Morphological productivity across speech and writing

Ingo Plag, Christiane Dalton-Puffer, Harald Baayen

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 and Lieber 1991, Baayen and Renouf 1995, Baayen and 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 afore-mentioned studies.

Studies in register variation have shown in great detail that there are 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 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. Furhermore, 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.[1]The data for our study come

[1] For ample discussion of the structual aspects of morphological productivity, see Plag 1999.

from the British National Corpus with an overall number of c. 100 million word tokens. 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. We offer some tentative explanations for these findings and discuss their 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% post-1975) with a written/spoken ratio of 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 sub-corpora. The so-called demographic part 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 part 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 catgories: education, business, public/institutional, and leisure. Table 1 gives a general overview of the relative sizes of the various parts of the BNC.

Table 1. Numeric composition of BNC (adapted from Burnard 1995:9)[2]

	number of word tokens
Written	89740544
Spoken Context Governed	6154248
Spoken Demographic	4211216

With over ten million words of spoken language the BNC certainly represents by far the largest source of computerised spoken data available. The well-established and widely use 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, however, 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 commerical and non-commercial institutions in the UK – this division is founded in a long-standing tradition of research into the differences between speech and writing.[3]

Even though the notion of 'typical speech' and 'typical writing' (or 'orality' and 'literacy' following Tannen 1982) continues to be useful and legitimate, it has become

[2] For a detailed account on the compositon and structure of the BNC see Burnard (1995) chapters 3 & 4.

[3] 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.

clear that a strict division between the linguistic characteristics of speech and writing is impossible as the division generalises 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). Other terms are also in use (‘genre’, ‘style’) but this is hardly the place to discuss the implications of the differing terminologies. It is, however, essential to state that register distinctions are not defined in linguistic terms but rest on participant relations, purpose, productions 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*)” (e.g. 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 integrate 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. Beaugrande and Dressler 1981, Kastovsky 1982, Lipka 1987, Akimoto 1991). However, this assumption appears somewhat relativised in view of the fact that the feature ‘nominalization’ appears to load significantly only on one of Biber’s seven factors or “dimensions of register variation” in English. According to Biber nominalization correlates significantly only with the “Elaborated Reference”-end of the dimension “Situation Dependent vs. Elaborated Reference” (Biber 1995:155), which seems to indicate that nominalization does not play a crucial role in register differentiation. Other languages studied in the same volume (Biber 1995), however, suggest that a broader view of word-formation might make a difference after all. For both Korean and Somali several linguistic features used in the analysis are word-formation features which then also appear in various factors. The choice of linguistic features is, after all, based on traditions of linguistic research and the **reason** why certain things are (not) studied are not exclusively determined by the structure of language alone. In other words, one can only find what one is looking for. (See section 2.3 for the choice of derivational features included in this study).

The discussion of 'register' in this section is important also in another respect. It should be clear by now that the structure of the BNC does not per se reflect *linguistic* differences. The one subsection which is possibly most homogeneous in this respect is the Spoken Demographic corpus as it contains only spontaneous conversations. It is not our aim here to study the BNC in terms of linguistically defined text-types (Biber 1995) but it is worth bearing in mind that the contextually defined 'registers' (or 'domains' in BNC terms) and linguistically defined 'text-types' are a case of cross-classification. In other words, we should not expect linguistic features to distribute evenly over the array of text-types assembled under each 'domain' heading. In this respect division by medium (speech-writing) is just as coarse-grained as division by domain. As the interest of the present paper is a first global view of the issue of register variation in word-formation, it was nevertheless thought legitimate to adhere to the pre-defined structure of the BNC. A more fine-grained analysis, however, would have to take into account the dimension of text-type which is not directly represented in the structure of the BNC. A first step into this direction would be to consider separately the imaginative and the informative written texts, since 'imaginative narrative' seems to be a text type which largely coincides with the common notion of 'fiction'. Interestingly, it has been suggested that fiction vs. non-fiction also behave differently with regard to the role of word-formation (Kastovsky & Kryk-Kastovsky 1997:469, Akimoto 1991:282, Indra 1990). Without going into detail, these suggestions mean that generalising over the written part as a whole, as we do in this paper, will downtone the differences between speech and writing. In other words, if we had compared just informative writing with the spoken parts of the BNC we probably would have found even sharper distinctions between the different registers than the ones described below.

2.3 Data

In this section we will discuss the rationale behind our choice of the linguistic features studied and describe the handling of the data. It was decided to focus the study on suffixal derivation. Raw data for thirty-eight suffixes were extracted via string search from the BNC word-frequency lists.[4] In theory, incorporating a word-class tag criterion in data extraction would have enabled us to filter out, for example, verbs like *to partition* from the nominal *–ion* data, thus producing files containing less irrelevant material. However, it turned out that tagging is unreliable for derived lexical items and cannot be employed in the extraction process.[5]

We selected fourteen derivational suffixes of which we expected that they would be at least moderately productive.[6] These suffixes are distributed over the following categories:

(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*

The main criterion for choice was the aim to complement Biber's only derivation-relevant feature 'nominalization by *–ion, -ity, -ness, -ment*' with an array of other derivational patters performing different morphosyntactic and morphosemantic functions. Other criteria such as the time needed for cleaning up the raw data also

[4] **The word-frequency lists were created by Adam Kilgarriff and can be obtained via FTP from the following site: ftp://ftp.itri.bton.ac.uk/pub/bnc.**

[5] 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% of all words are tagged erroneously and that a further 4.7 % of words carry ambiguous (or portmanteau) tags. Though we have not computed any figures and cannot supply percentages, it is clear from our data that derived words seem to attract both erroneous and ambiguous tags to a much greater extent.

[6] The morphological status of some of the items in (1) is perhaps controversial. Thus, derivatives with *-type* or *-like* could be argued to be compounds, and the nature of partitive *-ful* is questionable. It may look like an adjectival suffix, but it forms measure partitive nouns. The structural properties of these morphological categories 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.

played a role. Occurrences of the ornative *–ed* suffix, for instance, are hidden among countless tokens of past tenses and past participles (e.g. *open-ended, one-sided*) and could not be extracted under any reasonable cost-gain ratio.

Due to our 'string-search only' policy the raw data files contained a great deal of irrelevant material. All items that did not belong to the relevant morphological category were removed from the word lists. Where necessary we consulted the OED or checked items in their context in the BNC available on-line.[7] The following criteria were observed in the process:

- Complex lexical items with derivational affixes attached outside the suffixes in question were removed. This decision affected all derived adjectives used as adverbs, prefixed formations (e.g. *unavailable*) and compounds (e.g. *age-specificity*). Inflectional suffixes were ignored so that noun-plurals were subsumed under their respective singulars and verbs with overt inflectional endings were added to the uninflected tokens of the same type.

- In order to count as a token with a given suffix items had to fulfill the following conditions. The most obvious criterion was that semantically it belongs to the morphological category in question. 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, if anything, this skews data on the conservative side by excluding semantically opaque but formally analysable items from further consideration.

The only suffix where this procedure was not strictly followed 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 suffixe *–ster* (e.g. *gangster*), all the synthetic 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 data arising from **the string search** (V=48,476), we had to infringe our 'ignore tagging' principle by removing everything that was tagged as proper noun (NP0). This decision 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 if they were

[7] Simple searches can be conducted at the following web-site: http://thetis.bl.uk

partially wrongly tagged proper names. The results based on the *–er* data are therefore to be interpreted with caution.

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 words 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 processes like e.g. nominalizing *-th* (as in *length*) productivity seems to be a scalar concept. In other words, with some affixes it 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 and Baayen 1993, Baayen and Lieber 1991, Baayen and 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 (number of tokens = N) in the corpus. 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' **sorry to be such a dumbo - which is 'extent of use' now *V* or *I* ??.** 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 the significance of hapaxes for productivity.

In a sufficiently large corpus, the number of hapaxes in general approximates half the observed vocabulary size (e.g. Zipf 1935). Chitashvili and Baayen (1993:57) call this kind of distribution 'Large Number of Rare Events' distribution. They show that the frequency spectrum of whole texts closely resembles the frequency spectrum of productive morphological categories, and that productive morphological categories play a crucial role in anchoring a text in the Large Number of Rare Events zone (Chitashvili and Baayen 1993:126-132). Unproductive morphological categories show a completely different frequency distribution (cf. Chitashvili and Baayen 1993: 80-86, 125-126 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. This assumption receives strong support from the fact that high-frequency words are more likely to be stored in the mental lexicon than are low-frequency words (Rubenstein and Pollack 1963, Scarborough et al. 1977, Whaley 1978). Baayen and Renouf write that

> If 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 and Renouf 1996:74)

Having established the significant role of hapaxes in the determination of productivity, we can use the two principal measures to compute **two derived measures** of productivity, the so-called ‘extent of use’ (*I*) and 'productivity in the narrow sense (*P*)'. Given a suitable text corpus, the extent of use is the quotient of the number of types of a given morphological category sampled and the number of all word tokens sampled.

(2) $I = V^{\mathrm{aff}} / N^{\mathrm{all}}$

I is therefore a measure of how much a certain morphological category contributes to the overall vocabulary size.

Productivity in the narrow sense *P* is 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:

(3) $P = n_1^{\mathrm{aff}} / N^{\mathrm{aff}}$

Baayen and Lieber (1991:809-810) 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: chapter 2 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, large 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 and Aronoff 1988, Baayen and Lieber 1997, Plag 1999: chapter 5).

4. Results

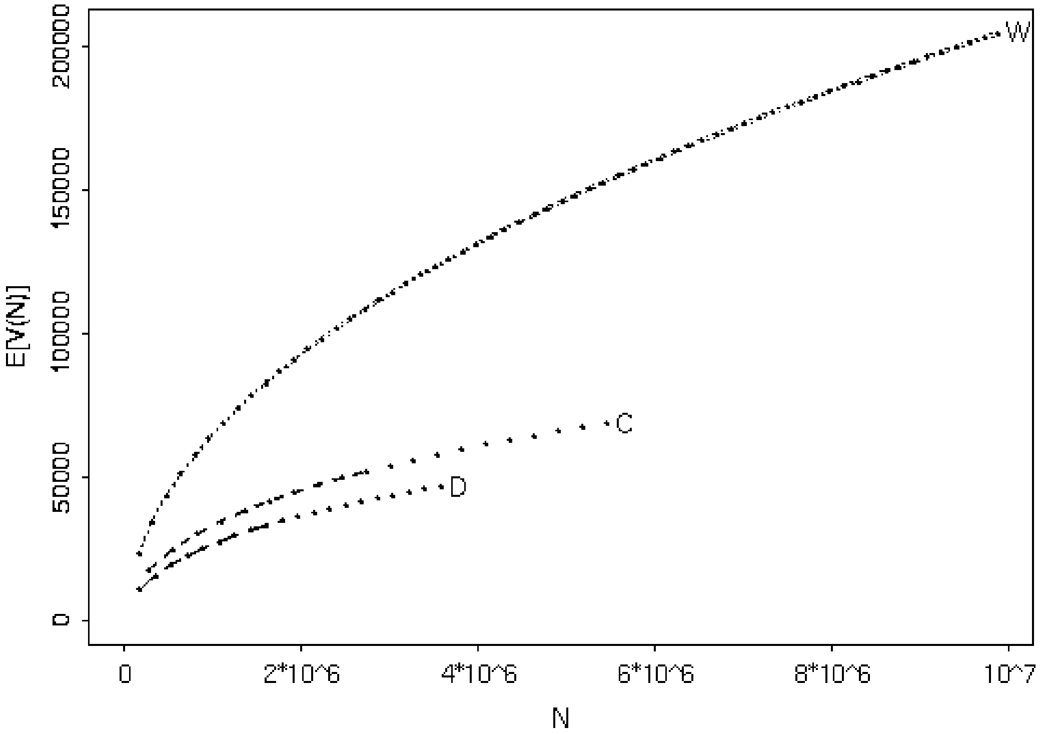
Having laid out the methodological and theoretical foundations for the present study we may now turn the results. In section 4.1 we will first develop some hypotheses concerning the relationship between lexical richness, lexical growth and derivational morphology and then look at the contribution of 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, 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. The graph shows the interpolated increase in the overall vocabulary size V^{all} as one reads through the corpus.[8] Thus, after having read (technically: ‘sampled’) for example 2 million word tokens, the W corpus exhibits approximately 100,000 different word 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 0.05 confidence interval is plotted in broken lines, but is so close to the curve that it is only clearly visible towards the right end of the W corpus curve). Note that, for expository reasons, the plot breaks off at 10 million tokens sampled, because the two spoken corpora end at c. 4.2 and 6.2 million tokens, respectively:

[8] We have use binomial interpolation for the estimation of vocabulary growth and size. Interpolation is appropriate because the BNC consists of a large number of unrelated small texts. See Baayen (1996) for a detailed discussion of the statistical problems involved with the application of binomial interpolation to running texts.

Figure 1



The difference in vocabulary growth as plotted in figure 1 empirically confirms 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 (see section 2 above). What has this to do with morphology? As already pointed out earlier, Chitashvili and 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* **cf my question on p.9 if V IS extent of use - what's happening with I?)** than in written registers, and that in context-governed speech, productivity is higher than in every-day 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.

4.2 The contribution of individual morphological categories to the vocabulary size

The behavior of the 15 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 1. 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

As the table clearly shows, *-like* is not only widely used (*V*=1713), but it is also massively used to coin new words, as is indicated by the high number of hapaxes. 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 4.5.

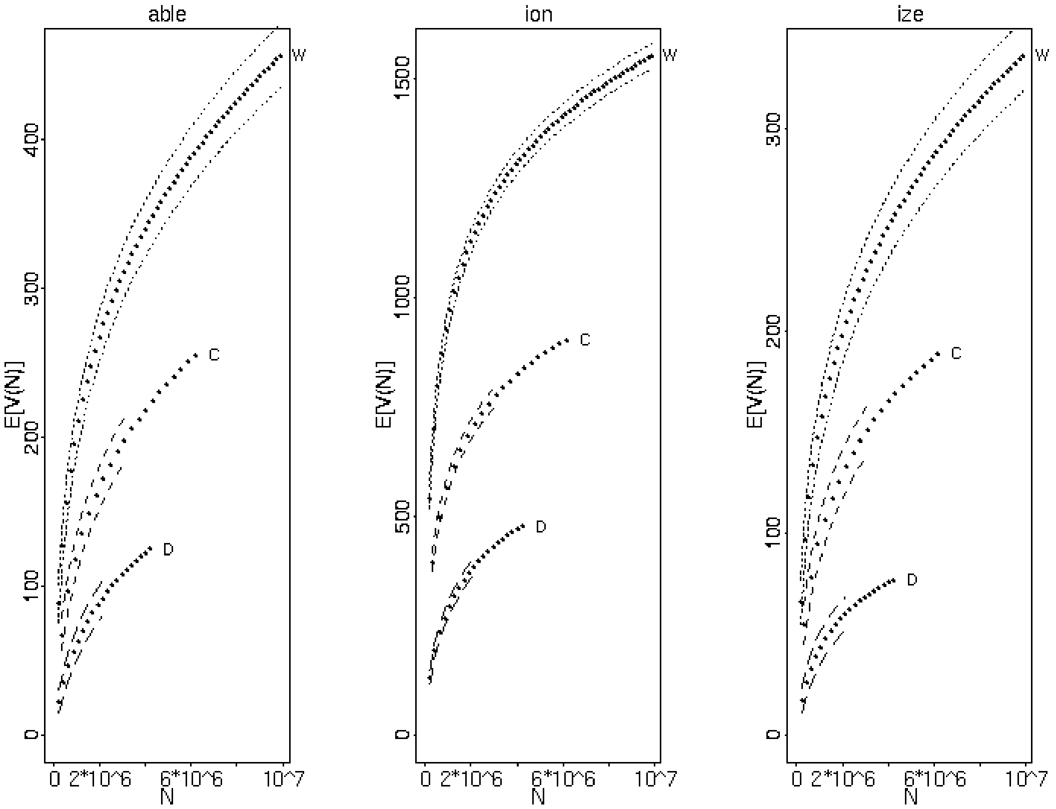
The other two suffixes in this group are also undoubtedly productive in the narrow sense in the W corpus, but not in the spoken registers. For example, *-type* is among the four most highly productive suffixes (n_1=574) we investigated, and *-free* (n_1=238) is in the same range as *-ize* (n_1=212), *-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.

The majority of the suffixes 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,*[9] *-ion, -ist, -ity, -ize, -ness* and *-less*. We have chosen the plots for *-able, -ize,* and *-ion* to illustrate the difference across registers. The plots for the other suffixes look very similar and are omitted for reasons of space.

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

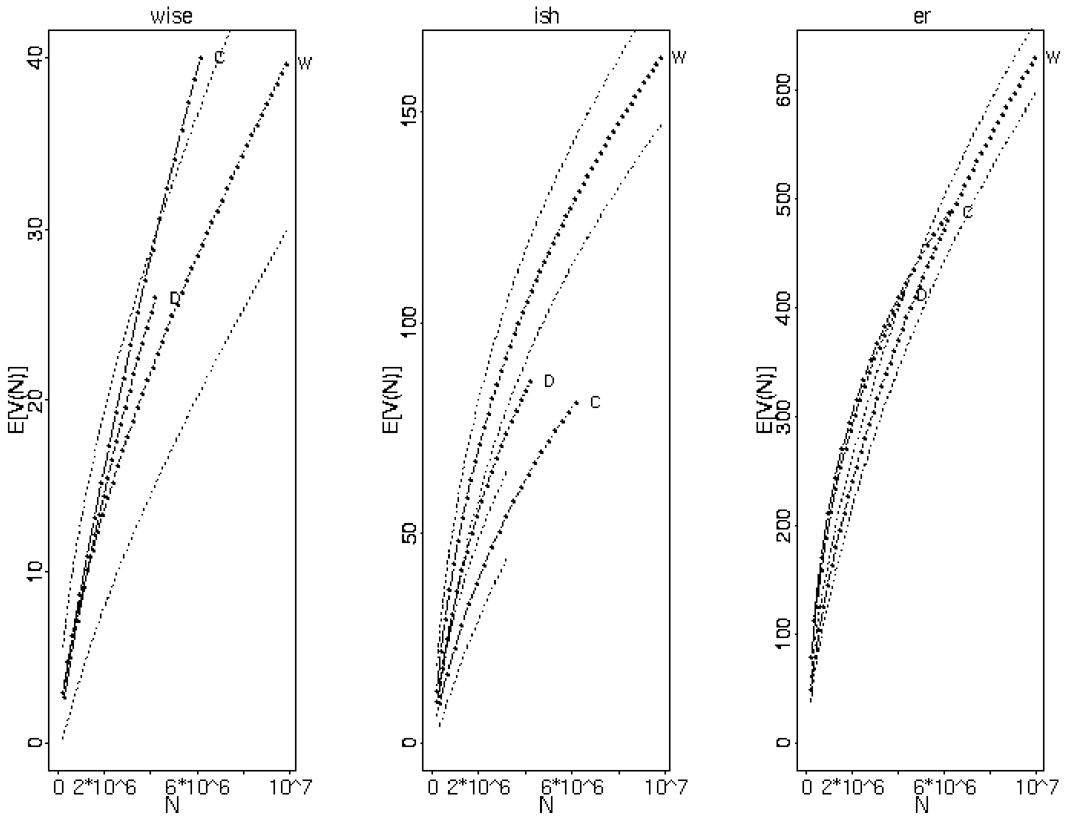
Vocabulary growth is plotted in the same way as in figure 1 above and can be interpreted analogously. The 0.05 confidence intervals are given by broken lines and can be read in such a way that two curves can be regarded as significantly different, if one is outside the confidence interval of the other.

Figure 2. Growth curves of *-able*, *-ion*, *-ize*



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

Figure 3. Growth curves of *-wise*, *-ish*, *-able*



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 is 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 counter-example to the general claim that derivational affixes are more productive in written than in spoken language.

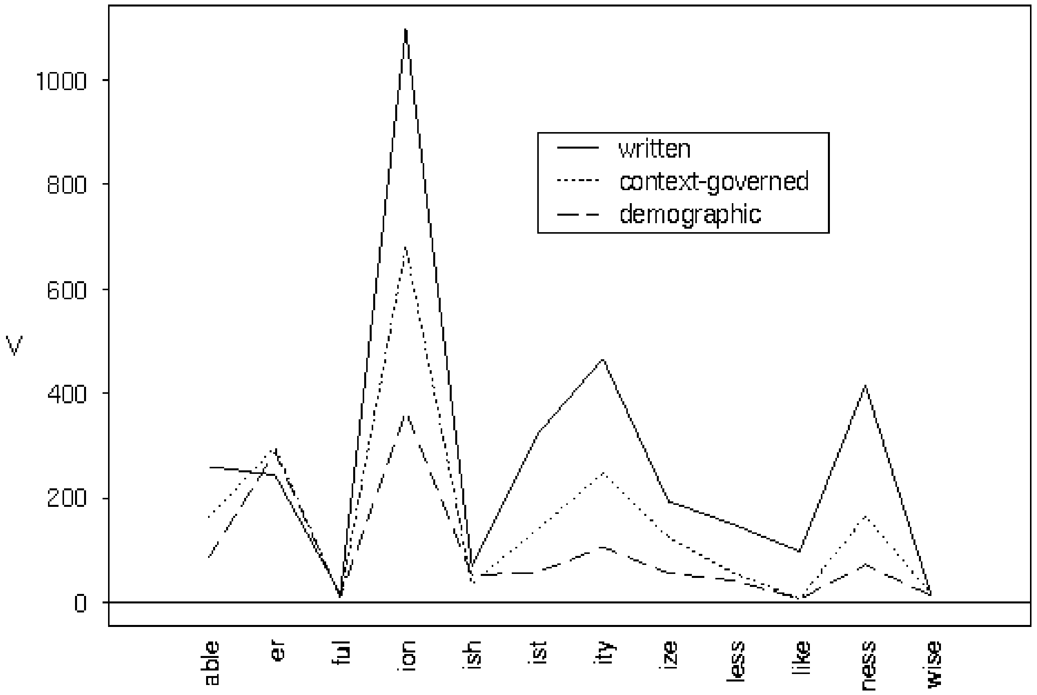
Moving on to *-ish*, we can state that it is the only suffix which is **used** significantly more extensively in every-day conversations than in context-governed speech. Still it is significantly less productive **than** in the W corpus.

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 corpus size is an impeding factor here. Thus the growth rate of the vocabularies of the C and D corpora at 4.2 and 6.2 million words, respectively, suggests that further sampling would lead to a 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

We are now in a position to give an overview of the differences between suffixes.
In the previous section we looked at the suffixes in their own terms, as it were. In this section we present a comparison of the contribution of individual suffixes to vocabulary growth. The problem is of course that the three corpora are not of equal size. It is therefore necessary to stop sampling at the point where the smallest corpus (i.e. the D corpus) ends. In technical terms, the interpolation of the growth curves for each suffix in each of the three corpora was time-logged at c. 4.2 million word tokens (i.e. the size of the D corpus) sampled. The following figure is based on the interpolation plots of individual suffixes as exemplified in figures 2 and 3 above. Figure 4 gives the extent of use of all suffixes across the three corpora. On the *y*-axis we **plotted the *mean* of V and not the absolute number of different types after having sampled through the whole corpus**. The reason for this decision was that the mean does not only reflect vocabulary size but also the shape of the curve, i.e. that is the rate of vocabulary growth.

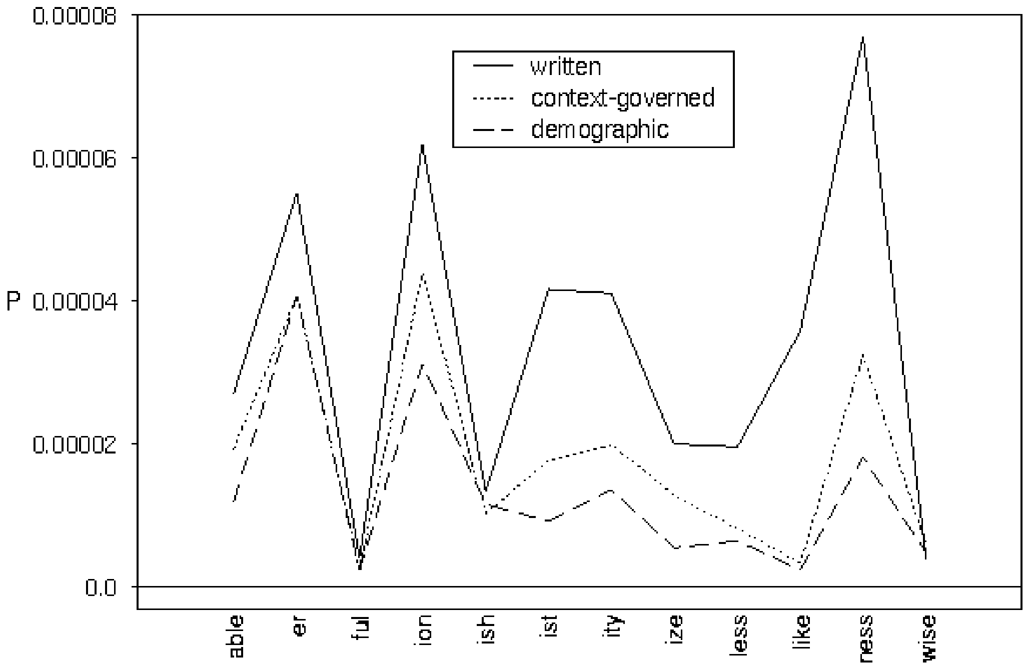
Figure 4



We can see two things: Firstly, the suffixes clearly differentiate register, as already pointed out in the previous section. 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.

Next, we will compare the different suffixes in terms of the *P*-measure, i.e. their productivity in the narrow sense. Recall that this measure estimates the probability of coming across new, unobserved types within the morphological category itself. The interpolation for the hapaxes was again time-logged at c. 4.2 million words.

Figure 5



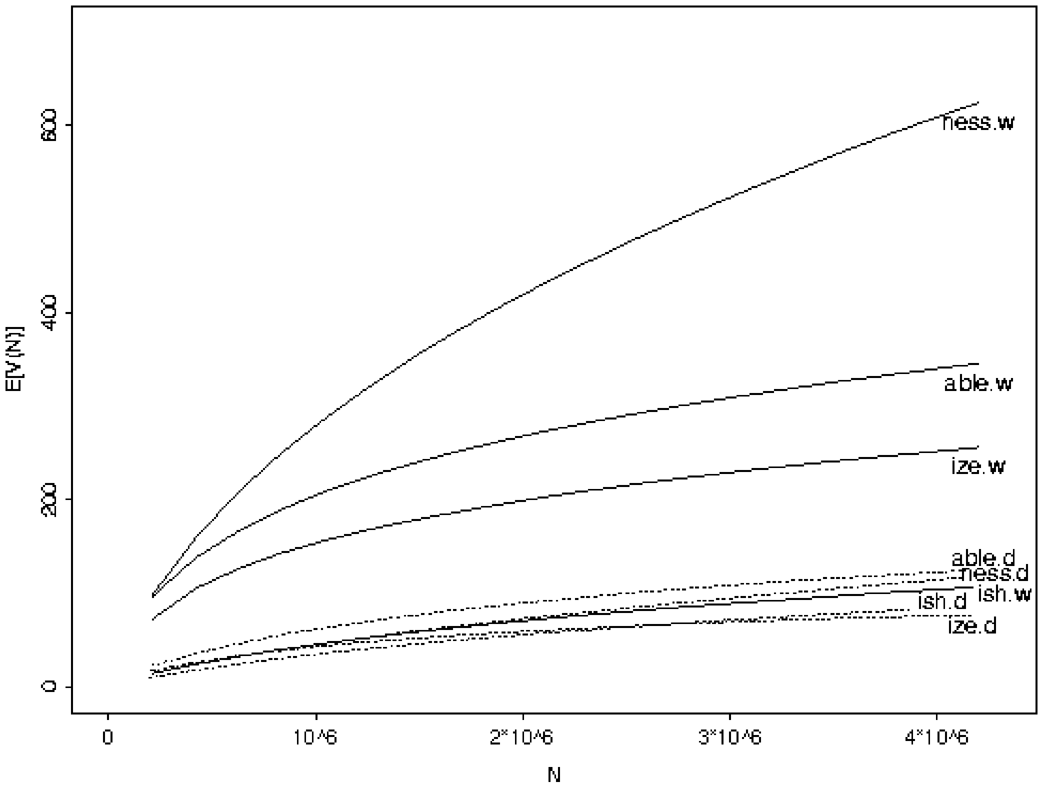
Comparing figure 5 to figure 4 we notice that different aspects of productivity are highlighted. Although the shape of the **diagram** differs considerably, the differences between the registers are largely preserved.

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) *-ity, -ist, -er* and especially *-ness* are more likely to be used in coining new words. The values for *-ness* in particular show that *-ness* has a great potential for the creation of neologisms but that these words are not so widely used (in comparison). The values for *-er* in Figure 4 reflect the problematic 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 on the other hand show **(delete however)** the greater potential to form new words in the written language.

4.4. Different suffixes across different registers

As mentioned in the introduction, work on the productivity of derivational affixes has not distinguished between registers. 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 degree of productivity of one and the same suffix may differ 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 it is the other way round. We will illustrate this point with the suffixes -*able* ,-*ize*, -*ish*, and -*ness* and the W corpus and the D corpus (in terms of extent of use). Consider the following figure:

Figure 6



(deleted: two sentences on able vs ize) For instance, saying that -*ness* is ‘more productive’ than -*able* is accurate only as long as we are solely looking at the W corpus. Overall, the productivity of -*ness* in W and D seems to straddle the productivity of -*able* in both corpora. Thus it makes little sense to state categorically that -*ness* is more productive than -*able*.

Concerning the suffixes -*ish* and -*ize* we can even observe a total reversal of their behaviour 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.

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, i.e. 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:

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

Baayen and 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) whereas in writing lexical, morphological and syntactic structure 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. In their article Baayen and Neijt refer to the Dutch nominalizing suffix *-heid*, the direct equivalent of English *-ness*. It seems, though, that other nominal suffixes may more readily be used in their labeling function. For example, derivatives **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 even less obvious. What is clear, however, is that structural restrictions cannot explain the register variation within one morphological category. It is thus difficult to envisage what structural constraints would restrict the possibility of coining and using words 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 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.

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 and Merlini Barbaresi (1994) and Schneider (1997) have provided important insights. Our findings suggest, however, that **more prototypical examples of** derivation 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.

References

Akimoto, Minoji. 1991. Deverbal nouns in grammar and discourse. In: Angela della Volpe (ed.) *17th LACUS Forum 1990*. Linguistic Association of Canada and the U.S.A., 281-290.

Baayen 1992. A quantitative approach to morphological productivity. In Geerd Booij and Jaap van Marle (eds.) *Yearbook of morphology 1991*. Dordrecht: Kluwer, 109-149.

Baayen, Harald. 1993. On frequency, transparency, and productivity. In Geerd Booij and Jaap van Marle (eds.) *Yearbook of morphology 1992*. Dordrecht: Kluwer, 181-208.

Baayen, Harald. 1994. Derivational Productivity and Text Typology. *Journal of Quantitative Linguistics* 1, 16-34.

Baayen, Harald. 1996. The effects of lexical specialization on growth curve of the vocabulary. *Computation and Linguistics* 22: 455-480.

Baayen, Harald and Anneke Neijt. 1997. Productivity in context: a case study of a Dutch suffix. *Linguistics* 35: 565-587.

Baayen, Harald and Rochelle Lieber. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29: 801-844.

Baayen, Harald and Rochelle Lieber. 1997. Word Frequency Distribution and Lexical Semantics", *Computers and the Humanities* 30, 281-291.

Baayen, Harald and Antoinette Renouf. 1996. Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language* 72.1, 69-96.

Beaugrande, Robert de and Wolfgang U. Dressler. 1981. *Introduction to text linguistics.* London: Longman.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Burnard, Lou (ed.) 1995. *Users' reference guide for the British National Corpus*. Oxford University Computing Service.

Chafe, Wallace L. 1985. Linguistic differences produced by difference between speaking and writing. In David R. Olsen, Nancy Torrence and Angela Hildyard (eds.) *Literacy, language and learning. The nature and consequences of reading and writing.* Cambridge, New York: Cambridge University Press, 105-123.

Chafe, Wallace L. and Jane Danielewicz. 1986. Properties of spoken and written language. In Rosalind Horowitz and S.J. Samuels (eds.) *Comprehending oral and written language.* New York: Academic Press, XX

Chitashvili, R.J. and Harald Baayen. 1993. Word frequency distributions. In G. Altmann and L. Hrebicek (eds.) *Quantitative text analysis.* Trier: Wissenschaftlicher Verlag, 54-113.

Dressler, Wolfgang U. and Merlina Barbaresi. 1994. *Morphopragmatics*

Enkvist, Nils E. 1977. Stylistics and text linguistics. In Wolfgang U. Dressler (ed.) *Current trends in textlinguistics*. Berlin, New York: de Gruyter, 174-190.

Ferguson, Charles A. 1994. Dialect, register and genre: working assumptions about conventionalization. In Douglas Biber & Edward Finegan (eds.) *Sociolinguistic persepctives on register* New York: Oxford University Press, 15-30.

Indra, Walter. 1990. Word-formation and text-cohesion. Unpublished M.A. thesis. University of Vienna.

Kastovsky, Dieter. 1982. Word-formation. A functional view. *Folia Linguistica* 16: 181-198.

Kastovsky, Dieter. 1986. "The Problem of Productivity in Word Formation", *Linguistics* 24, 585-600.

Kastovsky, Dieter & Barbara Kryk-Kastovsky. 1997. Morphological and pragmatic factors in text cohesion. in: Heinrich Ramisch and Kenneth Wynne (eds.) *Language in time and space. Studies in Honour of Wolfgang Viereck*. (ZDL Beihefte 97). Stuttgart: Steiner, 462-475.

Koch, Peter and Wulf Oesterreicher. 1994. Schriftlichkeit und Sprache. In Hartmut Guenter and Otto Ludwig (eds.) *Schrift und Schriftlichkeit*. (HSK 10.1) Berlin, New York: Mouton de Gruyter.

Lipka, Leonard. 1987. Word-formation and text in English and German. In Brigitte Asbach-Schnitker and Johannes Roggenhofer (eds.) *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik. Festgabe fuer Herbert E. Brekle zum 50. Geburtstag.* Tuebingen: Narr, 59-67.

Plag, Ingo. 1999. *Morphological productivity: structural constraints in English derivation*. Berlin, New York: Mouton de Gruyter.

Rubenstein, Herbert and Irwin Pollack. 1963. Word Predictability and Intelligibility. *Journal of Verbal Learning and Verbal Behavior* 2, 147-158.

Scarborough, Don - Charles Cortese - Hollis S. Scarborough. 1977. Frequency and Repetition Effects in Lexical Memory. *Journal of Experimental Psychology: Human Perception and Performance* 3, 1-17.

Schneider, Klaus Peter.1997. *'Size and Attitude'. Expressive Wortbildung und diminutivische Ausdrücke in der englischen Alltagskommunikation*. Unveröffentlichte Habilitationsschrift, Philipps-Universität Marburg.

Tannen, Deborah. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58: 1-21.

Tannen, Deborah. 1985. Relative focus on involvement in oral and written discourse. In David R. Olsen, Nancy Torrence and Angela Hildyard (eds.) *Literacy, language and learning. The nature and consequences of reading and writing.* Cambridge, New York: Cambridge University Press, 124-147.

Wells, Rulon. 1960. Nominal and verbal style. In THomas Sebeok (ed.) *Style in Language*. Cambridge, Mass.: MIT, 213-20.

Whaley, C. P. 1978. Word-nonword Classification Time. *Journal of Verbal Learning and Verbal Behavior* 17, 143-154.

Zipf, G. K. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

Appendix

List of type frequencies across subcorpora, cleaned files

Figures: V(N)(types) /N (tokens)/ n_1 (hapaxes)

Affix	demographic			context-governed			written		
	V(N)	N	n_1	V(N)	N	n_1	V(N)	N	n_1
