign concerns. The final column shows the proportion of noun occurrences in each category. The proportion of nouns in category "Colonial, Antipathy" is overwhelmingly high. Furthermore, Figure 3 illustrates how this category is dominant in most textbooks. It indicates that except for the textbooks of Korean residents in Japan, in all other textbooks, when it comes to talk about Japan, around 80% of any story is based on the colonial period, seemingly implying unpleasant feelings.

No evidence of sympathy can be traced and a neutral view is rather rare. Even in South Korean textbooks, a context which treats present-day Japan neutrally constitutes only about 10% of the references. Moreover in the textbooks of the KRJ community settled in Japan, the feeling towards Japan is "neutral" (80%) at best, but importantly never sympathetic towards their country of residence. *Ilbon* occurs in general in a negative context throughout the four textbooks. It is true that, quantitatively speaking, compared to the data in Table 3, the values of Table 5 are lower, and the frequency of "Japan" significantly varies between the

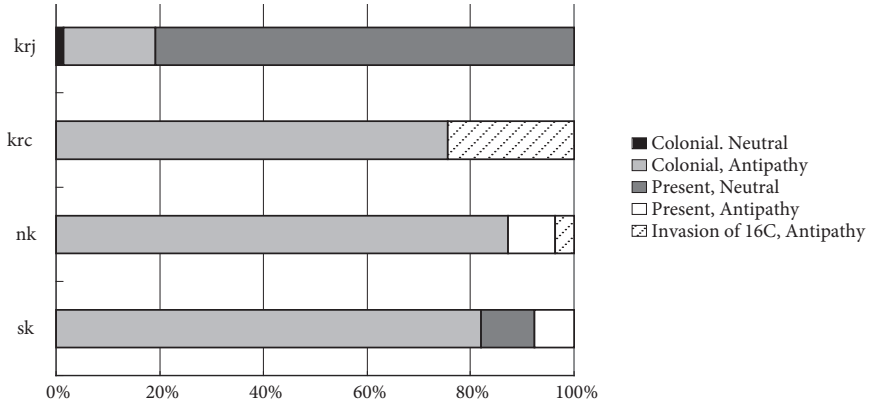


Figure 3. Image of Japan in the textbooks

four textbooks. Nevertheless in the use of *ilbon* a certain consensus among the four textbooks is observed. In general it conveys obvious unpleasant collective memories of past victimization.

4. Conclusion

Today South Koreans, North Koreans, Korean residents in China and Korean residents in Japan, are in socio-culturally and geopolitically different communities. Nevertheless the "Weness" in Koreans' minds persists across socio-political borders. In examining the Korean language textbooks currently used by primary schools in these four communities, I found two particular aspects of language use which are common throughout the four school textbooks. One is frequent collocation of the words *wuli* (we, our) and *mal* (language), thus *wuli mal* or "our language" that occurs in all textbooks as a semi-fixed expression. On the other hand, *ilbon* (Japan) occurs mostly in negative contexts. Crystallizing around these two keywords are the ultimate elements that generate and maintain the "Weness" of Koreans. They can be condensed into two factors: a common language and a common enemy. This "Weness" peculiar to Koreans is therefore internalized and reproduced through school education.

This study of language use touches on socio-political questions that need deeper exploration. Despite the fact that there have been no concerted efforts or normalization across the four communities and despite the obvious socio-cultural differences between them, there are clearly areas which can be identified as being of shared interest or shared preoccupations. Further study on language use of the four communities' textbooks may shed light on many more inter relationships.

References

Apple, M. 2000. *Official Knowledge*. London: Routledge.
Foucault, M. 1969. *L'archéologie du savoir*. Paris: Gallimard.
Fraysse-Kim, S. H. 2006. *School Textbooks and Ideology*. PhD dissertation in National University of Nagoya, Japan.
Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press Inc.
Lee, B. L. 1996. Correspondence Analysis. In *Vista the Visual Statistic System,* F. W. Young (ed). Online document (http://www.uv.es/prodat/ViSta/vistaframes/pdf/chap11.pdf).
Lee, K. 1984. *A new history of Korea*. Seoul: Ilchokak Publishers.
Pêcheux, M. 1975. Mises au point et perspectives à propos de l'analyse automatique du discours. *Langage* 37: 7–80.
Pêcheux, M. 1981. *La Langue Introuvable*. Paris: Maspero.
Pêcheux, M. 1983. *Language, Semantics and Ideology*. London: Macmillan.
Quirk, R. et al. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.
Sinclair, J. McH. 1966. Beginning the Study of Lexis. In *In memory of J. R. Firth*, C. E. Bazel, J. C. Catford & M. A. K. Halliday (eds), 410–429. Harlow: Longman.
Sinclair, J. McH. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.
Stubbs, M. 2001. *Words and Phrases*. Oxford: Blackwell.
Teubert, W. 2007. Parole-linguistics and the diachronic dimension of the discourse. In *Text, Discourse and Corpora,* M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds), 57–87. London: Continuum.
Williams, R. 1976. *Keywords: A vocabulary of culture and society*. Glasgow: Fontana/Macmillan.

General spoken language and school language

Key words and discourse patterns in history textbooks

Paola Leone
Università del Salento, Italy

The current study uses keyness to identify the essential lexicon and lexical patterns of history textbooks and match them to the out-of-school input to which a young learner might be exposed. Two frequency-ordered lists have been compared using the KeyWords component of WordSmith Tools v. 5.0 (Scott 2008). The first, compiled from a corpus of history textbooks used in lower secondary schools in Italy, named CoMaS (Corpus di Manuali di Storia), ascertains words needed to study history at school. The other, compiled from a published corpus of Italian spoken language, called LIP (Lessico di Frequenza dell'Italiano Parlato), specifies words frequent in daily communication. Results show discoursal, lexical, semantic, and morphological features which may be unfamiliar to the learner, and which should therefore be considered in a syllabus designed to develop students' ability to interpret and express historical discourse.

1. Definition and purposes of the research project

The study aims at identifying the Italian vocabulary students need for studying history in lower secondary school and predicting which words might present problems for non-native speakers. This is done by comparing a frequency-ordered word list compiled from a corpus constructed from thematically selected portions of seven history textbooks used in lower secondary schools in Italy (Leone & Grandi 2006) with the statistical occurrence profile of an oral natural communication corpus LIP (De Mauro et al. 1993). The research proposes to:

- explore a language variety used in schools;
- determine what vocabulary is essential for studying history in lower secondary school and reflect on criteria for identifying vocabulary that is likely to pose problems for non-native students.

The study is prompted by the increasing number of immigrant teenagers attending Italian schools.[1] With their parents they have left their native countries (particularly Albania, Morocco, Romania, China) mainly for economic reasons, since the late 1980s. With little or no competence in Italian, they face major learning difficulties.

Despite the significant percentage of non-native speakers (in the North of Italy, 7 out 100 students, according to 2005–2006 Ministero Pubblica Istruzione data), Italian schools do very little to assist these learners. Except in the schools where funding is available for special classes which remove students from mainstream lessons for remedial instruction, bilingual students are expected to develop language literacy advanced enough to study mainstream texts merely by listening to teachers' lessons and through extensive reading, starting from the Italian they have learnt during daily extra-curricular interaction. This is usually insufficient for them to keep pace with their native peers.

The dearth of appropriately designed second language instruction is in part due to the shortage of relevant data on classroom language and the language found in textbooks and learning materials. There is a need for applied linguistic research exploring the features of written text genres employed in schools, and the academic language demands of course-specific texts and study materials and of teachers' spoken discourse. Analysis of relevant corpora, as in the present research, can contribute to filling this gap.

2. Words for studying in a second language

Colombi and Schleppegrell (2002: 8–12) point out that advanced literacy can be reached if students can understand how meanings are construed in different content-area texts and how the concepts of school subjects are expressed in language. It goes without saying that words are fundamental for the construction of meaning: concepts can be expressed ungrammatically, but we cannot express anything verbally without vocabulary. A good knowledge of words in an L2 is relevant for oral and written comprehension, for the production of oral and of written discourse, and for the interaction in the language. Research attests to the positive correlation between vocabulary knowledge and the ability to understand a written text (Beck, Perfetti & McKeown 1982; Kameenui, Carnine & Freschi

1. Number of Immigrant teenagers in Italian schools 2005/2006: 424,683; 2006/2007: 501,445 (Ministero della Pubblica Istruzione, 2008. *Alunni con cittadinanza non italiana.* Online document at URL > http://www.pubblica.istruzione.it/mpi/pubblicazioni/2008/allegati/alunni_n_ita_08.pdf).

1982; Stahl 1983), and experimental studies (Hu & Nation 2000; Laufer & Sim 1985; Laufer 1989) have shown that text comprehension can be predicted by the percentage of known words in the text. Laufer (1989) showed that the knowledge of 95% of words in a text allowed Hebrew and Arabic native speakers to reach a score of at least 55% in a test of reading comprehension in English as L2. Hu and Nation (2000) set the percentage even higher: readers need to know 98% of the words in a text. What we can argue is that a high density of unknown words (i.e. more than one every twenty tokens) may well hinder adequate text comprehension (see also Laufer 1997; Nation 2001: 144–149, for a review).

Although the focus of this study is on vocabulary, lexical knowledge is of course not the only kind of knowledge a learner needs in order to study history, or indeed any school subject. For instance, to appreciate the demands on a reader in understanding a text and assimilating the content well enough to be able to express that information orally or on paper, we must consider the whole reading process and its dynamic and interactive nature, involving strategies at many levels from global to local (Carrell, Devine & Eskey 1998; Brodine 2001). Reading requires not only lexicogrammatical competence but also background knowledge, appropriate cognitive processes, and the ability to apply general reading strategies such as predicting content, guessing unknown words, inferring meanings, recognizing text type and structure, and understanding the main concepts (Laufer 1997). However, when teaching learners to cope with these manifold aspects of reading comprehension, the teacher can benefit from a more extensive awareness of the vocabulary involved.

3. CoMaS, a corpus of history textbooks

The CoMaS corpus (*Corpus di Manuali di Storia*; Leone & Grandi 2006), created by myself, consists of 121,734 tokens (120,038 tokens used for wordlist) and 13,777 types with a type/token ratio 11.48. It provides a sample of the language used in history textbooks for Italian lower secondary schools (ages 11–14).

CoMaS is composed of excerpts from seven lower secondary school textbooks: Gentile and Ronga (2001: vol. I, II, III), Mezzetti (1999: vol. II), and Paolucci & Signorini (1997: vol. I, II, III). Whilst the excerpts written for second-year students come from all these authors, the ones for first- and third-year students include no extracts from Mezzetti (1999), which was out of print during the second phase of our study. Consequently, in its present state the corpus consists mainly of second-year history textbooks (Table 1). Since CoMaS is a collection of extracts selected from the same text type, school textbooks dealing with historical events, it can be considered a specialized corpus (Gavioli 2005: 54–56; Hunston 2002: 14).

Table 1. Tokens and types in CoMaS

	Texts for the first year (n. 2)	**Texts for the second year (n. 3)**	**Texts for the third year (n. 2)**
Authors	Gentile & Ronga Paolucci & Signorini	Gentile & Ronga Mezzetti Paolucci & Signorini	Gentile & Ronga Paolucci& Signorini
Tokens	15,888	83,411	22,435
Types	3,886	11,206	4,573
Topics	The Fall of the Roman Empire, Feudalism	Different topics – History from 1400 to 1800	Fascism, Nazism

As no official data on sales were available from the Italian publishers association (AIE; Associazione Italiana Editori), the selection not only of the most studied topics but also of the most commonly used textbooks used to compile the corpus was based on suggestions made by a group of teachers. These informants included a university professor of history-teaching methodology, who was interviewed, as well as twelve secondary school teachers who work in four different Italian schools, who completed questionnaires.

The relevant passages were scanned and converted to machine-readable corpus format. Because the focus of the research was on the language alone, all images were omitted, as was formatting distinguishing titles and body text. While information-processing influenced by a text's multiple non-verbal dimensions would clearly be significant for an investigation of learners' reading comprehension strategies, for instance, the intention here was limited to selecting words to use in designing a vocabulary programme. As CoMaS is the focus in this article, it will be referred to as the target corpus.

4. The Lexicon of Spoken Italian: LIP

LIP (*Lessico di Frequenza dell'Italiano Parlato;* De Mauro et al. 1993*)* consists of 500,000 tokens (De Mauro et al. 1993: 29). It is currently the only available balanced, general corpus of transcribed spoken Italian which contains a large number of samples of different communicative situations, distributed in carefully planned proportions (five sections of around 100,000 words each). According to De Mauro et al. (1993: 40–41) it includes:

1. free-turn taking in face-to-face conversations (conversations at home; conversations at work; conversations at school and at university, etc.);

2. free-turn taking in non face-to-face conversations (telephone conversations, telephone conversations on the radio, recorded messages on the answering machine);
3. selected turn-taking in face-to-face conversations (legal debates, cultural debates, trade union meetings, labour meetings, primary school oral examinations; secondary school oral examinations, etc.);
4. unidirectional communication in the presence of the interlocutor(s) (primary school lessons, secondary school lessons, university lessons, etc.);
5. unidirectional communication either in distance or in deferred communication of an oral text (TV and radio programs).

LIP is a sample of the language input the learner may be expected to be exposed to outside school. It will be used as a reference corpus in the comparison with CoMaS *to define the distance* between everyday language usages and historical written school discourse, with the aim of underlining the linguistic features of this latter domain.

5. Research questions and methodology

The following questions guided the research:

1. What are the key words of a corpus of Italian history school textbooks (age 11–14) in comparison to a corpus of spoken Italian?
2. What are the differences between language usage in history textbooks and the oral input a young learner is likely to be exposed to outside of school?

The computational examination of the LIP and CoMaS corpora was carried out with the aim of identifying keyness, represented by the words that are unusually frequent in the school textbooks (represented by the target corpus, CoMaS) in comparison with the oral communication lexicon (represented by the reference file, the LIP word-list).

WordSmith Tools v. 5.0 KeyWords tool was used for comparing CoMaS with the LIP frequency-ordered list, selecting the statistical chi-square test (Scott 2000: 109–116; Tribble 2000: 78–84). Secondly, WordSmith concordancer was employed to examine target words in the various contexts in which they occur.

Before the corpora could be analysed, CoMaS and LIP and their respective frequency lists needed to be edited. First of all, the corpus transcriptions had to be made consistent. Accents required special attention, as in both corpora the same word might be found at times with a grave accent and at times with an acute accent (e.g. *perchè, perché*); these items were combined. Besides, because LIP is

a transcribed corpus of the spoken language, it was necessary to tag and subsequently filter out transcription notes (e.g. removing upper-case letters indicating all speakers). Furthermore, in order to compare the frequency-ordered lists, it was necessary to make choices regarding the use of dialect in LIP, employed during unplanned informal oral communicative situations. Specifically, dialectal forms whose pronunciation was dissimilar to the corresponding Italian words were excluded from the frequency list. For instance, while the dialectal *chiamma* was counted as *chiama* (he/she calls; call), *finiscimmo* for the Italian *finiamo* (we finish) was excluded.[2]

Taking into account the aim of making students better readers of history textbooks and also the fusional and clitic nature of the Italian language, the words were not lemmatised. In the reading process, the way words appear in a text is fundamental for comprehension. Hulstijn rightly contends that "Word recognition in reading is a process using orthographic information as its primary basis" (2001: 265). Moreover, like other fusional languages, Italian has very rich and varied inflectional and derivational morphology, so in usage some word-classes can change considerably from their lemma form. For instance, nouns and adjectives can be either masculine or feminine and their final vowel in the plural forms differs according to the gender (i.e. feminine *-a* changes into *-e* and masculine *-o* into *-i*; and there are exceptions as well). Verbs are conjugated differently according to mood, tense and aspect and in most tenses also vary according to the subject. *Minacciarono* (they threatened) is the perfective form, called "passato remoto", composed of the root *minacci* (*-are*) plus (*-arono*) the morpheme for the third plural person. Furthermore, Italian uses cliticization, which means for instance, that a verb such as *abbandonandone* is composed by the root *abbandon* (*-are*), fused with the morpheme of the gerund (*-ando*) plus the clitic pronoun *ne*. If we take into account all these aspects of the language, it is clear that lemmatization would have excluded word features that might well be influential in the comprehension process.

2. In order to compare the CoMaS and LIP frequency lists, the "case sensitivity option" of the software was not selected. The CoMaS frequency list was created without preserving the distinction between capitalized and uncapitalized words (i.e. after full-stop punctuation) and different typographical upper case fonts for headings, so that it could be compared with LIP, which was transcribed without using capital letters. As a consequence, the frequency lists do not take into account the difference between homographs such as common and proper nouns (i.e. monaco = *monk* and Monaco = *Munich*) nor do they indicate when the same word forms belong to different word classes.

6. The key word-list

In order to shed light on the linguistic features of historical discourse, the focus will be primarily on positive key words.

Table 2 shows the top 100 positive key words. Heading the list, the symbol "#", used by the software to replace all numbers cumulatively, shows, not surprisingly, that numbers, including dates, are frequent in history textbooks when compared with spoken language.

Table 2. Top 100 key words from the comparison of CoMaS and LIP

1	#	35	ESERCITO	69	ROMA
2	FU	36	GOVERNO	70	SPAGNA
3	ERANO	37	VENNERO	71	AUSTRIA
4	I	38	AD	72	STATI
5	FURONO	39	EBREI	73	CIVILTÀ
6	NEL	40	ITALIANI	74	COSTANTINO
7	IMPERO	41	VENEZIA	75	MEDIOEVO
8	ERA	42	INOLTRE	76	DEGLI
9	ITALIA	43	GRANDI	77	INGHILTERRA
10	DEL	44	POPOLO	78	TEDESCHI
11	VENNE	45	ESSI	79	III
12	SECOLO	46	CRISI	80	PARTITO
13	FRANCIA	47	MOLTI	81	SUO
14	CITTÀ	48	EGLI	82	DAL
15	IL	49	LORO	83	CONTRO
16	CHIESA	50	REPUBBLICA	84	EBBE
17	HITLER	51	FECE	85	NELLE
18	GERMANIA	52	GLI	86	MEDICI
19	RE	53	PACE	87	ARMI
20	PAPA	54	RIVOLUZIONE	88	UOMINI
21	LA	55	ROMANO	89	CASTELLO
22	MUSSOLINI	56	GUERRE	90	SUOI
23	IMPERATORE	57	EUROPEA	91	TRA
24	GUERRA	58	AVEVANO	92	VOLEVANO
25	DELL	59	VESCOVI	93	OCCIDENTE
26	DI	60	TERRITORI	94	VII
27	DEI	61	NUOVA	95	CARLO
28	DELLA	62	RINASCIMENTO	96	NAPOLEONE
29	DIVENNE	63	CONTADINI	97	PERCIÒ
30	REGNO	64	L	98	POPOLI
31	LE	65	FASCISTA	99	ALLA
32	AVEVA	66	INDIPENDENZA	100	DELLO
33	FASCISMO	67	AI		
34	POTERE	68	CONQUISTA		

The key word list highlights top frequency words such as:

- Link verbs/auxiliaries and full content verbs (e.g. *fu*, 'was'; *erano*, 'were'; *furono*, 'had been'; *divenne*, 'became'; *volevano*, 'wanted');
- Definite and indefinite articles (e.g. *la*, *il*, *i*; *un*, *uno*);
- Prepositions (e.g. *di*, *in*) and prepositions + definite articles (e.g. *del*, *della*);
- Proper names of countries (e.g. *Francia*, 'France');
- Common nouns (e.g. *Impero*, 'Empire');
- Qualitative adjectives which identify size/relevance (e.g. *grandi*, 'big/great') and the result of a changing process (e.g. *nuova*, 'new');
- Linking adjuncts (e.g. *inoltre*, 'besides').

We will discuss the lexicogrammatical features of some verbs, nouns and linking adjuncts to identify their function in construing the history textbook genre.

6.1 Framing the discourse: Time and transformation in historical narrative

The key word list calls attention to the fact that the "passato remoto" tense in the third person singular and plural (i.e *fu*, *furono*, *divenne*, *ebbe*; respectively, n. 2, 5, 29, and 84 in Table 2; 'was', 'were', 'became', 'had') is far more frequent in CoMaS, compared with the spoken varieties in LIP. This statistical finding is not surprising since in Italian oral communication the "passato remoto" is used only in Sicily: elsewhere in Italy the "passato prossimo" is preferred for such contexts.

While the third person singular and plural of the "passato remoto" of *essere*, main verb or auxiliary, proved to be exceptionally frequent in CoMaS, the verb *venire* appears as a key word only as *venne*, the third person singular of this tense. Further examination of the entire list shows that the comparison does not identify *vennero*, third person plural, as a key word. The concordancer reveals that *vennero* does indeed occur in the spoken language, but used as a full content verb, meaning *to come,* whereas in CoMaS it occurs only as an auxiliary on the CoMaS frequency-ordered list.

Divenne ('became') is a key word, evidence confirming that for phasing time "the concepts of change and continuity are central to meaning-making in history" (Coffin 2000: 149). History in discourse is represented as a cycle of events that have an inception, a duration and an end. Whilst the central phase, duration, is associated with continuity, the initial and the final stages of a process are related to transformations (Coffin 2000: 149), thus justifying the high occurrence of *divenne* ('became').

Predictably for a genre that needs to set time in the past, *secolo* ('century': n. 12 on the key word list) is a technical term prominently employed to discuss events that happened in a period of a hundred years (e.g. *la crisi del III secolo*,

'third century crisis'). Together with *day*, *month*, *year* and other expressions used when referring to important events such as the *French Revolution* or the *Middle Ages*, they build the time frame of narrative historical discourse (Coffin 2000: 147–149).

6.2 Causality and interpretation of events

A major feature of history texts is the narration, interpretation, and explanation of events, in which causal relations between past actions play an important role (Coffin 2000). In this regard, two items in the 100 top key word list, *perciò* ('therefore') and *inoltre* ('besides'), will serve to illustrate how linking expressions relate to this aspect of historical discourse, as well as some potential interpretation problems which they may pose for the young reader, and which the teacher needs to be aware of.

Perciò (n. 97 in Table 2; 'therefore; for this reason') constructs a causal framework by introducing either a result or a conclusion to some previous fact or event. Literally speaking, this conjunction builds the cause-effect pattern by means of anaphoric reference (*per* + *ciò*, 'for + that'), as can be seen in, "*La Germania venne considerata nei trattati di pace come unica responsabile della guerra. Fu sottoposta perciò a condizioni punitive…*" ('In the peace treaties Germany was considered the only country to blame for the war. For that reason it was subjected to punitive conditions…'). This need to identify the anaphor's antecedent complicates the reader's comprehension task more than, for instance, the causal conjunction *perché* ('because').

Inoltre (n. 42 in Table 2; 'besides, furthermore') links a discourse sequence to what was stated before, adding new events and sometimes emphasizing a further point considered relevant by the writer. It is thus a word that occasionally frames narration and argumentation in historical discourse in such a way as to veil the author's personal view point (e.g. *Nelle osterie spesso nascevano delle risse, e tra i nobili erano frequenti i duelli. Inoltre, a dispetto dei manuali di buone maniere, gli uomini erano grossolani, sporchi e scurrili*, 'In taverns people often used to brawl, and among nobles duels were frequent. Besides, in spite of good manners described in treatises, men were coarse, dirty and scurrilous').

6.3 Noun phrases

Di (of) and *di* + "definite article" are outstandingly frequent in the written corpus. Whilst *di* (n. 26 in Table 2) can be followed either by a noun group (i.e. *i sistemi di colonizzazione*; 'the systems of colonization') or by an infinitive form of a verb

(i.e. *molte probabilità di essere eletto*; 'good chances of being nominated'), *del, dell, di, dei, della* (namely n. 10, 25, 27, 28 in Table 2; 'of the') precede a noun. The high incidence of this preposition points to the pervasiveness of noun phrases with a strong tendency to post-modification. This linguistic characteristic, mentioned by Sinclair (1991; see also Scott & Tribble 2006: 99), is even more extensive in Italian than in English, since Italian does not permit the adjectival use of nouns as in the example above, *i sistemi di colonizzazione* 'colonization systems'. The recurrent presence of *di* also correlates with the high lexical density, i.e., "the much higher ratio of lexical items to total running words" (Halliday 1989: 61), and the subsequent compactness of information of written texts compared with a spoken corpus (Halliday 1989: 61–75).

6.4 "Aboutness"

Content words shed light on what is important in a text and on what is "indicative of its meaning, what it is about" (Scott & Tribble 2006: 58), namely, its "aboutness". In the history corpus, besides the time references like "century" mentioned above, key nouns relate to: (a) political or spiritual authorities (e.g. *re, imperatore* n. 19 and 23 in Table 2; 'king', 'emperor'); (b) political entities of government (e.g. *impero* n. 7 in Table 2; *empire*); (c) geographical spaces (e.g. *città* n. 14 in Table 2; 'town'); (d) conflicts (e.g. *guerra* n. 'war', 'army'). Proper names of geographical entities, people or historical periods (e.g. *Italia*, *Francia*, *Hitler*, *Rinascimento*; 'Italy', 'France', Hitler, 'Renaissance') are also recurrent. These findings confirm the prevalence of semantic fields relating to conflicts, government, places and time.

It is also important to notice the use of the word *città* (n. 14 in Table 2; 'town/towns'), which is rare in spoken language. This word might belong to the "highly available vocabulary", which includes words that are not frequent in the input but that are well known for their utility in everyday life.

7. Limitations of the study

This study was carried out by constructing a new text collection and comparing it with an existing corpus. The research methodology did not involve a tagging process, and comparison was simply between the word forms as they occurred in the two corpora, a choice made necessary because there exists no automated annotation software for Italian that would work both with written and spoken corpora. The possible alternative of manual tagging was rejected as overly time-consuming, at least for the present. This means that at this stage the frequency-ordered lists and key word list do not include information on the function of words in the two text varieties (e.g. *era* as link verb or as auxiliary) and on different word cat-

egories in homographs (see Note 2). Although the general profile of these findings calls for fuller investigation to be carried out using a concordancer, I would argue that the present school contexts – where the increasing presence of non-native speakers has intensified the teachers' need for immediate specific training to meet their needs – can still benefit from the information yielded by this and similar future research, in spite of some limitations.

One limitation of the study is corpus size: the corpus was too small to allow in-depth scrutiny of differences in an item's meanings, collocates or phraseology in similar contexts in historical discourse (Hunston 2002: 42–45). Still, the data furnishes a general idea of language usage in historical written school discourse, and paves the way for the more detailed study of word patterns and meanings which will emerge when other texts can be added to CoMaS. Corpus interrogation has already revealed information that takes us beyond mere intuition. For instance, the use of *because, therefore* and *besides* to build causal relationships and interpreting events, rather than other more complex phraseology, could only have come to light through empirical data inspection. Likewise, analysis of concordances highlights the use of *vennero* ('were; came') as an auxiliary, instead of as a full meaning verb as in spoken language. Since the aim of such studies is to develop a syllabus for teaching the content area language to non-native speakers and to increase teachers' awareness of some features of written historical discourse, even the limited number of instances provided by CoMaS are enough to illustrate clearly some significant characteristics and suggest how to search for others.

Whereas CoMaS was constructed specifically for this study, the reference corpus, LIP, was employed because it is the only balanced publicly available spoken corpus. Being compiled from oral monologues and dialogues that occur mainly in natural adult contexts, its language profile is arguably too restrictive for the purposes of this study, since it does not fully mirror the input that a teenager is exposed to. It would therefore be advantageous for future research to develop the Young LIP corpus using selected data from LIP in its present published form and adding new discourse samples. Until that time, we can take comfort from Scott & Tribble's claim (2006: 65): "In our experience, even the use of a clearly inappropriate reference corpus as in the case of the BNC for studying a Shakespeare play may well suggest useful items to chase up using the concordancer."

8. Conclusions

The language of history textbooks presents discoursal, lexical, semantic, and morphological features which may be unfamiliar to the learner and on which the teacher may need to focus in order to develop students' ability to understand and then use them. Expressions and verb forms are employed to sequence events in

time and to represent changes and transformations; adverbs and linking adjuncts, on the other hand, construe causality and in some contexts serve to show historical event interpretations.

The wide variety of genres that characterize historical discourse, such as autobiographies, biographies, factual accounts, expository, explanatory and evaluative texts (Oteíza 2003: 641), draws on a broad vocabulary ranging from everyday words to the language required to convey complex content and a high level of abstraction.

Unlike some other classroom subjects, the vocabulary of history is not monosemic, and hence is less "precise" than the vocabulary employed, for instance, in maths or specialized scientific discourse. One source of ambiguity is the fact that words in history textbooks occur in other contexts as well, including everyday communication (see examples discussed above), including terms referring to political or spiritual authorities (*king*, *emperor*) or political entities of government (*empire*).

> The language of history as a school subject (which is not, of course, to be confused with the language of historical research designed for specialists) is the least new and the most overused of all scholastic varieties. Each different meaning has been slowly laid down, one on top of the other, resulting in an almost "geological" stratification of meanings. (Deon 1986: 185, my translation)

Hence, historical discourse includes words that, being "the least removed from common human experience" (Edwards 1978: 55 in Coffin 2000: 15), are often "overused" in everyday life (Deon 1997: 44), acquiring new meanings in different social and historical contexts, what Deon terms "constituting semantic layers".

As a result, content-word meanings can be especially challenging to teach in a multilingual classroom. Authors of history textbooks should bear in mind the problems that may be created by terms whose meanings are not those current in the student's time or cultural context. What does "government" mean to a young apprentice historian starting to learn about the unification of Italy? What does "town" mean to a young learner from China? An intercultural perspective is essential when dealing with a lexis that entails diverse social and cultural meanings.

Further investigation of key words, including research using a concordancer, is needed in order to understand more fully their most common meanings, functions and patterns in different contexts. Concordancers can for instance provide such information as the typical contexts of common grammatical words or the word clusters occurring regularly in the semantic fields common in history textbooks, and compare these findings with everyday language. An in-depth analysis of the key words and their use in the two different varieties could be developed and then employed for purposes such as the following:

- developing teaching materials suitable for preparing minority students to cope with mainstream texts. The syllabus should focus on the function and content words most used in the school textbooks and on the word patterns they generate;
- training history teachers who work with multilingual classes, and teachers of Italian as L1 and L2. The comparison of oral data with data compiled from textbooks would enhance teachers' ability to reflect on the lexis used in the school context and see how it differs from the language employed for daily communication;
- preparing diagnostic tests to evaluate students' vocabulary knowledge, assessing the degree to which they have already acquired the words in the reading materials they will be expected to use (Have they *heard* the words "century" and "government"? Do they know the *definitions*? Can they *use* them?).

Ultimately, this research aims at allowing young immigrant students to feel more actively involved in school learning processes. However, I strongly believe that focussing attention on the teaching process and on the language it involves can result in benefits for all the students, native speakers included.

References

Beck, I. L., Perfetti, C. A. & McKeown, M. G. 1982. Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational* 74(4): 506–521.

Brodine, R. 2001. Introducing corpus work into an academic reading course. In *Learning with Corpora*, G. Aston (ed.), 138–176. Bologna: CLUEB.

Colombi, M. C. & Schleppegrell, M. J. 2002. Theory and practice in the development of advanced literacy. In *Developing advanced literacy in first and second languages: Meaning with power*, M. J. Schleppegrell & M. C. Colombi (eds), 1–19. Mahwah NJ: Lawrence Erlbaum.

Carrell, P. L., Devine, J. & Eskey, E. D. 1998. *Interactive approaches to second language reading*. Cambridge: Cambridge University Press.

Coffin, C. 2000. History as Discourse: Construals of Time, Cause and Appraisal. PhD dissertation, University of Edinburgh. Online document at URL > http://www.library.unsw.edu.au/~thesis/adt-NUN/uploads/approved/adt-NUN20010920.110615/public/01front.pdf.

De Mauro, T., Mancini, F., Vedovelli, M. & Voghera, M. 1993. *Lessico di frequenza dell'italiano parlato.* Milano: Etas Libri.

Deon, V. 1986. Analisi linguistica di alcuni manuali di storia per la scuola dell'obbligo. In *Prospettive didattiche della linguistica del testo*, S. Cargnel, G. F. Colmelet & V. Deon (eds), 183– 204. Firenze: La Nuova Italia.

Deon, V. 1997. Il manuale di storia fra divulgazione, parafrasi e storia generale. In *Il testo fa scuola. Libri di testo, linguaggi ed educazione linguistica*, R. Calò & S. Ferreri (eds), 41–60. Firenze: La Nuova Italia.

Gavioli, L. 2005. *Exploring corpora for ESP learning.* Amsterdam: John Benjamins.
Halliday, M. A. K. 1989. *Spoken and written language*. Oxford: Oxford University Press.
Hu, M. H. & Nation, P. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1): 403–430.
Hulstijn, J. H. 2001. Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehersal and automaticity. In *Cognition and Second Language Instruction*, P. Robinson (ed.), 258–286. Cambridge: Cambridge University Press.
Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.
Kameenui, E. J., Carnine, D. & Freschi, R. 1982. Effects of text construction and instructional procedures for teaching word meanings on comprehension and recall. *Reading Research Quarterly* 17(3): 367–388.
Laufer, B. 1989. What percentage of text-lexis is essential for comprehension? In *Special Language: From Humans Thinking to Thinking Machines,* C. Lauren & M. Nordman (eds), 316–323. Clevedon: Multilingual Matters.
Laufer, B. 1997. The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In *Second Language Vocabulary Acquisition. A rationale for pedagogy*, J. Coady & T. Hucking (eds), 20–34. Cambridge: Cambridge University Press.
Laufer, B. & Sim, D. D. 1985. Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals* 18(5): 405–411.
Leone, P. & Grandi, G. 2006. Le parole della storia e la comprensione del testo scolastico in italiano L2. *ITALS* IV (10): 69–92.
Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
Oteíza, T. 2003. How contemporary history is presented in Chilean middle school textbooks. *Discourse and Society* 14, 639–660.
Scott, M. 2008. *WordSmith Tools*. Version 5. Liverpool: Lexical Analysis Software Ltd.
Scott, M. 2000. Focusing on the text and its key words. In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard & T. McEnery (eds), Papers from the Third International Conference on Teaching and Language Corpora, 103–121. Frankfurt: Peter Lang.
Scott, M. & Tribble, C. 2006. *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam/Philadelphia: Benjamins.
Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Stahl, S. 1983. Differentiating word knowledge and reading comprehension. *Journal of Reading Behavior* 15(4): 33–50.
Tribble, C. 2000. Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard & T. McEnery (eds), 75–90, Papers from the Third International Conference on Teaching and Language Corpora. Frankfurt: Peter Lang.

School textbooks
Gentile, G. & Ronga, L. 2001. *Il Multilibro di Storia*, Editrice La Scuola, Brescia (vol. I, II, III).
Mezzetti, G. 1999. *La Storia e l'Ambiente*. Firenze: La Nuova Italia (vol. II).
Paolucci, S. & Signorini, G. 1997. *Il corso della storia. Dal Rinascimento alla fine dell'Ottocento*. Zanichelli: Bologna (vol. I, II, III).

Index

In the series *Studies in Corpus Linguistics (SCL)* the following titles have been published thus far or are scheduled for publication:

43 **PHILIP, Gill:** Colouring Meaning. Collocation and connotation in figurative language. *Expected March 2011*

42 **MINDT, Ilka:** Adjective Complementation. An empirical analysis of adjectives followed by *that*-clauses. *Expected February 2011*

41 **BONDI, Marina and Mike SCOTT (eds.):** Keyness in Texts. 2010. vi, 251 pp.

40 **PARODI, Giovanni (ed.):** Academic and Professional Discourse Genres in Spanish. 2010. xii, 255 pp.

39 **GILQUIN, Gaëtanelle:** Corpus, Cognition and Causative Constructions. 2010. xvii, 326 pp.

38 **MURPHY, Bróna:** Corpus and Sociolinguistics. Investigating age and gender in female talk. 2010. xviii, 231 pp.

37 **BALASUBRAMANIAN, Chandrika:** Register Variation in Indian English. 2009. xviii, 284 pp.

36 **QUAGLIO, Paulo:** Television Dialogue. The sitcom *Friends* vs. natural conversation. 2009. xiii, 165 pp.

35 **RÖMER, Ute and Rainer SCHULZE (eds.):** Exploring the Lexis–Grammar Interface. 2009. vi, 321 pp.

34 **FRIGINAL, Eric:** The Language of Outsourced Call Centers. A corpus-based study of cross-cultural interaction. 2009. xxii, 319 pp.

33 **AIJMER, Karin (ed.):** Corpora and Language Teaching. 2009. viii, 232 pp.

32 **CHENG, Winnie, Chris GREAVES and Martin WARREN:** A Corpus-driven Study of Discourse Intonation. The Hong Kong Corpus of Spoken English (Prosodic). 2008. xi, 325 pp. (incl. CD-Rom).

31 **ÄDEL, Annelie and Randi REPPEN (eds.):** Corpora and Discourse. The challenges of different settings. 2008. vi, 295 pp.

30 **ADOLPHS, Svenja:** Corpus and Context. Investigating pragmatic functions in spoken discourse. 2008. xi, 151 pp.

29 **FLOWERDEW, Lynne:** Corpus-based Analyses of the Problem–Solution Pattern. A phraseological approach. 2008. xi, 179 pp.

28 **BIBER, Douglas, Ulla CONNOR and Thomas A. UPTON:** Discourse on the Move. Using corpus analysis to describe discourse structure. 2007. xii, 290 pp.

27 **SCHNEIDER, Stefan:** Reduced Parenthetical Clauses as Mitigators. A corpus study of spoken French, Italian and Spanish. 2007. xiv, 237 pp.

26 **JOHANSSON, Stig:** Seeing through Multilingual Corpora. On the use of corpora in contrastive studies. 2007. xxii, 355 pp.

25 **SINCLAIR, John McH. and Anna MAURANEN:** Linear Unit Grammar. Integrating speech and writing. 2006. xxii, 185 pp.

24 **ÄDEL, Annelie:** Metadiscourse in L1 and L2 English. 2006. x, 243 pp.

23 **BIBER, Douglas:** University Language. A corpus-based study of spoken and written registers. 2006. viii, 261 pp.

22 **SCOTT, Mike and Christopher TRIBBLE:** Textual Patterns. Key words and corpus analysis in language education. 2006. x, 203 pp.

21 **GAVIOLI, Laura:** Exploring Corpora for ESP Learning. 2005. xi, 176 pp.

20 **MAHLBERG, Michaela:** English General Nouns. A corpus theoretical approach. 2005. x, 206 pp.

19 **TOGNINI-BONELLI, Elena and Gabriella DEL LUNGO CAMICIOTTI (eds.):** Strategies in Academic Discourse. 2005. xii, 212 pp.

18 **RÖMER, Ute:** Progressives, Patterns, Pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics. 2005. xiv + 328 pp.

17 **ASTON, Guy, Silvia BERNARDINI and Dominic STEWART (eds.):** Corpora and Language Learners. 2004. vi, 312 pp.

16 **CONNOR, Ulla and Thomas A. UPTON (eds.):** Discourse in the Professions. Perspectives from corpus linguistics. 2004. vi, 334 pp.

15 **CRESTI, Emanuela and Massimo MONEGLIA (eds.):** C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages. 2005. xviii, 304 pp. (incl. DVD).

14 **NESSELHAUF, Nadja:** Collocations in a Learner Corpus. 2005. xii, 332 pp.

13 **LINDQUIST, Hans and Christian MAIR (eds.):** Corpus Approaches to Grammaticalization in English. 2004. xiv, 265 pp.

12 **SINCLAIR, John McH. (ed.):** How to Use Corpora in Language Teaching. 2004. viii, 308 pp.

11 **BARNBROOK, Geoff:** Defining Language. A local grammar of definition sentences. 2002. xvi, 281 pp.

10 **AIJMER, Karin:** English Discourse Particles. Evidence from a corpus. 2002. xvi, 299 pp.

9 **REPPEN, Randi, Susan M. FITZMAURICE and Douglas BIBER (eds.):** Using Corpora to Explore Linguistic Variation. 2002. xii, 275 pp.

8 **STENSTRÖM, Anna-Brita, Gisle ANDERSEN and Ingrid Kristine HASUND:** Trends in Teenage Talk. Corpus compilation, analysis and findings. 2002. xii, 229 pp.

7 **ALTENBERG, Bengt and Sylviane GRANGER (eds.):** Lexis in Contrast. Corpus-based approaches. 2002. x, 339 pp.

6 **TOGNINI-BONELLI, Elena:** Corpus Linguistics at Work. 2001. xii, 224 pp.

5 **GHADESSY, Mohsen, Alex HENRY and Robert L. ROSEBERRY (eds.):** Small Corpus Studies and ELT. Theory and practice. 2001. xxiv, 420 pp.

4 **HUNSTON, Susan and Gill FRANCIS:** Pattern Grammar. A corpus-driven approach to the lexical grammar of English. 2000. xiv, 288 pp.

3 **BOTLEY, Simon Philip and Tony McENERY (eds.):** Corpus-based and Computational Approaches to Discourse Anaphora. 2000. vi, 258 pp.

2 **PARTINGTON, Alan:** Patterns and Meanings. Using corpora for English language research and teaching. 1998. x, 158 pp.

1 **PEARSON, Jennifer:** Terms in Context. 1998. xii, 246 pp.