English Language and Linguistics 3.2: 209–228. © Cambridge University Press 1999.

Morphological productivity across speech and writing[1]

INGO PLAG, CHRISTIANE DALTON-PUFFER, AND HARALD BAAYEN
University of Hanover, University of Vienna, Max Planck Institute for Psycholinguistics, Nijmegen

(Received 4 January 1999; revised 14 April 1999)

Claims about the productivity of a given affix are generally made without differentiating productivity according to type of discourse, although it is commonly assumed that certain kinds of derivational suffixes are more pertinent in certain kinds of texts than in others. Conversely, studies in register variation have paid very little attention to the role derivational morphology may play in register variation.

This paper explores the relation between register variation and derivational morphology through a quantitative investigation of the productivity of a number of English derivational suffixes across three types of discourse in the British National Corpus (written language, context-governed spoken language, and everyday conversations). Three main points emerge from the analysis. First, within a single register, different suffixes may differ enormously in their productivity, even if structurally they are constrained to a similar extent. Second, across the three registers under investigation a given suffix may display vast differences in productivity. Third, the register variation of suffixes is not uniform, i.e. there are suffixes that show differences in productivity across registers while other suffixes do not, or do so to a lesser extent. We offer some tentative explanations for these findings and discuss the implications for morphological theory.

1 Introduction

Corpus-based studies in the productivity of word formation have shown that large computer corpora can be fruitfully employed to find long-sought solutions to questions relating to the problem of morphological productivity (e.g. Baayen, 1992, 1993; Baayen & Lieber, 1991; Baayen & Renouf, 1995; Baayen & Neijt, 1997; Plag, 1999). These authors stated their claims about the productivity of a number of affixes without differentiating productivity according to type of discourse, although it is commonly assumed that certain kinds of derivational suffixes are more pertinent in certain kinds of texts than in others. It is presently unclear to what extent this common assumption is true or false and how it may have skewed the results in the aforementioned studies.

Studies in register variation have shown in great detail that there is a whole range of observable syntactic and lexical differences between different registers or text types, such that the clustering of such properties can even be used in defining a certain type of discourse (cf. Biber, 1995). However, very little attention has been

[1] We thank the anonymous referees of this journal and Bas Aarts for comments and helpful suggestions. The first two authors are indebted to the third author and to the Max Planck Institute for Psycholinguistics at Nijmegen for their hospitality and to the Max Planck Society for financial support.

devoted to the role derivational morphology may play in register variation. In many publications one can find cursory and sometimes implicit remarks on this topic, with nominalizations being unanimously regarded as typical of written, information-centered texts (e.g. Lipka, 1987; Koch & Oesterreicher, 1994: 591; Enkvist, 1977: 184; Kastovsky & Kryk-Kastovsky, 1997: 469). It is unclear whether this stands up to broader empirical testing and whether it can be generalized to other, non-nominalizing suffixes. Furthermore, if differences in the patterning of complex words in different text types can be detected, the relation of this patterning to the diverse functions of derivational morphology in language use remains to be determined.

This paper presents a quantitative investigation of the productivity of a number of English derivational suffixes across three types of discourse (written language, context-governed spoken language, and everyday conversations; see below). It is thus a study of the role of morphology in language use and is only secondarily concerned with the structural aspects of morphological productivity.[2] The data for our study come from the British National Corpus. Three main points emerge from the analysis. First, suffixes may differ enormously in their productivity within a single register, even when constrained structurally to a similar extent. Second, a given suffix may display vast differences in productivity across the three registers investigated in the present study. Third, register variation is not uniform for the suffixes we have studied, i.e. there are suffixes that show differences in productivity across registers while other suffixes do not, or do so to a lesser extent. We offer some tentative explanations for these findings and discuss the implications for morphological theory.

2 Methodology and data

2.1 The BNC

The data analyzed in this paper come from the British National Corpus (BNC, version 1.0). The BNC consists of c. 100 million word tokens of contemporary British English (89 percent post-1975) with a written/spoken ratio of approximately 9/1. Given the aims of this paper it is necessary to take a look at the different types of discourse represented in the corpus. The text samples in the 89+ million-word written corpus are classified into the two major categories 'fictional' and 'informative' with the latter splitting up into eight domains derived from the topical content of the samples (Arts, Belief and Thought, Commerce, Leisure, Natural Science, Applied Science, Social Science, World Affairs). The 10+ million words of spoken language form two distinct subcorpora. The so-called demographic corpus was gathered by having a demographically selected sample of speakers record their everyday conversations over the period of a week. The so-called context-governed

[2] For a recent discussion of the structural aspects of morphological productivity, see Plag, 1999.

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

Table 1. *The three subcorpora of the BNC (adapted from Burnard, 1995: 9)*[3]

	number of word tokens
Written	89,740,544
Spoken Context Governed	6,154,248
Spoken Demographic	4,211,216

corpus of the BNC consists of all types of spoken English other than spontaneous informal conversation, thus featuring samples from lectures, classroom interaction, news commentary, business meetings, sermons, legal proceedings, sports commentaries, and broadcast talk shows among many others. Similar to the written corpus, the context-governed spoken part is also subdivided according to real-world context. There are four categories: education, business, public/institutional, and leisure. Table 1 gives a general overview of the relative sizes of the three subcorpora of the BNC.

With over 10 million words of spoken language the BNC certainly represents by far the largest source of computerized spoken data available. The well-established and widely used London–Lund Corpus, by comparison, contains 1 million words. Large as the BNC may seem, for specific linguistic phenomena with relatively low frequencies, such as the questions of derivational morphology pursued in this paper, the 4 plus 6 million words quickly split up into rather small data-bases once further variables are introduced. This would be the case, for instance, if one wanted to find out about regional and/or gender differences. As the present paper aims at providing a first global view of register variation in word formation, it was decided to use the subdivisions of the corpus as predefined by the structure of the BNC. In the following section we will take a closer look at the implications of this decision.

2.2 The question of register

The most salient division of 'language' in the BNC is clearly that into speech and writing, i.e. the division according to the medium which is used for language production. Quite apart from the practicalities and technicalities of corpus production – the gathering of 10 million spoken words was possible only because of a joint effort of several commercial and noncommercial institutions in the UK – this division is founded in a longstanding tradition of research into the differences between speech and writing.[4]

Even though the notion of 'typical speech' and 'typical writing' (or 'orality' and

[3] For a detailed account of the composition and structure of the BNC see Burnard (1995: chapters 3 and 4).

[4] See Biber (1988: 47–58) for an overview and discussion. An even more longstanding tradition in this respect exists in education, where teaching the composition of written texts (we are not talking of the skill of writing itself) is a major item in curricula of all educational levels. Teaching the composition of oral texts, in comparison, plays a negligible role – at least in modern Western societies.

'literacy' following Tannen, 1982) continues to be useful and legitimate, it has become clear that a strict division between the linguistic characteristics of speech and writing is impossible as the division generalizes over several situational (and processing) constraints and a variety of communicative tasks (e.g. personal letters constitute a written genre with relatively oral situational characteristics; cf. Biber, 1988: 45). A more fine-grained analysis has to operate in a multidimensional space.

One of these dimensions is expressed through the topical and situational context in which language is produced. The compilers of the BNC have called this variable 'domain' (see section 2.1) while in linguistics 'register' seems to be the more common term (Ferguson, 1994; Biber, 1995). Note that register distinctions are not defined in linguistic terms but rest on participant relations, purpose, production circumstances, etc.

It is above all the work of Biber (e.g. 1988, 1995) which represents a systematic attempt to combine the study of register with the identification of typical linguistic features in a systematic way. Among the 67 linguistic features Biber uses in the analysis of English, there is only one which uncontroversially relates to the topic of word formation, that is the feature 'nominalizations (ending in *-tion, -ment, -ness, -ity*)' (Biber, 1988: 227). Wells (1960) claimed that nominalization marks a fundamental distinction between registers. Chafe and Danielewicz (1986) interpret nominalizations as markers of conceptual abstractness which can be used to condense information into fewer words and are thus particularly useful for conveying abstract (as opposed to situated) information. It seems that this diagnosis is the received wisdom of the linguistic community, as is witnessed by passing remarks on the functions of word formation in text, which mostly refer to nominalization (e.g. de Beaugrande & Dressler, 1981; Kastovsky, 1982; Lipka, 1987; Akimoto, 1991). In Biber's quantitative analysis it turns out that his only word-formation feature, i.e. 'nominalization', loads significantly only on one of his seven 'dimensions of register variation' of English text-types (Biber, 1995: 155). This is the dimension of 'Situated vs. Elaborated Reference', where a high degree of nominalization is typical of the latter. In other words, the only word-formation feature systematically investigated contributes to the distinction between registers only once. This suggests that word formation is not a major differentiating factor. However, Biber and his co-workers (1995) include diverse morphological categories in the study of Korean and Somali texts, implying that word formation may nevertheless have a more important role to play. The latter finding would be in line with other studies which claim that word formation is put to use in different ways in fictional and nonfictional texts (Kastovsky & Kryk-Kastovsky, 1997: 469; Akimoto, 1991: 282; Indra, 1990).

The aim of the present paper is to explore to what extent word formation differs across speech and writing using the three domains of the BNC (written, spoken context-governed, and spoken demographic language) as a first window on this aspect of register variation. Of these three domains, that of the spontaneous conversations in the subcorpus of spoken demographic language is perhaps most homogeneous, while the subcorpus of written language is quite heterogeneous and

covers a wide range of registers. Leaving the detailed analysis of the stylistic patterning of word formation across the full range of registers in the various subcorpora of the BNC for further study, the present paper seeks to establish to what extent word formation differs at the more abstract level of spoken versus written language.

2.3 The data

We extracted raw data for fifteen suffixes by means of string searches from the BNC word-frequency lists compiled by Adam Kilgarriff.[5] These suffixes, which are at least moderately productive, are distributed over the following categories:[6]

(1) abstract nouns: *-ity, -ness, -ion*
participant nouns: *-er, -ist*
measure partitive nouns: *-ful*
derived verbs: *-ize*
derived adjectives: *-able, -free, -ful, -ish, -less, -like, -type, -wise*[7]

The main criterion for choice was the aim to complement Biber's only derivation-relevant feature 'nominalization (*-ion, -ity, -ness, -ment*)' with an array of other derivational patterns performing different morphosyntactic and morphosemantic functions.

Kilgarriff's word-frequency lists provide counts for word forms and their word-category tags. We would have liked to use these word-category tags to separate, for instance, verbs like *to partition* from nouns such as *partition.* However, we found the word-category tagging to be too error-prone to be useful for our purposes.[8] We

[5] These word-frequency lists can be obtained via FTP from the following site: ftp://ftp.itri.bton.ac.uk/pub/bnc.

[6] The morphological status of some of the items in (1) is perhaps controversial. Thus, derivations with *-type*, *-free* or *-like* could be argued to be compounds, and the nature of partitive *-ful* is questionable. *-ful* may look like an adjectival suffix, but it productively forms measure partitive nouns. One referee suggests that *-ion* could in fact be two suffixes, one of them productive (*-ation*), the other unproductive (as in *insertion*, *production*). Be that as it may, the problem is even more complicated, since would-be productive *-ation* could in turn be argued to be two suffixes, one attaching to derived verbs ending in *-ize*, *-ify*, and *-ate*, the other attaching to certain nouns (see Plag, 1999: 68–9, 207–10, for details). In order to avoid questionable or arbitrary classifications, we decided to ignore these potential differences among *-ion* formations and include all kinds of derived *-ion* nouns in our analysis. If anything, the inclusion of putatively unproductive *-ion* derivatives has led to a potential decrease in the overall productivity of *-ion* derivatives. As it turned out, this potential decrease cannot have been very significant, since *-ion* emerged as one of the most productive suffixes nevertheless (see below).

In general, the structural properties of the morphological categories under investigation are certainly interesting by themselves, but will not be further elaborated on in this paper, because we focus on the *use* of derived words and not on their structural aspects.

[7] It can be argued that *-wise* is in fact an adverbial, and not an adjectival, suffix. This view rests on the controversial assumption that adjective and adverb are distinct syntactic categories. Nothing we say in this paper about *-wise* hinges on the classification of *-wise* as either adverbial or adjectival.

[8] Tagging is discussed on the current BNC web-page (http://info.ox.ac.uk/bnc/what/gramtag.html). There it is said that only c. 1.7 percent of all words are tagged erroneously and that a further 4.7 percent of words carry ambiguous (or portmanteau) tags. Though we have not computed any figures and cannot

therefore analyzed the raw frequency lists for our fourteen suffixes by hand, removing irrelevant items and consulting the *OED* and checking words in their context in the BNC, where necessary.[9]

The only suffix for which we were forced to make use of the word-category labels was agentive *-er*. The *-er* files contained the highest amount of irrelevant data such as verbs (e.g. *to cater*), words from other languages, especially French and German, occurrences of the suffix *-ster* (e.g. *gangster*), all the comparatives of adjectives (e.g. *larger, higher*) and a large number of names originating from occupational terms (e.g. *Wheeler, Stocker, Thatcher* etc.). Given the large amount of word types arising from the string search (V = 48,476), we discarded all words tagged as proper nouns, adjectives, and verbs. Especially the decision to discard items tagged as proper nouns unavoidably led to the potential loss of relevant data because of wrongly tagged items. On the other hand, with words that are both current as agent nouns *and* proper nouns (such as *Walker*) not all tokens tagged as common nouns were checked for whether they were partially wrongly tagged proper names. The results based on the *-er* data are therefore to be interpreted with caution.

In order to count as a token with a given suffix, an item had to fulfill the following conditions. Firstly, it should belong to the morphological category in question both formally and semantically. Secondly, the base either had to be an independent word of Modern English (e.g. *conform–conformity*) or needed to occur as a bound item in at least one other derivative (e.g. *baptize–baptism*). Note that we have been conservative here by excluding semantically opaque but formally analyzable items from further consideration (e.g. *organize*). Complex lexical items with derivational affixes attached outside the suffixes in question were removed, as, strictly speaking, they do not belong to the morphological categories of these affixes. This decision affected all derived adjectives used as adverbs (e.g. *purposefully*), prefixed formations (e.g. *unavailable*) and compounds (e.g. *performance-artist*). Inflectional suffixes were not discarded but collapsed with their base forms, so that, for instance, noun-plurals were subsumed under their associated singulars.

3 Measuring morphological productivity

In order to estimate the role of a particular morphological category in a given text or text type a quantitative analysis of the productivity of the pertinent affixes in this text or text type needs to be carried out. Productivity is generally loosely defined as the possibility to coin new complex words according to the word-formation rules of a given language. The main methodological problem with measuring the degree of productivity of a given affix is to operationalize the notion of 'possibility' mentioned in the above definition of productivity. Apart from truly unproductive derivational

supply percentages, it is clear from our data that derived words seem to attract both erroneous and ambiguous tags to a much greater extent. This type of error-proneness is inevitable because we are looking at rare events, where statistical tagging is not performing at its best.

[9] Simple searches can be conducted at the following web-site: http://sara.natcorp.ox.ac.uk/lookup.html.

processes like e.g. nominalizing *-th* (as in *length*), productivity seems to be a scalar concept. In other words, with some affixes one is more likely to encounter newly formed words than with others, a fact that makes productivity a probabilistic notion which is susceptible to statistical analysis.

Baayen and co-workers (Baayen, 1992; Baayen, 1993; Chitashvili & Baayen, 1993; Baayen & Lieber, 1991; Baayen & Renouf, 1996) have developed a number of corpus-based statistical measures of productivity which all rely on the existence of more or less representative and sufficiently large samples of computerized texts. What exactly counts as sufficiently large is not easy to determine but even relatively small corpora like the Dutch Eindhoven Corpus (600,000 words of written text) seem to yield interesting results (Baayen, 1992, 1993).

There are three principal statistical measures available on the basis of which further analyses (such as the ones to be presented in section 4) can be carried out. The first of these measures is the number of tokens N of a given morphological category, which is calculated by counting how often words of a given morphological category are used in the corpus (number of tokens = N). The second measure is the number of types V of a given morphological category, which is calculated by counting how many different words belonging to the category occur in the text (number of types = V). V is also referred to as 'extent of use'. The third important measure is the number of words of the given category that occur only once in the corpus (so-called hapax legomena, or hapaxes for short), which can be interpreted as an indication of how often a suffix is used to coin a hitherto unattested word, i.e. a neologism. Why should hapaxes, i.e. the new, unobserved types, tell us anything about productivity? After all, the new, unobserved types could only be rare words, and not neologisms. There are, however, strong arguments for claiming that hapaxes are significant for productivity.

In a sufficiently large corpus, the number of hapaxes in general approximates half the observed vocabulary size (e.g. Zipf, 1935). Chitashvili & Baayen (1993: 57) call this kind of distribution a 'Large Number of Rare Events' distribution. This is a kind of distribution with so many low-probability words that special care is required for statistical analysis. In other words, the vocabulary as observed in texts itself, without making any distinction with respect to morphological structure, is so productive that special statistical measures are required. What is interesting from a morphological point of view is that this productivity of the vocabulary as a whole is driven by its productive word-formation rules. The shape of the word-frequency distributions of productive word-formation rules is similar to the shape of the word-frequency distribution of texts, while the shape of the word-frequency distributions of unproductive word-formation patterns is qualitatively quite different (cf. Chitashvili & Baayen, 1993: 80–6, 125–6 for the difference between productive nominal *-ness* and unproductive verbal *en-*). The crucial assumption now is that the number of hapaxes of a given morphological category correlates with the number of neologisms of that category, so that the number of hapaxes can be seen as an indicator of productivity. Note that we do not claim that a hapax legomenon *is* a

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

neologism. A hapax legomenon is defined with repect to a given corpus. When this corpus is small, most hapax legomena will be well-known words of the language. However, as the corpus size increases, the proportion of neologisms among the hapax legomena increases, and it has been shown that it is precisely among the hapax legomena that the greatest number of neologisms appear (Baayen & Renouf, 1996). From a statistical viewpoint, the hapax legomena play an essential role for gauging the probability that new forms will be encountered that have not been observed before in the corpus.

This approach to measuring morphological productivity receives strong support from the fact that high-frequency words (e.g. *happiness*) are more likely to be stored in the mental lexicon than are low-frequency words (e.g. *pretensionlessness*: see Rubenstein & Pollack, 1963; Scarborough et al., 1977; Whaley, 1978). Baayen and Renouf write that

> [i]f a word-formation pattern is unproductive, no rule is available for the perception and production of novel forms. All existing forms will depend on storage in the mental lexicon. Thus, unproductive morphological categories will be characterized by a preponderance of high-frequency types, by low numbers of low-frequency types, and by very few, if any, hapax legomena, especially as the size of the corpus increases. Conversely the availability of a productive word-formation rule for a given affix in the mental lexicon guarantees that even the lowest frequency complex words with that affix can be produced and understood. Thus large numbers of hapax legomena are a sure sign that an affix is productive. (Baayen & Renouf, 1996: 74)

Having established the significant role of hapaxes in the determination of productivity, we can use the number of hapaxes and the number of tokens to calculate a derived measure of productivity known as 'productivity in the narrow sense', defined as the quotient of the number of hapax legomena n_1 with a given affix and the total number of tokens N of all words with that affix:

(2) $P = n_1^{\text{aff}} / N^{\text{aff}}$

Baayen & Lieber (1991: 809–10) explain the idea behind P as follows. 'Broadly speaking, *P* expresses the rate at which new types are to be expected to appear when *N* tokens have been sampled. In other words, *P* estimates the probability of coming across new, unobserved types, given that the size of the sample of relevant observed types equals *N*.'

Although there are certain problems involved in the sampling of relevant tokens and types (see Plag, 1999: chapters 2 and 5 for discussion), the productivity P of an affix can be calculated and interpreted in a rather straightforward fashion. A large number of hapaxes leads to a high value of P, thus indicating a productive morphological process. Conversely, larger numbers of high-frequency items lead to a high value of N, hence to a decrease of P, indicating low productivity. These results seem to be exactly in accordance with our intuitive notion of productivity, since high frequencies are indicative of the less-productive word-formation processes (Anshen & Aronoff, 1988; Baayen & Lieber, 1997; Plag, 1999: chapter 5).

4 Results

Having laid out the methodological and theoretical foundations for the present study we may now turn to the results. In section 4.1 we will first develop some hypotheses concerning the relationship between lexical richness, lexical growth, and derivational morphology, and then look at the contribution of individual morphological categories to the overall vocabulary size and growth in different registers in section 4.2. We then consider the differences between these morphological categories (section 4.3), before section 4.4 presents differences across categories and registers. Section 4.5 summarizes the results and discusses the implications of the findings.

4.1 The contribution of derived words to overall vocabulary size and growth

In figure 1, we have plotted the vocabulary growth in the three subcorpora of the BNC, irrespective of morphological complexity, using Kilgarriff's frequency list, which implies a type definition in which each wordform–tag combination represents a type. The graph shows how the vocabulary size, i.e., the number of types, shown on the vertical axis, increases as one reads through the tokens of the corpus, plotted on the horizontal axis. The number of types plotted is not the actual number of types in the corpus. The corpus and its subcorpora consist of large numbers of unrelated text fragments, i.e. they have no intrinsic textual order. What we plot, then, is the expected vocabulary size E[V(N)], the number of different types one may expect to count on average for a great many different orderings of the text fragments in a given subcorpus.[10] By means of the resulting vocabulary growth curves we can easily compare the three subcorpora of the BNC for a range of different values of corpus sizes N. In this way, we avoid the problem of having to compare directly the small subcorpora of spoken English with the large subcorpus of written English, without distortion due to the substantial differences in the overall sizes of the subcorpora (see table 1) and the concomitant substantial differences in V and P. Figure 1 also plots 95 percent confidence intervals around the vocabulary growth curves, for technical reasons up to half the total subcorpora sizes. A 95 percent confidence interval for V gives the range of values that V is most likely to have when calculated for new corpora of the same design and size. These confidence intervals are very small, leading to a hardly visible narrow band around the curves of V. For expository reasons, the plot for W (89 million tokens) breaks off at 10 million tokens sampled, because the two spoken corpora already end at c. 4.2 and 6.2 million tokens, respectively. Thus, after having read (technically: 'sampled') for example 2 million word tokens, the W corpus exhibits approximately 100,000 different word

[10] We have used binomial interpolation for the estimation of vocabulary growth and size. Binomial interpolation is a technique for estimating the vocabulary size V(M) for corpus sizes $M < N$. It assumes that words appear randomly and independently in texts, and that each word has a fixed and unvarying probability of being sampled. See Baayen (1996) for a detailed discussion of the statistical problems involved with the application of binomial interpolation to running texts.

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

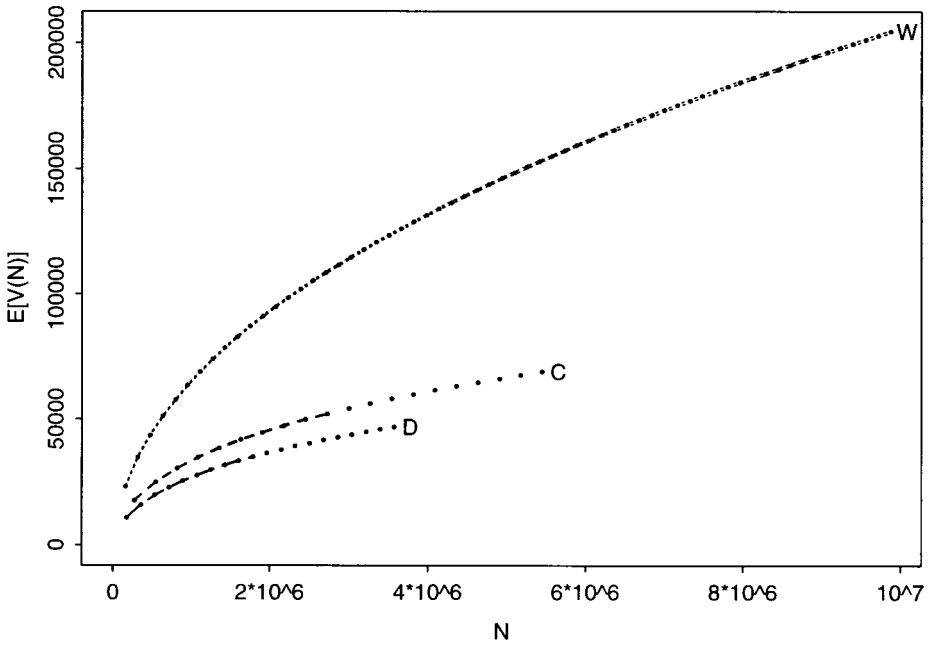

Figure 1. The average number of types E[V(N)] (calculated by means of binomial interpolation) as a function of the number of tokens N for the written subcorpus (W), for the context-governed subcorpus (C), and for the demographic subcorpus (D). 95 percent confidence intervals are also plotted for all three registers, for technical reasons up to half the size of the subcorpora C and D. (2*10^6 = 2,000,000; 4*10^6 = 4,000,000; . . . ; 10^7 = 10,000,000)

types, whereas the context-governed corpus (C corpus) and the demographic corpus (D corpus) exhibit less than half the vocabulary size at that point of sampling. The differences between the corpora are all statistically highly significant.

The differences in vocabulary growth as plotted in figure 1 empirically confirm the assumption about written and spoken registers that can be found in the literature, namely that written registers are lexically much richer than spoken registers. What has this to do with morphology? As already pointed out earlier, Chitashvili & Baayen (1993) claim that vocabulary growth in large texts is primarily due to derivational morphology. If this claim is correct, one can make the prediction that the differences between the three registers as given in figure 1 result from differences in the productivity of derivational morphology. We thus hypothesize that in spoken registers, derivation is much less productive (at least in terms of extent of use V) than in written registers, and that in context-governed speech, productivity is higher than in everyday conversations. Although these hypotheses are intuitively highly plausible, no detailed empirical description is available to confirm or refute them. As will be shown in the following sections, the prediction is confirmed by the BNC data.

Table 2. *Distribution of* -free, -like, -type *in the three subcorpora of the BNC*

Affix	demographic			context-governed			written		
	V(N)	N	n_1	V(N)	N	n_1	V(N)	N	n_1
-free	4	4	4	9	19	6	415	2297	238
-like	12	13	12	26	41	21	1713	9700	1071
-type	3	3	3	11	12	10	689	1209	574

4.2 The contribution of individual morphological categories to vocabulary size

The behavior of the fifteen suffixes under investigation is not uniform. First there is a group of suffixes which are only widely used in written language and hardly ever occur in spoken registers: *-type, -like,* and *-free*. Table 2 summarizes the relevant figures for the three suffixes in the three corpora.

Table 2 shows that *-like* is not only widely used (V = 1713), but also that it is massively used to coin new words, as indicated by the high number of hapaxes ($n_1 = 1071$). In fact, *-like* has the highest number of hapaxes of all suffixes under investigation in the W corpus. This shows that the lack of productivity in the spoken corpora cannot be attributed to structural factors (i.e. productivity restrictions imposed by the grammar), a fact to which we will return in the discussion in section 5.

The other two suffixes in this group are also undoubtedly productive in the narrow sense in the W corpus, but not in the spoken registers. For example, *-type* is among the four most highly productive suffixes ($n_1 = 574$) we investigated, and *-free* ($n_1 = 238$) is in the same range as *-ize* ($n_1 = 212$), *-less* ($n_1 = 272$), and *-ish* ($n_1 = 262$). For information on V, N and n_1 for all affixes, the reader may consult table A in the appendix. To summarize, there is a group of three suffixes which almost exclusively occur in written texts, which are clearly productive, and nevertheless hardly appear in the spoken domains.

The majority of the suffixes in our study form a group in which each individual suffix shows significant differences in the extent of use across all three corpora. This group consists of *-able*, (partitive) *-ful,*[11] *-ion, -ist, -ity, -ize, -ness* and *-less*. We have chosen the plots for *-able*, *-ize*, and *-ion* to illustrate the kind of differences between the subcorpora. The large dots represent the average number of types E[V(N)] for N tokens, with N ranging from 0 to the total number of tokens observed for a given affix in a given subcorpus of the BNC. The small dots represent 95 percent confidence intervals. Two curves can be regarded as significantly different, if one is outside the confidence interval of the other.

[11] The adjective-forming suffix *-ful* (e.g. *beautiful*) is unproductive in terms of any of the productivity measures in all three corpora.

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

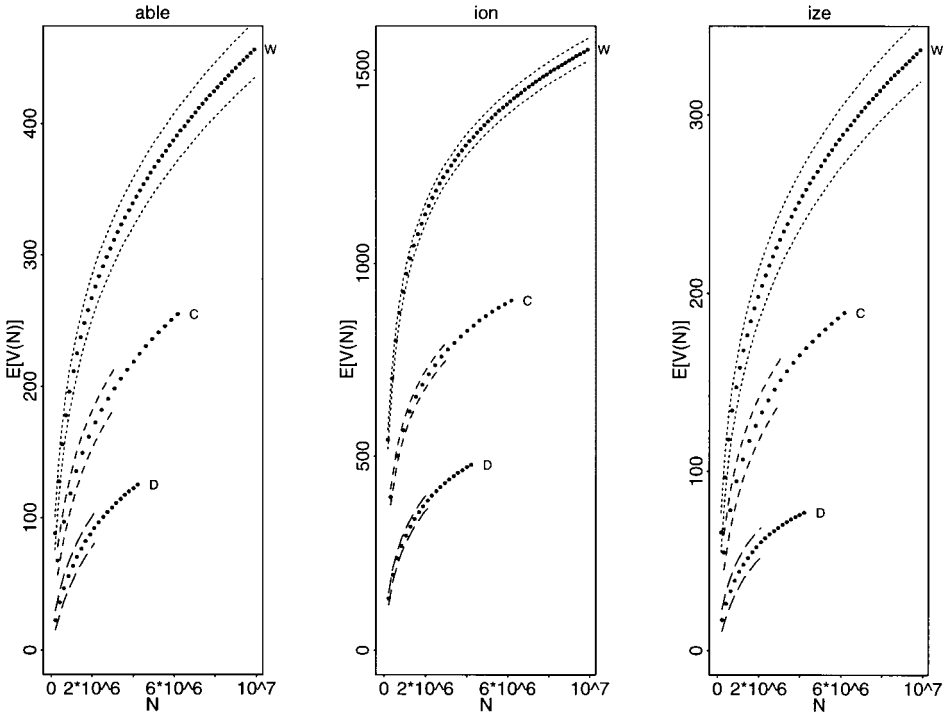

Figure 2. The number of types E[V(N)] as a function of N (large dots) for the three suffixes *-able*, *-ion*, *-ize*. 95 percent confidence intervals are shown by means of small dots

Finally, there are three suffixes that each show a peculiar patterning across registers, *-wise, -ish,* and *-er*. Their vocabulary growth curves are plotted in figure 3. We will discuss each in turn.

The suffix *-wise* contrasts with all suffixes mentioned so far in that it is at least as productive in spoken as in written registers. The growth curve for the C corpus moves out of the confidence interval of the W corpus, which means that it is significantly more widely used in context-governed speech than in written language. Although the number of observations is rather small, it comes out clearly that *-wise* is a counterexample to the general claim that derivational affixes are more productive in written than in spoken language.

Moving on to *-ish*, we can state that it is the only suffix which is used significantly more extensively in every-day conversations than in context-governed speech. Still it is significantly less productive than in the W corpus.

We end our discussion of register differences of individual suffixes with some remarks on *-er*, which also shows an idiosyncratic patterning. It appears to be more productive in the spoken registers. However, the shape of the curves suggests that the small sizes of the spoken corpora may distort the first impression that the right panel of figure 3 leaves us with. Upon closer inspection, the growth curves of the

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

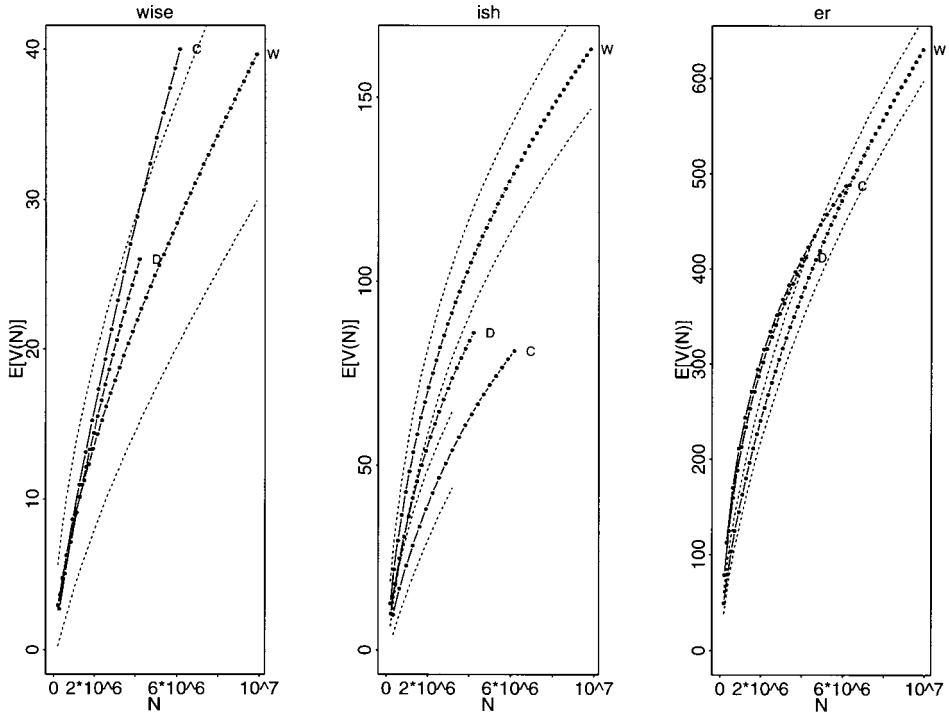

Figure 3. The number of types E[V(N)] as a function of the number of tokens N for the suffixes *-wise*, *-ish*, *-er*, with 95 percent confidence intervals

vocabularies of the C and D corpora at 4.2 and 6.2 million words, respectively, grow less quickly than that of the written language. The curve of the latter is very steep, while those of the spoken registers suggest that further sampling would lead to an even further flattening of the curves. This indicates that, given a larger spoken corpus, *-er* would emerge as less productive in speech than in writing.

4.3 Differences between suffixes

In this section we present a comparison of the contribution of individual suffixes to the growth of the vocabulary of a given subcorpus as a whole. Again, we have to face the problem that our subcorpora have substantially different sizes. We have solved this problem by comparing the subcorpora for the largest range of token sizes N that they have in common. Since the demographic corpus is the smallest subcorpus, our range is $[1,N_D]$. For this range, we calculated the expected number of types with a given suffix at twenty equally spaced intervals. This provided us with the information how many types in our fifteen affixes are expected to occur after reading 210,000, 420,000, 630,000, …, 4,200,000 word tokens of each of our three domains. For each of these domains, we then calculated the average number of types

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

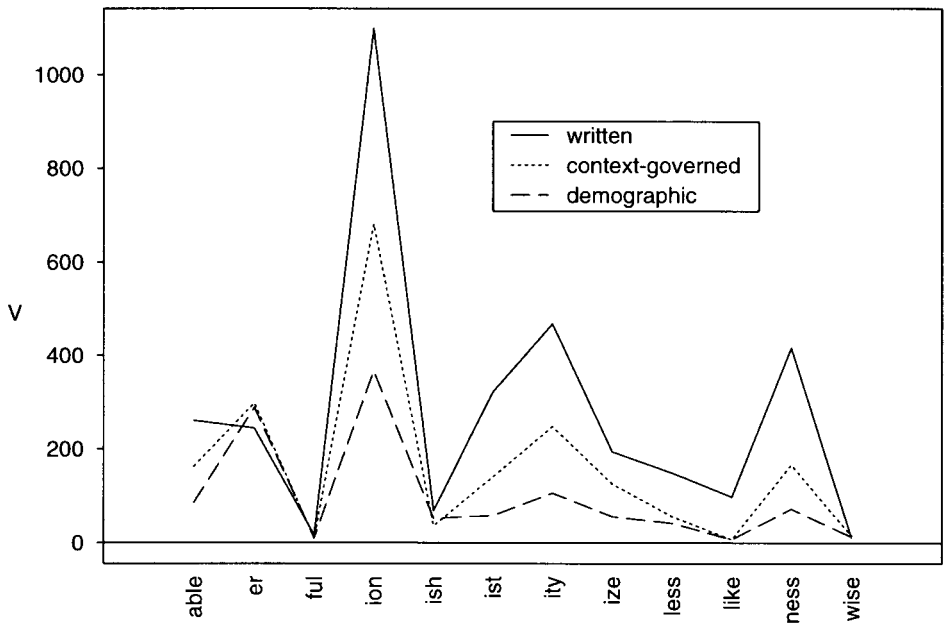

Figure 4. Average number of types V as a function of affix and register, averaged over twenty equally spaced measurement points in the interval [0, 4.2 million]

and the average growth rate P ('productivity in the narrow sense', see above). By taking averages instead of, for example, the final values at N = 4.2 million, we take the shape of the vocabulary growth curve into account as well, instead of only the V and P values for N = 4.2 million. Figures 4 and 5 visualize the results obtained for our selection of suffixes. Figure 4 plots for each affix (on the horizontal axis) the number of types averaged over our twenty measurement points, using solid lines for the written texts, dotted lines for the context-governed spoken language, and dashed lines for the demographic spoken language. Figure 5 presents a similar plot, but now averaging over the twenty P-values.

We can see two things. Firstly, the suffixes tend to yield more types in the written than in the spoken registers. They can be used as quantitative markers of this dimension of register variation. They clearly differentiate our three registers: the graph of the written texts is almost always above that of the context-governed spoken language, which in turn tends to be above that of the demographic texts. Secondly, the suffixes differ considerably in the extent to which they contribute to vocabulary size. Derived nouns clearly make a much larger contribution than the other patterns. *-able* and *-ize* are the runners-up. Other suffixes, like *-ful, -ish* and *-wise*, contribute very little to the overall vocabulary size.

Turning to figure 5, we now plot on the vertical axis the probability P that a new type with a given affix is sampled after having read N tokens of a given subcorpus, again averaged over twenty measurement points. This figure allows us to gauge to

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

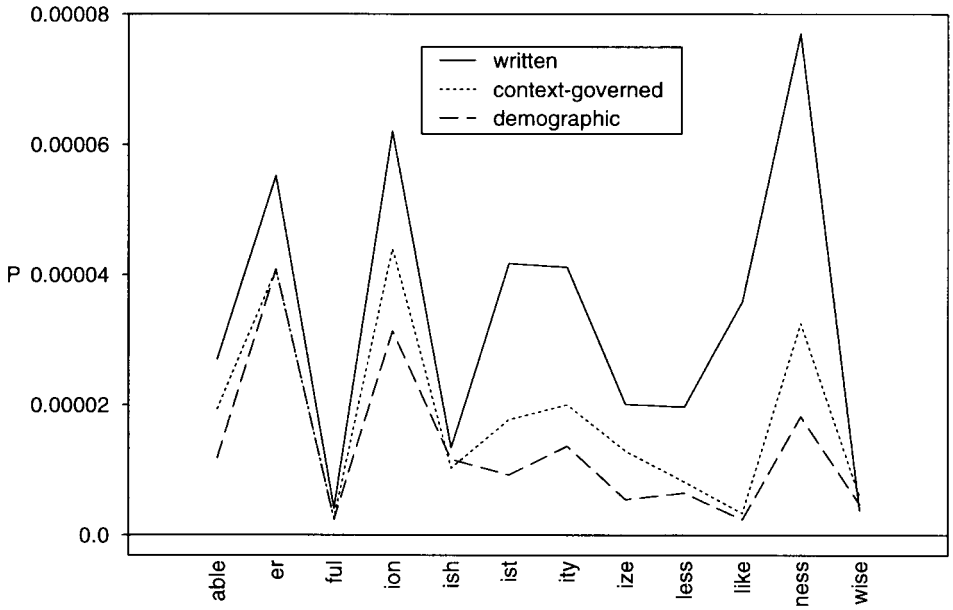

Figure 5. Average growth rate P as a function of affix and register, averaged over twenty equally spaced measurement points in the interval [0, 4.2 million]

what extent the various suffixes may be expected to give rise to new, unobserved types when the subcorpora are increased.

Comparing figure 5 to figure 4 we notice that different aspects of productivity are highlighted. Although the shape of the diagram differs considerably, the differences between the registers are largely preserved.

Figure 5 shows far more pronounced peaks for all nominal suffixes except *-ion*. While *-ion* nominals are more widely used than others (cf. figure 4), *-ness* is more likely to be used in coining new words. *-ion* is also strong in coining new words, but the P values of *-ity, -ist,* and *-er* are closer to the value of *-ion* than is the case with the respective V values in figure 4. The values for *-er* in figure 4 reflect the interesting growth curves for this suffix discussed above (figure 3, right-hand panel): the mean value of V (written) is smaller than the mean values of the spoken corpora. The P values in figure 5 on the other hand show the greater potential of *-er* to form new words in the written language.

4.4 Different suffixes across different registers

Theoretically oriented studies on the productivity of derivational affixes have not called attention to the influence of register on productivity. In other words, whatever the productivity measure employed, the results have been interpreted to express the degree of productivity of affix X 'as such'. Our study shows, however, that the

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

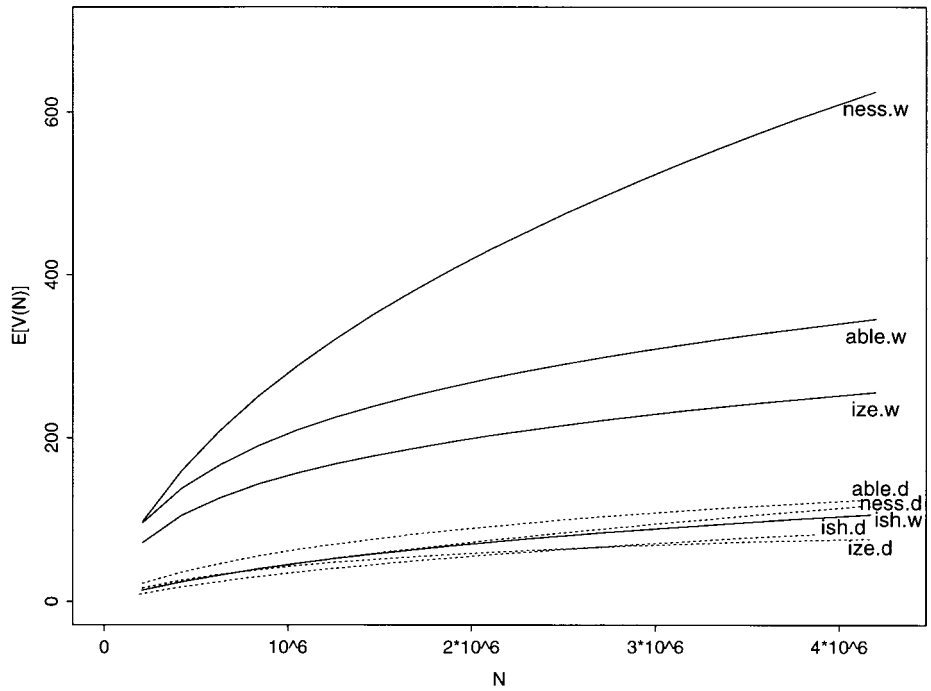

Figure 6. The expected vocabulary size E[V(N)] for *-ness, -able, -ize*, and *-ish* in the written and demographic subcorpora of the BNC as a function of the size in tokens N of these subcorpora

degree of productivity of one and the same suffix may vary according to which register we are looking at. This variation may have the peculiar consequence that in register X suffix A may be more productive than suffix B, whereas in register Y the reverse is the case. We will illustrate this point with the suffixes *-able, -ize, -ish,* and *-ness* and the W corpus and the D corpus, simply by looking at their extent of use V.

Consider figure 6, which plots the expected vocabulary size as a function of the size of the written and demographic subcorpora. Thus, figure 6 shows that we may expect some 200 types in *-able* in the written subcorpus for a sample comprising 1 million tokens. For *-ness*, on the other hand, 1 million tokens yield nearly 300 different types. Interestingly, saying that *-ness* is 'more productive' than *-able* is accurate only as long as we are solely looking at the W corpus, where it is more productive than *-able*. But in the spoken language of the demographic subcorpus, *-able* is slightly more productive than *-ness* in terms of its extent of use. Overall, the productivity of *-ness* in W and D seems to straddle the productivity of *-able* in both subcorpora. Thus it makes little sense to state categorically that *-ness* is more productive than *-able*.

Concerning the suffixes *-ish* and *-ize* we can also observe the reversal of their behavior in W and D. While *-ish* is less productive than *-ize* in the W corpus, it is more productive than *-ize* in the D corpus.

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

5 Conclusion

Our results can be summarized as follows. First, we have shown that the productivity of a given suffix may differ across different registers. In fact, the vast majority of the suffixes under investigation behave in this way. Secondly and conversely, it can be stated that registers differ in the amount of derivational morphology being used. Thirdly, the register-related patterning of the suffixes is not uniform.

How can this kind of hitherto undocumented register variation be explained? We can offer a functional explanation for the high productivity of abstract nouns in the written language. Derivational morphology has two important functions, among others. The first of these is the so-called reference function, i.e. the condensation of information for the purposes of facilitating reference to things mentioned in the previous discourse. The second, that is, the so-called labeling function, is the creation of a (new) name for an entity or an event (see Kastovsky, 1986 for more detailed discusssion, though couched in different terminology). The following example from Kastovsky (1986: 595) illustrates the referential function:

(3) ... and whether your own conversation doesn't sound a little *potty*. It's the *pottyness*, you know, that's so awful.

Baayen & Neijt (1997) have shown that the referential function is typical of certain kinds of abstract nouns, for example Dutch *-heid*, which is more or less equivalent to English *-ness*. Since the referential function is frequently needed in written discourse, this can explain both the extensive use and the productivity in the narrow sense of nominalizations in the corpus. What lies behind this phenomenon is undoubtedly the different conditions under which oral and written texts are produced and perceived (cf. Tannen, 1985: 128). With its strong anchoring in physical context, orality has other means of maintaining reference (establishing common ground, paralinguistic possibilities, prosody, deixis) whereas in writing lexical, morphological and syntactic structure *alone* have to do the job (e.g. Chafe, 1985).

It may well be the case, though, that nominalizing suffixes do not all behave in exactly the same way. As noted above, in their article Baayen and Neijt refer to the Dutch nominalizing suffix *-heid*. It seems, though, that other nominal suffixes may more readily be used in their labeling function. For example, derivatives ending in *-ity* are very often found in technical or scientific texts, where they are used to encode field- or domain-specific concepts. This clearly is a question for further research.

With morphological categories other than nominalization, explanations are difficult to find. What our study shows is that structural restrictions cannot explain the register variation within one morphological category. It is difficult to envisage what structural constraints would restrict the possibility of coining and using words ending in *-like* to the written modality, for example. In general terms, all suffixes that significantly differ in productivity across registers pose a problem for exclusively structural explanations of productivity.

This finding would seem to add a new dimension to the discussion of productivity

https://doi.org/10.1017/S1360674399000222 Published online by Cambridge University Press

restrictions, a discussion which so far has been conducted predominantly on the structural plane. With reference to English derivation the debate has centered on morphonological, morphosyntactic, and morphosemantic concerns (see e.g. Plag, 1999). The results of our study suggest, however, that pragmatic or cultural factors are also of considerable importance (see also Baayen, 1994).

The problem now is to determine the nature of these factors. In the field of evaluative morphology, which suggests itself as a promising research area in this respect, Dressler & Merlini Barbaresi (1994) and Schneider (1997) have provided important insights. For example, Schneider (1997) shows that the productivity of diminutives in English depends on the type of discourse and illocution. Thus, diminutives are most likely to occur in phatic communication and vocative speech acts (subject to further pragmatic constraints).

'Speech situation' is also an important factor in Merlini Barbaresi & Dressler (1994). In addition, they point out that the degree to which the pragmatic meanings of morphological processes are derivable from their semantic meanings may vary. Consequently, certain morphological rules cannot be fully described unless an autonomous pragmatic meaning is postulated.

Our findings suggest that more prototypical examples of derivation than those addressed by Schneider, Merlini Barbaresi, and Dressler are equally susceptible to the influence of pragmatic constraints. The challenge for future research is to extend the study of the pragmatics of morphology to a broader range of morphological categories, and to study in greater detail how context and cotext affect the use of complex words.

Authors' addresses:

Ingo Plag
Englisches Seminar
Universität Hannover
Königsworther Platz 1
D-30167 Hanover
Germany
e-mail: plag@mbox.anglistik.uni-hannover.de

Christiane Dalton-Puffer
Institut für Anglistik und Amerikanistik
Universität Wien
Spitalgasse 2, Hof 8
A-1090 Vienna
Austria
e-mail: christiane.dalton-puffer@univie.ac.at

Harald Baayen
Max-Planck-Institut für Psycholinguistik
Wundtlaan 1
NL-6525 XD Nijmegen
The Netherlands
e-mail: baayen@mpi.nl

Appendix: List of suffixes and their frequencies across the BNC subcorpora

V(N): number of types, N: number of tokens, n_1: number of hapax legomena

Affix	demographic			context-governed			written		
	V(N)	N	n_1	V(N)	N	n_1	V(N)	N	n_1
-able	125	815	49	255	5021	89	933	140627	311
-er (noun)	412	3160	171	487	5157	189	1823	40685	792
-free	4	4	4	9	19	6	415	2297	238
-ful ('measure')	18	76	8	21	94	15	136	2615	60
-ful ('property')	53	1820	13	75	3753	15	154	77316	22
-ion	479	5620	131	905	50607	183	2392	1369116	524
-ish	86	218	46	81	232	45	491	7745	262
-ist	87	501	38	227	2583	77	1207	98823	354
-ity	149	1372	57	349	15468	89	1372	371747	341
-ize	77	1293	23	189	3617	57	658	100496	212
-less	61	335	27	93	609	31	681	28340	272
-like	12	13	12	26	41	21	1713	9700	1071
-ness	118	918	76	310	4037	159	2466	106957	943
-type	3	3	3	11	12	10	689	1209	574
-wise	26	45	19	40	98	31	183	2091	128

References

Akimoto, M. (1991). Deverbal nouns in grammar and discourse. In della Volpe, A. (ed.), *17th LACUS Forum 1990*. Linguistic Association of Canada and the USA. 281–90.

Anshen, F. & M. Aronoff (1988). Producing morphologically correct words. *Linguistics* **26**: 641–55.

Baayen, H. (1992). A quantitative approach to morphological productivity. In Booij, G. & J. van Marle (eds.), *Yearbook of morphology 1991*. Dordrecht: Kluwer. 109–49.

Baayen, H. (1993). On frequency, transparency, and productivity. In Booij, G. and J. van Marle (eds.), *Yearbook of morphology 1992*. Dordrecht: Kluwer. 181–208.

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

Baayen, H. (1996). The effects of lexical specialization on the growth curve of vocabulary. *Computational Linguistics* **22**: 455–80.

Baayen, H. & A. Neijt (1997). Productivity in context: a case study of a Dutch suffix. *Linguistics* **35**: 565–87.

Baayen, H. & R. Lieber (1991). Productivity and English derivation: a corpus-based study. *Linguistics* **29**: 801–44.

Baayen, H. & R. Lieber (1997). Word frequency distribution and lexical semantics. *Computers and the Humanities* **30**: 281–91.

Baayen, H. & A. Renouf (1996). Chronicling the *Times*: productive lexical innovations in an English newspaper. *Language* **72**: 69–96.

Beaugrande, R. de & W. U. Dressler (1981). *Introduction to text linguistics.* London: Longman.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation*. Cambridge: Cambridge University Press.

Burnard, L. (ed.) (1995). *Users' reference guide for the British National Corpus*. Oxford University Computing Service.
Chafe, W. L. (1985). Linguistic differences produced by difference between speaking and writing. In Olsen, D. R., N. Torrence & A. Hildyard (eds.), *Literacy, language and learning. The nature and consequences of reading and writing.* Cambridge, New York: Cambridge University Press. 105–23.
Chafe, W. L. & J. Danielewicz (1986). Properties of spoken and written language. In Horowitz, R. & S. J. Samuels (eds.), *Comprehending oral and written language.* New York: Academic Press.
Chitashvili, R. & H. Baayen (1993). Word frequency distributions. In Altmann, G. & L. Hrebicek (eds.), *Quantitative text analysis.* Trier: Wissenschaftlicher Verlag. 54–113.
Dressler, W. U. & M. Barbaresi (1994). *Morphopragmatics. Diminutives and intensifiers in Italian, German, and other languages*. Berlin, New York: Mouton de Gruyter.
Enkvist, N. E. (1977). Stylistics and text linguistics. In Dressler, W. U. (ed.), *Current trends in textlinguistics*. Berlin, New York: de Gruyter. 174–90.
Ferguson, C. A. (1994). Dialect, register and genre: working assumptions about conventionalization. In Biber, D. & E. Finegan (eds.), *Sociolinguistic perspectives on register*. New York: Oxford University Press. 15–30.
Indra, W. (1990). Word-formation and text-cohesion. Unpublished MA thesis, University of Vienna.
Kastovsky, D. (1982). Word-formation. A functional view. *Folia Linguistica* **16**: 181–98.
Kastovsky, D. (1986). The problem of productivity in word formation. *Linguistics* **24**: 585–600.
Kastovsky, D. & B. Kryk-Kastovsky (1997). Morphological and pragmatic factors in text cohesion. In Ramisch, H. & K. Wynne (eds.), *Language in time and space. Studies in honour of Wolfgang Viereck*. (ZDL Beihefte 97). Stuttgart: Steiner. 462–75.
Koch, P. & W. Oesterreicher (1994). Schriftlichkeit und Sprache. In Guenter, H. & O. Ludwig (eds.), *Schrift und Schriftlichkeit*. (HSK 10.1). Berlin, New York: Mouton de Gruyter.
Lipka, L. (1987). Word-formation and text in English and German. In Asbach-Schnitker, B. & J. Roggenhofer (eds.), *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik. Festgabe für Herbert E. Brekle zum 50. Geburtstag.* Tübingen: Narr. 59–67.
Plag, I. (1999). *Morphological productivity: structural constraints in English derivation*. Berlin, New York: Mouton de Gruyter.
Rubenstein, H. & I. Pollack (1963). Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior* **2**: 147–58.
Scarborough, D., C. Cortese & H. S. Scarborough (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance* **3**: 1–17.
Schneider, K. P. (1997). *'Size and Attitude'. Expressive Wortbildung und diminutivische Ausdrücke in der englischen Alltagskommunikation.* Habilitationsschrift, Philipps-Universität Marburg.
Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language* **58**: 1–21.
Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In Olsen, D. R., N. Torrence & A. Hildyard (eds.), *Literacy, language and learning. The nature and consequences of reading and writing.* Cambridge, New York: Cambridge University Press. 124–47.
Wells, R. (1960). Nominal and verbal style. In Sebeok, T. (ed.), *Style in language*. Cambridge, MA: MIT. 213–20.
Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior* **17**: 143–54.
Zipf, G. K. 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.