Chapter 25. Frequency and corpora*

Péter Rácz, Viktória Papp, Jennifer Hay
New Zealand Institute of Language, Brain and Behaviour

This chapter considers the role of the linguistic corpus in morphology. We review the ways corpora are typically used, the inherent challenges of corpus-based morphology in synchronic as well as diachronic work, and the insights corpus linguistics gives to morphological theory. Work on corpora often explores the morphological relevance of frequency, and so results and challenges in this domain are highlighted.

1 What is a Corpus?

Empirical work in morphology draws its data from three primary sources: corpora, experimental results, and grammaticality judgements. An important attribute of a word form that is best approximated through the use of corpora (as opposed to the other two sources) is its pattern of usage: its frequency and distribution of use. This chapter gives an overview of what a linguistic corpus is, how it can be used to assess a word form's frequency and distributions, and how these are relevant to morphological theory.

As Manning & Schütze (1999) note, the loosest interpretation of a corpus can be found in statistical natural language processing, where *corpus* refers to a set of data from a certain domain of interest. In this sense, a collection of saccades recorded with an eye tracker, like the Dundee Corpus (Kennedy et al., 2003), is also a corpus, even if not a linguistic one.

Kilgarriff & Grefenstette (2003) offer a slightly more restrictive interpretation of a corpus as 'a collection of texts when considered as an object of language or literary study' (p. 334.). Kilgarriff and others, such as Zséder et al. (2012), argue that any large collection of text (often gathered from the web) can be a valid subject of linguistic research because its size, along with the available computational and statistical methods, makes up for its lack of sampling and control. Gries & Newman (to appear) add, however, that a prototypical corpus is expected to meet further requirements. It needs to be machine-readable to allow researchers a fast extraction of the patterns they are looking for. It should also be representative of a certain dialect, sociolect, or register, and it has to be balanced so that the aspects of the dialect (and other relevant factors) are sampled equally and thus weigh equally in an analysis. And while a corpus is often defined as 'collection of texts', it is important to note that 'text' is interpreted fairly broadly. Both the medium (including audio or video recordings) and the register or genre of a corpus (spanning from transcripts of informal interviews to literary texts) can vary (cf. Gries to appear(a)).

Kučera & Francis (1967) collected the Brown University Standard Corpus of Present-Day American English during the course of the 1960s. This pioneering work of corpus linguistics is one of the earliest examples of what Gries calls a prototypical corpus, with a sample of written American English containing one million words from five hundred samples in fifteen genres. The

*The authors would like to thank Márton Sóskuthy, Sascha Wolfer, and Martin Hilpert for their help.

arrival of corpora like the Brown Corpus, and others that followed, has grounded a significant bulk of empirical morphological research on the link between frequency, morphological structure and productivity.

It is usually argued that if we use large enough corpora, like CELEX (Baayen et al., 1993) or the British National Corpus, the frequency of most forms approximates their probability in everyday use. While this is probably true for more frequent word forms, one has to take note of the genre and the sources of a corpus before relying on it for distributional data. For example, different communities of practice can have different vocabularies with markedly different frequency patterns. Automatically collected, web-based corpora are filtered using spell-checkers, which eliminate most word-level variation, and corpora composed of sociolinguistic interviews will overuse the past tense because interviewees are prompted to talk about past events. (A detailed overview can be found in Gries 2011, Gries to appear(b) and Gries & Newman to appear.)

Corpora tend to build on written sources, and the written genre is notably different from speech in many respects, such as sentence length, the use of grammatical functions like present perfect or the passive voice, and, most importantly, vocabulary use. As Connine et al. (1990) have shown, frequency effects can be medium-specific, leading us to erroneous conclusions on the processing time of a word in an audio task if it has a larger frequency in written than in spoken language. Furthermore, not everyone shares the same frequency distributions. Auer et al. (2000) show that processing patterns are notably different for hearing and deaf people, which they attribute to the difference in language experience. Walker & Hay (2011) find a correlation between whether a word is typically used by older or younger speakers and its processing speed when produced by an older or a younger speaker. Words typically used by younger speakers are processed faster if they are heard as spoken by a younger person and *vice versa*. This shows that both the background of the individual and the context of the utterance have an effect on frequency patterns in lexical processing.

It is clear that special care needs to be taken when assuming that frequencies collected from a particular corpus generalise to a particular speaker or speaker population that the researcher is trying to understand.

The new-found interest in linguistic corpora, whatever form they may assume, can partly be linked to the emergence of *usage-based functionalism* in theoretical linguistics. Proponents of this school argue that patterns of usage are crucial to understanding representation. As such, patterns which are only observable in the *parole*, such as the frequency or probability of a linguistic unit, are central to an adequate description of its behaviour (cf. e.g. Hay et al. 2003 for a summary). Morphemes have been shown to have different phonological patterning, to resist morphological change to a different degree, to be produced and processed more easily and to display a different degree of productivity, all depending on their frequency of use (cf. e. g. Gries to appear(a) for a summary). Usage-based functionalism, in turn, is also heavily influenced by the last thirty years of research in psychology and categorisation theory. Understanding the link between the frequency and distributions of a morphological pattern in a particular corpus on the one hand, and its familiarity or productivity (as measured by psycholinguistic experiments, for example) on the other hand, is far from trivial. While many are agreed that patterns and frequencies of usage are important, there are many ways in which the notion of frequency can be construed and measured.

2 What is Frequency?

It may, on the surface, seem unnecessary to devote a small section of this chapter to discussing what we actually mean when we use the word 'frequency'. Surely frequency is simply a count of occurrence of how many times an object of interest appears in a corpus? There are, however, many different types of 'objects of interest' that can be relevant to morphological work. It therefore seems sensible to delineate some different types of frequency that have been discussed and/or deemed relevant in the literature.

Take a simple word-form such as *unzipping*. There are a number of different frequencies that can be extracted from a corpus which may be relevant to understanding this word's behaviour. The number of times that the whole word is represented – the **word-form frequency**, or **surface frequency** – is the frequency of *unzipping* in the corpus. This is sometimes taken as related to the strength of the representation of this morphologically complex word in the mental lexicon. The **lemma frequency** is the frequency of the 'dictionary' form, collapsing together the various inflectional variants. Thus, the lemma frequency would be the combined frequency of *unzip, unzips, unzipping,* and *unzipped*. This too, is argued to relate to representation strength. Whether lemma frequency or word-form frequency is more important depends to some degree on one's theory about storage of inflected forms. However, experimental work suggests that at least some inflected forms are stored (Alegre & Gordon, 1999).

The word form is made up of a root (*zip*), and two affixes (*un-* and *-ing*). Each of these have their own frequency in the corpus, and each of these frequencies are relevant to the word's processing, storage, and the productivity of its parts. The frequency of an affix can be assessed by either *types* or *tokens*. The number of types is the number of observed distinct words containing that affix. *Un-* for example, occurs in many different types: *untie, undo, unweave* etc. The number of different distinct words is the **type frequency**. Each of these types occurs with different frequencies. *Undo*, for example, is quite frequent, whereas *unweave* is likely to be encountered very infrequently. The total number of observations for *un-* is its **token frequency**. The token frequency will be contributed to much more by frequent words than infrequent words. As will be outlined below, the relationship between affix type and token frequency is an important consideration in assessing morphological productivity. Of course, not all morphological processes involve affixation, and any other morphological patterns can also be assessed in terms of **pattern frequency**, both by types and by tokens. Any type that is observed just once in a corpus is known as a **hapax legomena**, and such forms have special status for the assessment of morphological productivity (Baayen, 1992b).

In a bimorphemic word such as *unzip*, it has been argued that the **relative frequency** between the surface frequency and the **base frequency** (*zip*) is also relevant, with forms where the base is significantly less frequent than the whole word less likely to remain robustly decomposed (Hay, 2001).

The base word, *zip*, also occurs in other words and compounds such as *zipper* and *ziplock*, which together form a word family. The number of distinct types in the family is known as the **word family size** (de Jong et al., 2000). The total number of observed tokens is known as **word family frequency** (de Jong et al., 2000). Other groups of relevant words are also sometimes quantified, such as the size of 'similar' groups of words, in the form of lexical neighbourhoods or lexical gangs.

Frequencies are extracted from corpora, and typically assumed to represent human experience in some form. The corpus count is generally regarded to be a measure of **objective frequency**. This is related to, but is not the same as **subjective frequency**, which is the intuition that an individual has about the relative frequency or familiarity of words or other linguistic objects (see,

e.g. Balota et al. 2001).

Some morphological work has also considered the frequency of phonological objects. This is particularly true of the phonemes near a morphological boundary. *Unzip*, for example, contains a low frequency phonological sequence *-nz-*. Such transitions have been argued to work as a decomposition cue, which helps reinforce the morphological complexity of a word like *unzip* (Hay, 2000). The consideration of frequencies of phonological sequences is known is **phonotactic frequency**. For phonotactics, it appears to be type frequency, rather than token frequency, which provides the most important metric.

A general methodological practice when using frequencies is to use **logarithmic frequency** instead of raw frequency. Baayen (2000) stresses that this is closer to how humans are likely to perceive frequency (with differences between lower frequencies appearing more salient, so that one pineapple versus five pineapples is more salient than 1001 pineapples versus 1005 pineapples). It also more closely approximates the assumptions required for linear regression.

All of the above concepts can be expressed as raw frequency, but they are often usefully recast as a **probability**. Starting with observed frequency distributions and then expressing them as probabilities has a number of advantages. First, it facilitates comparisons across data-sets. If an object is observed 50 times in a data-set, this number doesn't (by itself) tell us how many times we might expect to find it in a different, differently sized, data-set. However if we express this as the probability of it occurring in a given data-set (e.g. 50/1,000,000 = 0.0005), then this can be usefully interpreted in other domains. Second, it enables flexible redefinition of the reference set, in the form of the **conditional probability**. For example, while the probability of encountering *zip* in a data-set might be quite low, the conditional probability of encountering it, given that we have just encountered the prefix *un-* is likely to be considerably higher. Third, probabilities of events can be combined to establish the probability of a (possibly unobserved) joint event. Given the probability of *-n*, for example, and the probability of *z-*, we can calculate the probability of the two segments occurring next to each other, simply by chance. This can define the **expected frequency** of the co-occurrence. The expected frequency is sometimes compared to the observed frequencies in order to establish whether a given pattern is occurring to a greater or lesser degree than we might expect by chance (Frisch et al., 2004).

Moscoso del Prado et al. (2004b) assume an information theoretic approach to the relationship between word frequency and word strength. They draw a parallel between various measures of word frequency and the information theoretic notion of average unpredictability or **entropy**. Moscoso del Prado et al. (2004b) propose a new measure of the information content of a word, its **information residual**, combining gauges such as surface frequency and base frequency, inflectional ratio, cumulative root frequency, and morphological family size. In a series of experiments, they show that the information residual predicts lexical decision response latencies better than any other measure.

Milin et al. (2009) extend the information theoretic approach. Starting out on the basis of work done on Serbian nominal inflection by Kostic (1995); Kostic et al. (2003), they propose measures for the information content of morphological paradigms and lexical classes. The main insight of Kostić is that one cannot resort to relative frequencies of exponents in an inflectional paradigm, because the functional load of these will also differ, potentially contributing to the costs of lexical processing. The case of derivational morphology is even more complex, because derivational classes can be open and are consequently often unclear in size. Even more complex (and overlapping) inflectional paradigms, like verbal inflection, propose further challenges.

Before we turn to the question of how frequency might contribute to, for instance, morphological productivity, we have to acknowledge the conceptual leap inherent in establishing a connection

between frequencies of word forms in a language corpus and its *cognitive* aspects, reflected, for example, in its familiarity in an individual language user.

One way to join our corpus data with a finding that, for example, an English speaker is more comfortable with *wughood* than with *wugdom* is to assume that the frequencies we observe in a corpus are available to the individual speaker. *Exemplar-based* models of the lexicon (cf. Langacker 1987) are thus attractive to morphologists in this respect because they offer an account in which frequency is an integral property of the word's representation (see, e.g. Hay & Baayen 2005)

A common set of assumptions in the literature is thus: (1) representations are updated with experience, and thus the frequency with which a form is encountered impacts its representation in some way. (2) the corpus we use to gain frequency data is a balanced sample of the ambient language, therefore the frequencies found in the former reflect 'frequencies in the world'; and (3) consequently, the *strength* of the word entry in the individual's lexicon is related (in some way) to its frequency in the corpus, and we can use corpus frequency as a rough proxy for cognitive frequency, or representation 'strength'.

Note, however, that results such as reported by Tily et al. (2009) indicate that a simple raw-frequency based view of human lexical storage may not be tenable, simply due to the sheer capacity that we would need to assume in order to account for the extent of the variation that is apparently affected by linguistic practice. Baayen (2011a) and Baayen & Hendrix (2011) explore an alternative to simplistic rich-memory models of human storage, accommodating the observed detail without a combinatorial explosion of the required capacity. This work will be discussed more in section 7.

3 Frequency and Morphological Processing

A large bulk of experimental research has concentrated on the effect of word frequency on the lexical processing of the individual.

Primary research questions in this arena include: (1) Do (various types of) morphologically complex forms have stored representations (in which case we might see effects of word-from frequency in experiments); or are they actively composed in speech production and decomposed in speech perception? (2) what evidence is there associations between various types of morphological relatives in the lexicon (as might be evidenced by affixal priming effects); and (3) (related to (1) and (2)) to what degree are the sub-parts of morphologically complex forms active during speech perception and production? (For examples see, e.g. Baayen et al. 1997, Bertram et al. 2000, Taft 1979).

This literature is extensive, and we do not outline it in any detail here. However, as outlined in Hay and Baayen (2005), there is good evidence for rather extensive storage of morphologically complex forms, and for wide variation in the degree to which these are decomposed by the individual (both in representation, and - interlinked - during speech perception). Higher degrees of decomposition seem to be facilitated by phonological and semantic transparency, and a high frequency of the parts (e.g. the base and affix) relative to the whole (see e.g. Hay 2000). A low probability phonotactic sequence across the morphological boundary also seems to facilitate decomposition in English, which is likely a reflex of the more general strategy used by English listeners, who tend to use phonotactics to segment words from the speech stream (Aslin et al. 1998, Saffran et al. 1996). The balance of storage and computation is thus variable across word-forms (and likely across listeners and contexts). The relative contribution of storage and computation/decomposition has been modelled in a variety of ways (Baayen et al. 1997; Burani & Caramazza 1987; Gonnerman & Andersen 2007; Schreuder & Baayen 1995).

There is no shortage of interesting and relevant work that tries to establish the domain of

influence of related (in some way) forms upon one another. For example Taft (1979) reports the results of an experiment showing that reaction times in visual lexical decision to monomorphemic words in English are codetermined by the frequencies of the inflected variants. Baayen et al. (1997) observe a similar effect for Dutch. This essentially means that the lexical processing of a word form is not only affected by its token frequency but also by the words belonging to the same *lemma*. The view of the lexical family as based on similarity in form is extended and partly shifted by Schreuder & Baayen (1997), who argue that '[f]amily size appears to be an indicator of the extent to which a noun is incorporated in the network of semantic relations linking concepts in the mental lexicon' (p. 135.). They also observe that, according to their results on the processing of Dutch singulars with varying frequencies of related plurals, the main effect is not the token frequency of the related forms, but rather, their type frequency, i.e. the size of the lexical family. They see family size as a later, central effect in processing, due to its semantic nature.

Dabrowska (2008) reports on the significant effect of type frequency and neighbourhood density in adult Polish speakers' speed and aptitude in supplying dative forms of unfamiliar nouns. Neighbourhood density is similar to family size in the sense that it is a measure of the amount of words that are similar to the target, but, in this case, the similarity is purely formal and has nothing to do with the semantics. The relevance of similarity of forms, often formulated in terms of competing production *schemata*, has been shown in other languages as well, including English past tense formation (Prasada & Pinker, 1993) or German plural formation (Köpcke, 1988).

Moscoso del Prado et al. (2004a) look at other languages to further support the relevance of family size in lexical processing. They report family size effects from Finnish and Hebrew, which, alongside the available Dutch and English data, leads them to argue that the organization of related words in morphological paradigms is an important factor in lexical processing. They also note that the type of the morphology used by the language is reflected in family size effects as well. In Hebrew, which has a non-concatenative inflectional system of word formation, the semantically related family members lead to facilitation while the semantically unrelated ones give rise to inhibition for words with homonym roots. In Finnish, which has a rich agglutinative inflection system, lexical processing of a complex word is only co-determined by the subset of words directly derived from the complex word. These patterns constitute further evidence for the role played by semantic factors in the family size effect.

This large literature clearly indicates that frequency is relevant to the processing and representation of morphological forms. It follows from this that we might expect to observe frequency-based effects in people's usage of morphologically complex forms too, and it is with respect to this question that corpora are particularly useful.

4 Frequency and Morphological Productivity

Aronoff (1980), in a classic study on the psycholinguistic reflexes of productivity, discusses how the topic of word formation had mostly been eschewed in theoretical linguistics, precisely because – unlike in the case of syntactic or phonological rules – it does not suffice to say that a word formation rule is either *obligatory* (in expressing a certain function, for example) or *variable*. We also have to account for the extent of the variability of the word formation rule, which is only possible if we can approximate the distributions of its outputs. This is achievable in two ways. We can look at what human subjects do with novel word formations (whether they find them acceptable, and to what extent, or whether they show larger latency in making this decision, and so on). We can also look at the *frequency* of word forms in a corpus to gain a picture of their *probability*.

In a comprehensive outline of the problems related to the concept, Baayen (2009) gives a tentative definition of productivity as a property of a lexical set that is capable of obtaining new members. In this sense, the set of verbs in English is a fully productive one, since it can always accommodate new members – any English word can function as a verb, including novel ones like *Google*, *Tumblr*, or *Netflix*. In comparison, the set of articles is non-productive, and is unlikely to attain new members beside *a(n)* and *the*.

The innovative early work of Joan Bybee and Harald Baayen emphasised the relevance of both type and token frequency in morphological productivity. Bybee (1985), Bybee (1995), Baayen (1992a), and Baayen (1992b) all argue that high type frequency is related to high morphological productivity, but that some types matter more than others. Very high-token-frequency forms do not tend to contribute overly to the productivity of a pattern (as they are more fused, and less decomposable than lower frequency forms (Bybee 1995). Indeed, a large number of low frequency types is a relatively good indicator of high productivity ((Baayen, 1992b)). Another reason we tend to focus on types to assess productivity, as Baayen reflects, is that categories that are otherwise less productive (novel formations would be, for example, rejected by speakers to a larger extent) can supersede more productive categories in terms of total frequency of *tokens*. The reason for this is that the vast majority of tokens in a corpus of a natural language belong to a tiny minority of types, a distribution first observed by Zipf (1935).

Baayen (1992a) points out that productivity is likely a graded phenomenon as shown by the fact that sets are able to increase in size to varying extents. This is supported by the evidence surveyed by Hay & Baayen (2005). He adds that if we move beyond a binary interpretation of productivity, we have to tackle various interpretations of it. He takes note of three. *Realised productivity* is the size of a category, measured in number of *types*, which we can assess using a representative corpus. *Expanding productivity* indicates the speed with which the category acquires new members. This is usually measured by counting the number of types belonging to the category that only occur once in the corpus, the *hapax legomena*. *Potential productivity*, which indicates the extent to which the category is *saturated*, is estimated by dividing the number of hapax legomena with the number of total types in the corpus.

The use of *hapax legomena* stems from the habit of word distributions to belong to the LNRE class of distributions. LNRE is the acronym for Large Number of Rare Events. LNRE classes have the property of extremely uneven frequency distributions: a few types have a very large token count while the overwhelming majority of the types is extremely rare. For instance, the work of Baayen & Tweedie (1998) and Baayen (2000) reveals that monomorphemic content words (nouns, adjectives and verbs) are outside the LNRE zone, but that word frequencies of affixes, for instance have prototypical LNRE distributions. The LNRE zone, according to Baayen, is the range of sample sizes in a corpus where we keep finding previously unseen words, no matter how large the sample size is, which makes it is hard to predict the future growth rate. The relationship between increasing token and and type frequencies also indicates that the types cannot be compared at different token frequencies. Further, the growth rate is also systematically decreasing as word form frequency becomes larger. For word frequency estimations, even in corpora of tens of millions of words are generally within the LNRE zone. For details on modelling LMRE distributions, see Evert & Baroni (2006).

As Hay & Baayen (2002) note, one way of interpreting the productivity of an affix is as a function of the frequency of the decomposed forms in the lexicon. If the particular form is accessed through its parts, the suffix is activated, which means that its lexical storage is updated with a novel instance. Therefore, the form contributes to the relative strength (and thus productivity) of the suffix. In comparison, if a form is accessed as a whole, the suffix is not 'recognised', and if the

suffix only ever occurs in word forms that are not parsable into parts, it 'dies out' and only remains as a non-productive ending. This is the case of the nominal ending *-th* in English, the umlaut plural in German, or certain cases of French *liaison*. Productivity, argue Hay and Baayen, is thus directly reinforced by decomposition in speech perception. Any factors that facilitate decomposition in perception will facilitate long-term productivity of an affix.

Baayen & Renouf (1996), in a longitudinal corpus study of morphological productivity, also observe that the productivity of an affix varies significantly with the morphological structure of the base word to which it attaches. That is, not all words contribute equally to the productivity of a suffix. Both the internal structure and, most importantly, the relative frequencies of the words have to be taken into consideration when assessing the productivity of a morphological class. (We return to diachronic aspects of morphological productivity in section 6.)

Plag et al. (1999) have a look at morphological productivity in three parts of the British National Corpus, *written*, *context-driven spoken*, and *everyday spoken* language. They use two productivity measures, a measure of how much a certain morphological category contributes to the overall vocabulary size and the rate at which new types are to be expected to appear when N tokens have been sampled. They note how size differences between sub-parts of a corpora can constitute a considerable difficulty when using measures of productivity. They also find that the written register has a greater propensity to form new words and that there is variation in the extent to which suffixes are used across registers in general.

5 Bursty words and bursty productivity

It is important to note that word forms do not have a uniform distribution throughout our linguistic experience, and this is also true for the collected corpora. As Gries (2008) notes, the words *HIV*, *keeper*, and *lively* are about equally frequent in the British National Corpus, but *HIV* has a different distribution, being concentrated in just a small set of the parts that make up the corpus, unlike the latter two, which are more equally dispersed. Differences in dispersion relate to the *niche* of the word, that is, the types of people who use it and the registers or genres it occurs in, amongst other things.

Baayen (1994) approaches the question the other way around, performing text type categorisation based on morphological productivity using a principal components analysis. He finds that literary authors can be differentiated to some extent when we rely on morphological productivity alone, such as by measures of their use of Latinate versus Germanic derivative suffixes.

This all suggests that morphological productivity does not behave uniformly throughout any given corpus. Altmann et al. (2009) and Pierrehumbert (2012) develop the concept of *burstiness*, the property of word frequency to be higher in a given domain. Working with a part of the USENET archive, Altmann et al. (2009) find that content words have a non-uniform distribution deviating from the exponential distribution we would expect them to have (the distribution first observed by Zipf 1935).

They argue that the extent to which words have bursts and lulls in the overall distributions crucially relates to the extent to which word meanings have differential contextualizations. To put it very simply, a word that can be used in a variety of contexts will have a more uniform distribution, whereas a word that is specific to just one or two contexts will be very bursty in these but much less frequent in others, even if the two words have similar overall token frequencies.

Burstiness relates strongly to productivity, because a word form which has a more symmetrical (or *even*) overall distribution is easier to learn. A morphological pattern confined to one context

will then be less productive than another pattern that has a similar overall frequency but occurs in multiple contexts.

The same point is made by Stefan Gries. Gries (2008) argues that, due to the uneven dispersion of forms in corpora, the use of raw frequencies is highly misleading. He goes on to review and compare various dispersion measures, noting that part-based ones are problematic if the parts are not similar in size, that some are sensitive to the number of corpus parts, which can be arbitrary, and that difference-based measures are, in general, sensitive to ordering. He proposes the dispersion measure DP (from 'deviation of proportions') which is based on computing the size of corpus parts and all pairwise absolute differences of observed and expected values of the frequency of the form in these corpus parts. Gries (2009) extends the discussion on DP and its applications.

Researchers including Baayen, Gries, and Pierrehumbert have very different aims in going into the analysis of word frequency distributions in corpora. A primary relevance of recent work to the morphologist is that modelling the behaviour of morphological classes based on simple overall type frequencies is misleading, since the way a morphological class is dispersed in a corpus (and in everyday speech) may be as important as how often it comes up in it.

Lexical neighbours, related to the target word by similarity of either function or form, are also used in machine learning, specifically, in models of lexical productivity. *Nearest-neighbour* algorithms interpret words as consisting of a number of features and predict an unknown feature of the word based on the (known) behaviour of other words which are most similar to it in terms of the known features. The nearest-neighbour algorithm is the main working principle of the Tilburg Memory Based Learner (Timbl) (Daelemans et al., 2007), and is implemented in various other suites of machine learning software, like Weka (Witten et al., 1999). Other models of lexical productivity include Skousen's (2002) Analogical Modeling.

6 Diachronic Corpus Morphology

This section gives an overview of quantitative corpus morphology in the diachronic domain. The gradual refining of methods in corpus-based synchronic morphology is mirrored by work in diachronic morphology, which also shifted from dictionary-based work treating productivity as a binary primitive to a corpus-driven approach treating productivity as a complex, gradient epiphenomenon arising from language change. The trend manifests itself most clearly in the increasing number of multi-century diachronic corpora, such as the Helsinki Corpus, ARCHER, the Corpus of Early English Correspondence (CEEC), Corpus Histórico del Español en México, the Corpus Del Espanõl, or the Mainz Newspaper Corpus, just to mention a few.

Although it has been claimed that their small size and other characteristics pose problems for analysis, historical language corpora present an authentic, context-based section of language and as such, they are now an essential tool in diachronic morphology. In this section, we outline the challenges of research on morphological change and give a brief overview of the methodologies developed in response to these challenges.

Dictionaries are widely used by morphologists as evidence for changing patterns of word-formation (see e.g. Anshen & Aronoff, 1997; Aronoff & Anshen, 2001; Bauer, 1994). However Bauer (2001) points out some of the potential pitfalls of work based on the Oxford English Dictionary. For example it is not clear how long a word was used for or how frequent it was – first citations can be earlier than the general use and last citations can artificially postdate the vitality of a pattern.

Similarly, a rare persisting word may accumulate as many citations in the Oxford English Dictionary as common but short-lived words. Plag and others point out that dictionaries typically

aim for a non-comprehensive account of the language, prioritizing less compositional and more idiosyncratic forms, leaving out much of the predictable forms, which are precisely the locus of productive behaviour (Baayen & Renouf, 1996; Plag, 1999; Dalton-Puffer & Cowie, 2002). As a result, dictionaries are useful in distinguishing productive and non-productive processes, but fail to give us a good picture of the gradience within the productive set. Baayen & Renouf (1996), similarly to Plag, discourage the practice of combining corpus data with non-corpus-based dictionary data, as the presence of compositional forms in the dictionary is likely arbitrary.

Later works also rely on dictionaries in sorting out the role of hapaxes versus neologisms in evaluating diachronic productivity. Trips' Criterion of Productivity (Trips, 2009, p. 38) requires that 'A productive series of formations is defined as the occurrence of formations with a morphological category with at least two hapaxes where a hapax is a new type built by a new rule and a new type exploiting that new rule.' Säily's (Säily 2008; Säily & Suomela 2009) approach is to make sure that the suffixed word has an extant base attested in the time period in which the types are counted. Furthermore, in order to avoid lexicalised forms creeping into the set of hapaxes, she sets an (arbitrary) 'age' threshold for the word in the given time period.

Because of the various problems with dictionary research, corpus work has become increasingly popular. The measurement of morphological change in early corpus-based work was often operationalised through observed changes in type frequencies over time in diachronic corpora. For example, early works often compared percentages of the types of a given affix out of all words, or similar measures "normalizing" over N words across time intervals (e.g. Kučera, 2007; Kaunisto, 2009; Baker, 2010; Berg, 2011; Ciszek, 2012). Besides making it impossible to evaluate the statistical significance of the differences, this practice also ignores two possible issues concerning the distributions of types. The first is that the number of types may grow at different rates for different processes. The second is that the number of types may also grow at a different rate than the number of tokens in the corpus. As a solution, early and/or low computation works often had the assumption that the amount of data is approximately the same size within the time intervals in the corpus and therefore type or token frequency obtained from each interval can be compared directly. Then, if the type counts differ by an order of magnitude, it may be possible to draw conclusions without paying attention to statistical significance, e.g. Dalton-Puffer (1996, p. 106).

Thanks to the rise of larger diachronic corpora, token frequencies became available for models of morphological processes across time. The difference between type and token frequency is as relevant to the study of language change as it is to the synchronic study of morphological productivity. It may shed light on different aspects of the synchronic representation of the word form or pattern during the time intervals considered in the diachronic corpus.

Work on diachronic change shows that token frequency can be a conservative force protecting high-frequency structures from analogical levelling (Bybee & Thompson, 1997). For instance, in English there has been continuous pressure to regularise irregular verb forms. Since the time of Old English, nearly 200 verbs have lost the stem vowel alternation and have adopted the regular past tense form. Synchronically, we find that most of the verbs that are still irregular are very frequent. Bybee & Thomson's hypothesis is that the frequent use has strengthened their representation in memory, which is why they have resisted the pressure from analogical change.

While token frequencies are often not directly available, indirect or intuitive frequencies have been used for their explanatory power. As an example, Enger (2004) found that an intuitive singular-plural bias explained the contemporary results of analogical processes in the gender/declension predictability of Norwegian. Based on dictionary-derived type counts, the declension of most Norwegian nouns is predicted on the basis of gender (the so-called *gender first* nouns). For some nouns, however, gender can be predicted on the basis of declension (the *declension first* nouns).

According to Enger, whether gender or declension is taken as basic can be explained through intuitive token frequency disparities: if the plural form is more frequent than the singular (e.g. *berry* ∼ *berries*), the plural will be taken as the basic form and gender will be predicted on the basis of the declension. If, however, the singular form is more frequent, it will be taken as basic and the declension can be predicted from the gender.

The data in this generation of diachronic morphological works often cross-tabulates the values and significance is evaluated with a chi-square test or parametric regression models, often in Varbrul (e.g. Nevalainen, 2000; Pappas, 2001; Laitinen, 2008).

Only a few studies apply modern corpus-based methodology to assess productivity (as discussed in Section 3) diachronically. This may be due to the fact that diachronic corpora – as shown above – present a number of immediate methodological problems before one can move on to the calculation of the productivity of a given process, or it may be because that commonly available statistical methods have to be reconsidered and adapted to diachronic data.

Baayen's corpus-based approach defines productivity as the likelihood of observing a new type when sampling a sufficiently large corpus. In his measure of potential productivity, gradually increasing number of new types (type frequency, V) are seen as a function of token frequency (N): with the increasing number of tokens (i.e. an increase in corpus size), the number of types will also increase. This relation gives rise to the definition of the *vocabulary growth curve* and to the notion of *vocabulary growth rate*, the latter being calculated by the proportion of hapaxes ($V1$) to the overall number of tokens Baayen & Renouf (1996). In this approach, which yields a synchronic snapshot of productivity, an unproductive morphological category is characterised by few if any hapaxes, especially as the size of the corpus increases, and the vocabulary growth curve flattens out until the vocabulary is exhausted. Conversely, the availability of a productive word formation process guarantees that complex words of even the lowest frequency can be produced and understood if they display the process. Therefore, a large number of hapaxes is a strong evidence that the process is productive in the given time period in the corpus.

When it comes to quantifying the degree of productivity in a morphological change across time, Lüdeling & Evert's (2005) recommendation – based on Baayen (1992c) – is to calculate the synchronic growth rate in each time period first, then compare the degree of synchronic productivity across time. Comparing the confidence intervals around the vocabulary growth curves across time periods then offers a useful diagnostic to detect change: overlap between the confidence intervals of the growth curves of a process in the measured time periods suggests that there is no significant change in the productivity. However, in order to shed light on the (potentially qualitative) aspects of diachronic rivalry between forms, a similar diachronic analysis is needed to evaluate the selectional preference of the process, such as when the same noun displays a diachronically different pluralizing strategy, or when a certain verb paradigm changes the pivot to model other members of the paradigm after. To bring an example for growth curve comparisons, Stichauer (2009) calculated the vocabulary growth and Zipf-Mandelbrot estimates, then used interpolation to show the diachronic development of the Italian deverbal nominal suffixes *-mento*, *-zione* and *-gione* between 13th to the 16th century in the Letteratura Italiana Zanichelli corpus (LIZ 4.0).

Hilpert & Gries (2009) set out to provide statistical tools tailored to analysing trends in frequency changes in multi-stage diachronic corpora. (Gries & Hilpert, 2008; Hilpert, 2011; Gries & Hilpert, 2012). They recommend Kendall's tau to detect trends and propose two iterative algorithms which can be used to periodise significant trends.

Gries & Hilpert (2010) draw on the syntactically parsed Corpus of Early English Correspondence (PCEEC) to explore the morphological shift from 3sg *(e)th* (as in *he giveth*) to *(e)s* (as in *he gives*). They use an iterative algorithm to derive periods, using text frequencies of the variant

suffixes. Based on the clustering of the dataset, Gries and Hilpert distinguish five intervals. This periodisation is one of the explanatory variables that is subsequently fed into a regression analysis that models the change in allomorphy, taking into account both language-external factors (derived periods and author gender) and language-internal factors (phonological context of the suffix) to predict the observable variation. The model has a 95% success rate.

Medina Urrea (2009) track the development of Spanish affixes in the 16-20th century. Medina Urrea operationalised the concept of affixality by quantifying frequency, entropy, economy and compositionally competing word forms (squares) of the affixes. The resulting diachronic morphological profiles provide new insight into the development of Spanish dialects, such as the emergence of Mexican Spanish before the 18th century. Additionally, Medina Urrea calculated pairwise Euclidean distances between the diachronic states and dialectal morphological profiles to study the evolution of Mexican Spanish as a distinct dialectal system and the development of Peninsular Spanish.

Nonparametric methods have also been used in diachronic morphology to some extent. For example, following Baayen's (Baayen, 2000, p. 24-32) computation of Monte Carlo confidence intervals for the accumulation curves of some lexical characteristics, Säily (2008) and Säily & Suomela (2009) calculate accumulation curves for types including hapaxes and then use Monte Carlo sampling to calculate the upper and lower bounds of the curves to compare the type accumulation of female writers against in certain time periods.

Gries & Hilpert (2010) successfully merged a traditional corpus linguistic tool, collocations, with quantitative diachronic morphology. Their model accounts for the *horror aequi* effect, which is interpreted in this context as a sort of Obligatory Contour Principle acting across the word boundary, disfavouring the verbal prefix that ends in the sound the word to the immediate right of the verb begins with (e.g. *he give[th th]anks* vs. *he give[s th]anks*).

Finally, Chapman & Skousen's Analogical Modeling (AM) (2005) provides explicit constraints on analogy that allow for the explanation of how morphological changes begin, which forms most likely serve as patterns for analogy, and which forms are most likely to change. In AM the likelihood of being selected as an analogue is calculated for each competing exemplar based on three properties: proximity in a network, gang effect of surrounding items having the same behaviour and heterogeneity of the surrounding exemplars intervening. In a promising test, AM was given the task of using forms containing negative prefixes for one time period to predict the prefixes that adjectives would take in the subsequent time period. For each of the roughly seventy-year periods in the corpus, AM was able to predict valid prefixes about 90 percent of the time.

With the faster development of methods than corpora, it is common that the same variables (often mined from the same corpora) constitute a test case for increasingly more sophisticated tools and models. For examples see the treatment of the *subject ye/you* shift in Nevalainen (2000), Nevalainen & Raumolin-Brunberg (2003), Raumolin-Brunberg (2005), and Hinneburg et al. (2007); the *3sg -(e)th/-(e)s* shift in Nevalainen (2000), Gries & Hilpert (2010); and the competition between nominalizing *-ity* and *-ness* in Aronoff & Anshen (2001), Dalton-Puffer & Cowie (2002), Säily (2008), and Säily (2011).

7 Going Forward: Is Frequency really Frequency?

As outlined above, there are a plethora of morphological effects which appear to be associated with frequency, and so there is a large amount of work on this topic. However morphological work has engaged somewhat less than desirable in dialogue about other potential effects with which frequency is confounded. One of these, for example, is Age of Acquisition (see, e.g. Johnston & Barry 2006).

Earlier learned words appear to behave differently in a number of different domains, and age of acquisition is, of course, highly correlated with lexical frequency. A second factor is predictability in context, and the effects of repetition, both of which have been shown to be relevant in speech production (Bell et al., 2009). These are, of course, highly relevant, because frequent words are more likely to have been recently produced, and are also more likely to be contextually predictable.

There is also the question of whether the frequency of exposure to a word is the most relevant factor, or whether how familiar a word appears to an individual is more important. In an extensive review of preceding literature, supported by experimental data, Gernsbacher (1984) arrives at the conclusion that *subjective familiarity* is a better predictor of various measures of lexical processing than objective frequency. Subjective familiarity is, quite simply, the extent to which an individual participant finds a word form familiar, measurable on a one-dimensional scale.

Connine et al. (1990) have another look at the effect of frequency and familiarity in lexical processing. They find that both objective frequency and subjective familiarity affect lexical decision and naming tasks in both reading and listening. They note, however, that objective frequency (based on written corpora, such as the Brown corpus) is a better predictor of processing in a reading task (where the medium is kept constant) and that the importance of familiarity *vis-à-vis* objective frequency increases considerably in a delayed naming task. Balota et al. (2001) call attention to the inherent problems of the *subjective familiarity* concept, which arguably conflates distinct measures of familiarity, including semantics (whether the individual understands the word) and orthography (whether the word has a highly irregular spelling). To give an example, *birthday* and *architecture* have roughly the same token frequency in the Celex corpus (352 and 340, respectively), but English speakers will know what the former is, while some might be uncertain about the meaning of the latter, resulting in a possible difference in familiarity ratings, one that has little to do with the probability of encountering either. Similarly, the fact that *both* is pronounced with a diphthong, unlike all other common English words ending in *-oth*, like *goth, moth, cloth, froth*, might retract from its similarity – though, in this case, its frequency will probably make up for this.

Balota et al. find a strong correlation between subjective estimates of word *frequency* and objective frequency, which leads them to conclude that frequency counts from corpora can be used to approximate word frequency in the ambient language of the individual.

Given the range of possible confounds, whenever we see what appears to be a direct relationship between frequency and behaviour, we should always consider the degree to which we can be confident that itself is really the driving force.

Most startling, in this regard, is the recent work of Harald Baayen and colleagues, within the naive discriminative learning paradigm. For example, Baayen (2011b) re-examines the ubiquitous word frequency effect which has been reported for lexical decision times. He shows that 90% of variance in word frequencies in his data-set is actually predictable from other properties. These properties involve contextual measures capturing the contextual distribution of the word (such as syntactic and morphological family size, syntactic entropy and dispersion). Once these properties are accounted for, the frequency of the word has very little explanatory power in predicting reaction times. Baayen argues that the word frequency effect is 'an epiphenomenon of learning to link form to lexical meaning'.

If frequency effects are in fact largely artefactual, this would certainly recast many of the questions investigated in the literature outlined in this chapter, most of which situate frequency effects directly the representation – presupposing models which Baayen refers to as containing a 'counter in the head' (Baayen, 2011b).

Indeed in Baayen et al. (2011) Baayen and collaborators carefully discuss many of the frequently reported frequency effects in morphology, and demonstrate that they emerge from their simple

learning model. Their model does not include representations of morphemes, and does not directly encode frequency. Frequency effects for complex words and phrases emerge in the model without the presence of any overt whole-word or whole-phrase representations. They observe that their model 'can be viewed as a formal, computational implementation of the notion of analogy in word and paradigm morphology' (Baayen et al., 2011) (cf. e.g. Blevins 2003). This body of work seems to show that frequency is not important because it is stored in representations, and active in production and perception processes, but rather that it is an epiphenomenon of the effects of contextual predictability in learning.

This clearly provides many pathways for exploration and – if true – would turn much of the literature outlined in this chapter on its head. The observed correlations and phenomena in this literature will provide an important baseline for investigations, but many questions now remain about the representations and processes driving such phenomena. As Baayen has shown, careful work needs to be done to understand the mechanisms through which contextual factors and frequency are related, and the degree to which each drives or reflects morphological learning, production, perception and representation. The careful use of corpora will remain central in such explorations, as will carefully implemented computational models.

References

Alegre, Maria & Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40. 41–61.

Altmann, Eduardo G., Janet B. Pierrehumbert & Adilson E. Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One* 4(11). e7678.

Anshen, Frank & Mark Aronoff. 1997. Morphology in real time. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1996*, 9–12. Dordrecht: Kluwer.

Aronoff, Mark. 1980. The relevance of productivity in a synchronic description of word formation. In Jacek Fisiak (ed.), *Historical morphology. Papers prepared for the International Conference on Historical Morphologsy held at Boszkowo, Poland, 15-18 March 1978*, 71–82. De Gruyter Mouton.

Aronoff, Mark & Frank Anshen. 2001. Morphology and the lexicon: lexicalization and productivity. In Andrew Spencer & Arnold M. Zwicky (eds.), *The Handbook of Morphology*, 237–247. Oxford: Blackwell.

Aslin, Richard N., Jenny R. Saffran & Elissa L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9(4). 321–324.

Auer, Edward T., Lynne E. Bernstein & Paula E. Tucker. 2000. Is subjective word familiarity a meter of ambient language? a natural experiment on effects of perceptual experience. *Memory and Cognition* 28(5). 789–797.

Baayen, R. Harald. 1992a. On frequency, transparency and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 181–208. Kluwer.

Baayen, R. Harald. 1992b. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 109–149. Springer.

Baayen, R. Harald. 1992c. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 109–149. Dordrecht: Kluwer.

Baayen, R. Harald. 1994. Derivational productivity and text typology. *Journal of Quantitative Linguistics* 1(1). 16–34.

Baayen, R. Harald. 2000. *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In A Luedeling & M Kyto (eds.), *Corpus Linguistics. An international handbook*, 909–919. Berlin: Mouton De Gruyter.

Baayen, R. Harald. 2011a. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328.

Baayen, R. Harald. 2011b. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5. 436–461.

Baayen, R. Harald, Ton Dijkstra & Robert Schreuder. 1997. Singulars and plurals in dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37(1). 94–117.

Baayen, R. Harald & Peter Hendrix. 2011. Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. In *Empirically examining parsimony and redundancy in usage-based models, lsa workshop*, .

Baayen, R. Harald, P. Milin, D. Filipovic Durdevic, P. Hendrix & M. Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Baayen, R. Harald, Richard Piepenbrock & Hedderik van Rijn. 1993. The CELEX lexical database.

Baayen, R. Harald & Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language* 72. 69–96.

Baayen, R. Harald & Fiona J. Tweedie. 1998. Sample-size invariance of LNRE model parameters: Problems and opportunities. *Journal of Quantitative Linguistics* 5(3). 145–154.

Baker, Paul. 2010. Diachronic variation. In *Sociolinguistics and corpus linguistics*, chap. 3, 57–80. Edinburgh University Press.

Balota, David A, Maura Pilotti & Michael J Cortese. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition* 29(4). 639–647.

Bauer, Laurie. 1994. *Watching English change*. London: Longman.

Bauer, Laurie. 2001. *Morphological productivity* (Cambridge Studies in Linguistics 95). Cambridge: Cambridge University Press.

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60. 92–111.

Berg, Thomas. 2011. A diachronic frequency account of the allomorphy of some grammatical markers. *Journal of Linguistics* 47(1). 31–64.

Bertram, R., Robert Schreuder & R. Harald Baayen. 2000. The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26. 1–23.

Blevins, James P. 2003. Stems and paradigms. *Language* 79(4). 737–767.

Burani, C. & A. Caramazza. 1987. Representation and processing of derived words. *Language and Cognitive Processes* 2. 217–227.

Bybee, Joan. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10.

Bybee, Joan & Sandra Thompson. 1997. Three frequency effects in syntax. In *Proceedings of the 23rd annual meeting of the Berkeley Linguistics Society: General session and parasession on pragmatics and grammatical structure*, 378–388. Berkeley: Berkeley Linguistics Society.

Chapman, Don & Royal Skousen. 2005. Analogical Modeling and morphological change: the case of the adjectival negative prefix in English. *English Language and Linguistics* 9(2). 333–357.

Ciszek, Ewa. 2012. The Middle English suffix -ish: Reasons for decline in productivity. *Studia Anglica Posnaniensia* 47(2–3). 27–39.

Connine, Cynthia M, John Mullennix, Eve Shernoff & Jennifer Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(6). 1084–1096.

Dabrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers – productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58(4). 931–951.

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch. 2007. TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide. Tech. rep. ILK Research Group Technical Report Series no. 07-07.

Dalton-Puffer, Christiane. 1996. *The French influence on Middle English morphology: a corpus-based study of derivation* (Topics in English linguistics 20). New York: Mouton de Gruyter.

Dalton-Puffer, Christiane & Claire Cowie. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In J. E. Díaz Vera (ed.), *A changing world of words: Studies in English historical lexicography, lexicology and semantics*, 410–437. Amsterdam: Rodopi.

de Jong, Nivja, Robert Schreuder & R. Harald Baayen. 2000. The morphological family size effect and morphology. *Journal of Memory and Language* 42. 390–405.

Enger, Hans-Olav. 2004. On the relation between gender and declension: a diachronic perspective from Norwegian. *Studies in Language* 28(1). 51–82.

Evert, Stefan & Marco Baroni. 2006. The zipfR library: Words and other rare events in R. In *Presentation at useR! 2006: The Second R User Conference, Vienna, Austria*, Vienna, Austria.

Frisch, Stefan A., Janet B. Pierrehumbert & Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22. 179–228.

Gernsbacher, Morton A. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113(2). 256–281.

Gonnerman, Seidenberg M. S., L. & E. Andersen. 2007. Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General* 136. 323–345.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.

Gries, Stefan Th. 2009. Dispersions and adjusted frequencies in corpora: further explorations. *Language and Computers* 71(1). 197–212.

Gries, Stefan Th. 2011. Frequency tables: tests, effect sizes, and explorations. In Dylan Glynn & Justyna A. Robinson (eds.), *Polysemy and synonymy: Corpus methods and applications in cognitive linguistics*, Amsterdam: John Benjamins.

Gries, Stefan Th. to appear(a). *Companion to Cognitive Linguistics* chap. Corpus and quantitative methods. London, New York: Continuum.

Gries, Stefan Th. to appear(b). Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us. In Irma Taavitsainen, Merja Kytö, Claudia Claridge & Jeremy Smith (eds.), *Developments in English: expanding electronic evidence*, Cambridge University Press.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3(1). 59–81.

Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(3). 293–320.

Gries, Stefan Th. & Martin Hilpert. 2012. Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*, Oxford: Oxford University Press.

Gries, Stefan Th & John Newman. to appear. *Research methods in linguistics* chap. Creating And Using Corpora. Cambridge University Press.

Hay, Jennifer B. 2000. *Causes and consequences of word structure.* New York and London: Routledge.

Hay, Jennifer B. 2001. Lexical frequency in morphology: is everything relative? *Linguistics* 39(6). 1041–1070.

Hay, Jennifer B. & R. Harald Baayen. 2002. Parsing and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2001*, 203–235. Springer.

Hay, Jennifer B. & R. Harald Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.

Hay, Jennifer B., Stefanie Jannedy & Rens Bod (eds.). 2003. *Probabilistic linguistics.* MIT Press.

Hilpert, Martin. 2011. Diachronic collostructional analysis: how to use it and how to deal with confounding factors. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics* (Topics in English linguistics 73), 133–160. Berlin: De Gruyter.

Hilpert, Martin & Stefan Th. Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquistition. *Literary and Linguistic Computing* 24(4). 385–401.

Hinneburg, Alexander, Heikki Mannila, Samuli Kaislaniemi, Terttu Nevalainen & Helena Raumolin-Brunberg. 2007. How to handle small smaples: Bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing* 22(2). 137–150.

Johnston, Robert & Christopher Barry. 2006. Age of acquisition and lexical processing. *Visual Cognition* 13. 789–845.

Kaunisto, Mark. 2009. The rivalry between English adjectives ending in *-ive* and *-ory*. In R. W. McConchie, Alpo Honkapohja & Jukka Tyrkkö (eds.), *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX 2)*, 74–87. Somerville, MA: Cascadilla Proceedings Project.

Kennedy, Alan, Robin Hill & Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th european conference on eye movement*, .

Kilgarriff, Adam & Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3). 333–347.

Köpcke, Klaus-Michael. 1988. Schemas in German plural formation. *Lingua* 74. 303–35.

Kostic, Aleksandar. 1995. Information load constraints on processing inflected morphology. In Laurie Beth Feldman (ed.), *Morphological aspects of language processing*, 317–344. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Kostic, Aleksandar, Tanja Markovic & Aleksandar Baucal. 2003. Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. Harald Baayen & Robert Schreuder (eds.), *Morphological structure in language processing* (Trends in Linguistics. Studies and Monographs (TiLSM) 151), 1–44. Amsterdam: Walter de Gruyter.

Kučera, Henry & Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Kučera, Karol. 2007. Mapping the time continuum: A major raison d'étre for diachronic corpora. In *Proceedings of corpus linguistics birmingham 2007*, Birmingham: University of Birmingham.

Laitinen, Mikko. 2008. Sociolinguistic patterns in gramnmaticalization: *he, they*, and *those* in human indefinite reference. *Language Variation and Change* 20. 155–185.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, vol. 1. Stanford university press.

Lüdeling, Anke & Stefan Evert. 2005. The emergence of non-medical *-itis*. Corpus evidence and qualitative analysis. In S. Kepser & M. Reis (eds.), *Linguistic evidence. Empirical, theoretcial, and computational perspectives*, 315–333. Berlin: Mouton de Gruyter.

Manning, Christopher D & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Medina Urrea, Alfonso. 2009. Toward a comparison of unsupervised diachronic morphological profiles. *Language and Computers* 71. 29–45.

Milin, Petar, Victor Kuperman, Aleksandar Kostic & R. Harald Baayen. 2009. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 214–252. Oxford: Oxford University Press.

Moscoso del Prado, Martin, Raymond Bertram Fermin, Tuomo Haikio, Robert Schreuder & R. Harald Baayen. 2004a. Morphological family size in a morphologically rich language: The case of finnish compared to dutch and hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30. 1271–1278.

Moscoso del Prado, Martin, Raymond Bertram Fermin, Aleksandar Kostic & R. Harald Baayen. 2004b. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94(1). 1–18.

Nevalainen, Terttu. 2000. Gender differences in the evolution of Standard English: Evidence from the Corpus of Early English Correspondence. *Journal of English Linguistics* 28(1). 38–59.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Longman.

Pappas, Panayiotis A. 2001. The microcosm of a morphological change: Variation in *thelo* + infinitive futures and *ethela* + infinitive counterfactuals in Early Modern Greek. *Diachronica* 18(1). 59–92.

Pierrehumbert, Janet B. 2012. Burstiness of verbs and derived nouns. In Diana Santos, Krister Lindén & Wanjiku Ng'ang'a (eds.), *Shall we play the festschrift game? essays on the occasion of Lauri Carlson's 60th birthday*, 99–115. Springer.

Plag, Ingo. 1999. *Morphological productivity: structural constraints in English derivation* (Topics in English linguistics 28). New York: Mouton de Gruyter.

Plag, Ingo, Christiane Dalton-Puffer, R. Harald Baayen et al. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.

Prasada, S. & S. Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes* 8(1). 1–56.

Raumolin-Brunberg, Helena. 2005. The diffusion of subject *you*: A case in historical sociolinguistics. *Language Variation and Change* 17(1). 55–73.

Saffran, Jenny R., Elissa L. Newport & Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35(4). 606–621.

Säily, Tanja. 2008. *Productivity of the suffixes -ness and -ity in 17th century English letters: A sociolinguistic approach*. Helsinki University of Helsinki Masters Thesis.

Säily, Tanja. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7(1). 119–141.

Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: the case of women, men and *-ity* in early English letters. In Antoinette Renouf & A. Kehoe (eds.), *Corpus linguistics: Refinements and reassessments*, 87–109. Amsterdam: Rodopi.

Schreuder, Robert & R. Harald Baayen. 1995. Modeling morphological processing. In Laurie Feldman (ed.), *Morphological aspects of language processing*, 131–154. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Schreuder, Robert & R. Harald Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language* 37(1). 118–139.

Skousen, Royal. 2002. *Analogical modeling: An exemplar-based approach to language*. Amsterdam: John Benjamins.

Stichauer, Pavel. 2009. Morphological productivity in diachrony: The case of deverbal nouns in *-mento, -zione* and *-gione* in Old Italian from the 13th to the 16th century. In Fabio Montermini, Gilles Boyé & Jesse Tseng (eds.), *Selected proceedings of the 6th décembrettes: Morphology in Bordeaux*, 138–147. Somerville, MA: Cascadilla Proceedings Project.

Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7(4). 263–272.

Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari & Joan Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2). 147–165.

Trips, Carola. 2009. *Lexical semantics and diachronic morphology: The development of -hood, -dom and -ship in the history of English* (Linguistische Arbeiten 527). Berlin: De Gruyter.

Walker, Abby & Jennifer B. Hay. 2011. Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology* 2(1). 219–237.

Witten, I.H., E. Frank, L. Trigg, M. Hall, G. Holmes & S.J. Cunningham. 1999. Weka: Practical machine learning tools and techniques with Java implementations. In *ICONIP/ANZIIS/ANNES*, vol. 99, 192–196.

Zipf, George Kingsley. 1935. *The psycho-biology of language: an introduction to dynamic philology*. The MIT Press.

Zséder, Attila, Gábor Recski, Dániel Varga & András Kornai. 2012. Rapid creation of large-scale corpora and frequency dictionaries. In Nicoletta Calzolari (ed.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 1462–1465. Istanbul, Turkey: European Language Resources Association (ELRA).