The emergence of productive non-medical *-itis*: corpus evidence and qualitative analysis

Anke Lüdeling and Stefan Evert

1 Introduction

No natural language has a closed vocabulary (Kornai 2002). In addition to mechanisms to add to the base vocabulary, like borrowing, shortening, creativity etc. the productivity of morphological processes can form new complex entries. Some word formation processes can be used to form new words more easily than others. This fact, called morphological productivity, has been recognized for a long time and discussed from many points of view (see for example Aronoff 1976; Booij 1977; Baayen and Lieber 1991; Baayen 1992; Plag 1999; Bauer 2001; Baayen 2001; Nishimoto 2004).

This paper is concerned with evidence for different aspects of morphological productivity. Our claim is that the problem of productivity can only be understood when different kinds of evidence – quantitative and qualitative – are combined. We will try to understand more about the interaction of qualitative and quantitative aspects of morphological productivity. We illustrate our claim by looking at a morphological element that has not received much attention in morphological descriptions yet: German *-itis*.[1]

1.1 Qualitative productivity

In this section we want to introduce two different ways of looking at the qualitative aspects of productivity: categorial models and similarity-based models.

In generative models for linguistic competence, every rule[2] is either valid or not, i.e. a rule produces the ‘grammatical’ expressions – complex words in our case – of a language. A rule states whether a process is *available* (Bauer 2001) in a given language. In our model, rules can refer to every linguistic property of a lexical entry – a consequence of this are, of course, complex lexical entries where information on all linguistic levels can be associated with each word. For the sake of simplicity we assume a basic item-and-arrangement model for German word formation here.[3]

These rules refer to linguistic categories and are thus categorial. In a competence model every rule is 100% productive, i.e. every item that belongs to a category given in a rule can be inserted. The rules do not refer to linguistic experience or frequencies of complex words or the like. In a lexical component that goes with a competence model, only irregular complex items are stored.[4]

Next to rule- or constraint-based competence systems there are morphological models that are based on similarity: existing words are grouped according to some similarity criterion. The proportional formula introduced by Greek grammarians defines one instance of a similarity measure, analogy. The term analogy has been used in morphology in at least two different ways (for an overview see Becker 1990). For the Young Grammarians, analogy was a regularization process in the formation of new groups and elements, and thus in language change, be it in syntax or in morphology (compare Paul 1920: Chapter 5).[5] A different view is given in Pinker (1999), where analogy is a process that is used for exceptions only and is a process totally different from rules. In Pinker's model, regularly formed words are not stored. Analogical models in the sense of Paul, which we want to adopt here, in contrast to competence models, are based on linguistic experience: we assume that instances even of regularly formed types are stored, grouped, and these groups can serve as examples after which new elements can be moulded.

Rule-based and similarity-based models for morphology have in common that they are based only on the different *types* of complex words produced by a morphological process.

1.2 Quantitative productivity

As stated above, some rules form new words easier than others. In a generative competence model, this notion is not expressible. Nonetheless, many authors have tried to associate quantitative terms such as 'highly productive' or 'semi-productive' with generative word formation rules without specifying how they would fit into a generative model. See (Plag 1999: 12) for an overview. The intuition is that the different 'degrees' of productivity are due to the number of restrictions for a word formation process and also to the number of possible bases. Some authors have even given formulae for measuring productivity: (Aronoff 1976: 36), for example, states that "There is a simple way to take such restrictions into account: we count up the number of words which we feel could occur as the output of a given WFR [word formation rule, AL&SE] (which we can do by counting the number of possible bases for that rule), count up the number of actually occurring words formed by that rule, take a ratio of the two, and compare this with the same ratio for another WFR. In fact, by this method we could arrive at a simple index of productivity for every WFR: the ratio of possible to actual words." The formula he suggests is thus

$$I = \frac{\textit{number of attested words}}{\textit{number of possible words}} \tag{1}$$

Even if it were possible to count the number of possible bases and the number of actual bases, this formula would not yield the intended result: what we would get is a static number. But what does it mean to state that, say, 38% of all words that can possibly be formed by a rule have already been formed? This number would not tell us whether the rule will ever form a new word, i.e. it will not allow a statement on the productivity of a rule and certainly no predictions. This issue is also discussed by Baayen (1989) who states that Aronoff's measure I can be seen as expressing the *unproductivity* of the word formation process for an unproductive affix (Baayen 1989: 30).

(Booij 1977: 5) suggests a different way of computing productivity: "The degree of productivity of a WF-rule can be seen as inversely proportional to the amount of competence restrictions on that WF-rule." To realize this idea one would have to come up with a theory of how to count and rank restrictions (does the restriction "X combines only with verbs" have the same status as the restriction "X combines only with bisyllabic elements" etc.?). If we had such a theory the formula would again yield a static number.

To circumvent such problems, more sophisticated quantitative models have been proposed, which take both the number of *types* of complex words and the number of *tokens* of these words into account - counted on a given corpus (see below). The most influential ones are the statistical models developed by Baayen (see, among others, Baayen 1989; Baayen and Lieber 1991; Baayen 1992, 1993a,b, 1994, 2001, 2003). Here, the quantity of interest is the readiness with which a morphological rule will form a new complex word. It can be operationalized by the concept of vocabulary growth, i.e. how often new word types are encountered when an increasing amount of text is sampled. We will return to measures of vocabulary growth and productivity models in Section 3. Before that, we describe the properties of the element *-itis* and explain why this qualitative analysis needs to be complemented with quantitative information.

2 *-itis*

We will now briefly describe the properties of our test case -*itis*. We chose -*itis* because it is part of two very different word-formation processes: the rule-based, or categorial, medical -*itis* and the similarity-based non-medical -*itis*.

2.1 Medical *-itis*

The German morphological element -*itis* is originally used in medical contexts with the meaning 'inflammation (of)'. It is always bound and combines productively with neoclassical elements denoting body parts, e.g. *Arthritis* 'inflammation of the joints' or *Appendizitis* 'inflammation of the appendix'. Most of the elements it combines with occur only in bound form (often called a formative), it is therefore difficult to assign them a part of speech. However, from their semantics, it could be argued that they are nominal elements. A rule for medical -*itis* could look like

$$N \leftarrow \mathit{Formative}_{\text{neoclassical}}[[\text{body-part}]] + \text{-}\mathit{itis} \tag{2}$$

2.2 Non-medical *-itis*

-*itis* can be used in non-medical contexts in a different function. Well-known examples of this 'non-medical -*itis*' are *Telefonitits* 'excessive use of the telephone' or *Subventionitis* 'excessive subsidizing'. In contrast to medical -*itis* it is difficult to characterize non-medical -*itis* in categorial terms. It combines mostly with neoclassical elements but (in recent years, see below) more and more also with native elements cf. *Fresseritis* 'eating too much', names as in *Wehneritis* 'being too much like Wehner (a German politician in the 1960s and 1970s)' or English elements *Bestselleritis*. Categorially, the non-head can be a noun, as in *Zitatitis* 'citing too much', a verb as in *Aufschieberitis* 'procrastinating too much', or an adjective as in *Exklusivitis* 'wanting exclusive interviews, articles etc. too often (journalistic context)' or even phrases as in *Vielzuvielitis*, lit.: much-too-much-itis "wanting too much". -*itis* attracts and bears stress and wants to follow an unstressed syllable. Where the non-head ends in a stressed syllable, sometimes the allomorph -*eritis* is used, cf. *Filmeritis* 'watching too many movies'. Where the non-head ends in a vowel, a linking element is inserted, as in *Tangolitis* 'playing too many tangos'. Semantically, non-medical -*itis* is rather vague – its meaning can be described as 'doing too much of X' where 'X' is some activity related to the meaning of the non-head. This vague paraphrase shows already that the non-head is interpreted 'verbally' rather than 'nominally' independent of it's actual part of speech. Note that the meaning of non-medical -*itis* is, of course, not independent of the meaning of medical -*itis*: we suspect that medical -*itis* was generalized to mean 'illness' (instead of referring specifically to an inflammation). One indication for this is the fact that non-medical -*itis* collocates with words such as *akut* 'acute', *chronisch* 'chronic' or *leiden an* 'suffer from'.

It is not easy to write a categorial rule for non-medical -*itis* like the one above for medical -*itis*. We believe that non-medical -*itis* is a good case of a similarity-based process. One piece of evidence is that non-medical -*itis* words are to a certain extent stylistically marked which medical -*itis* words are not.

2.3 Goals of the quantitative analysis

The qualitative analysis of -*itis* shows that we have evidence for two morphological processes with different properties. Qualitative evidence does not suffice, however, to explain their productivity. We want to look at two aspects of productivity, (1) do both processes differ with respect to productivity and (2) (how) does the productivity of each process change over time?

It has been argued that categorial and similarity-based morphological processes exist next to each other. If so, can we see differences in their quantitative behaviour? As stated above, in a competence model every rule is fully productive. The rule we formulated says that all neoclassical formatives that denote body parts can be inserted. This cannot be directly compared to a similarity-based process where one can calculate type-token relationships. In the remainder of this paper we will therefore use the same model, based on type-token statistics, for both processes (see Section 3.2). This means that we will only look at the *output* – the complex words – of the two processes. If the two processes are really fundamentally different, we would expect to see quantitative differences in their output: the productivity for the rule-based process should be higher and more constant. The statistical analysis assumes a homogeneous model – we would therefore expect to get better goodness-of-fit values for the rule-based process than for the much more heterogeneous similarity-based process.

Another issue of interest is the short-term diachronic change of productivity. The hypothesis would be that the established medical rule-based use of -*itis* does not change over time but non-medical -*itis*, which is similarity-based and therefore dependent on the stored examples can show short-term qualitative changes as well as changes in productivity. Again, this cannot be expressed in a competence model. We will suggest different ways of looking at what could be called ’diachronic productivity’ below.

Our quantitative analysis of -*itis* is based on the full 980 million word corpus “Textbasis für das digitale Wörterbuch der deutschen Sprache” (henceforth *Textbasis*) collected by the Berlin Brandenburgische Akademie der Wissenschaften. This corpus is an opportunistic collection of newspaper data, literature, informative texts, scientific texts and spoken language from the 20th century.[6] The theoretical problems in using an opportunistic corpus of this sort are addressed below. In addition, there are a number of practical problems, which are described by Evert and Lüdeling (2001). The data have been manually cleaned up according to the guidelines given there.

It is important to keep in mind that quantitative measures of productivity are closely tied to the corpus on which they are based. The precise question to which they provide an answer can be paraphrased in the following way: how likely is it that previously unseen word types (formed by the process being studied) will appear when additional (similar) text is sampled? Our interest in the phenomenon of productivity, however, is at its core a cognitive one – we want to understand how a speaker of a language knows that she can use a morphological rule to form a new word or phrase. Quantitative productivity is an observable reflex of this knowledge, namely the readiness of speakers to form new words, but it is also influenced by many other factors. In particular, our results apply only to the particular situation that is represented by *Textbasis* (mostly journalistic writing).

However, it is also possible to give the corpus data a cognitive interpretation: We assume a model of word formation that incorporates qualitative and quantitative knowledge about word formation processes. This model is based on the idea that knowledge about the productivity of a morphological process depends on a speaker’s *linguistic experience*. This implies that both qualitative and quantitative aspects of productivity change with the change of experience. Corpus data – in particular the number of different words already formed by a given process and the apparent readiness of forming new words – can then be seen as a model for the speaker’s linguistic experience. Such an assumption is problematic in many respects, of course: no existing corpus comes close to representing the experience of a native speaker, let alone an opportunistic collection such as *Textbasis* or the recently very popular “World Wide Web as a corpus”. In this paper we therefore only measure and compare the productivity of the two processes involving -*itis* within *Textbasis* without claiming to provide a corpus-based model for linguistic experience.

3 Measuring morphological productivity

3.1 Vocabulary growth

The statistical models of Baayen (1989, 1992, 2001, 2003) link the degree of productivity of a morphological process to the rate of vocabulary growth, i.e. to how frequently new word types that are formed by the process are encountered when an increasing amount of text is sampled. If the degree of productivity changes over time, there should be a corresponding change in the vocabulary growth rate.

For a corpus with a publication date for each document (as in the case of *Textbasis*[7]), a natural approach is to scan the corpus in chronological order. The vocabulary size V of a given word-formation process at a given time t, given as $V(t)$, is the number of different word types (formed by the process) found in the part of the corpus up to the time t. Figure 1 shows vocabulary growth curves, graphs of $V(t)$ against t, for medical (left panel) and non-medical (right panel) -*itis* nouns in *Textbasis*. The slope of these vocabulary growth curves represents the rate at which new types appear in the corpus.

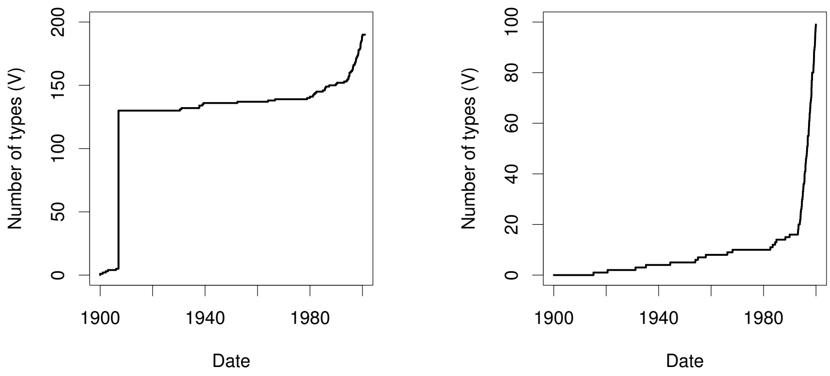

Figure 1: Vocabulary growth of -*itis* throughout the 20th century (left: medical -*itis*, right: non-medical -*itis*)

Taken at face value, the steep rise of both vocabulary growth curves towards the end of the century seems to indicate that both medical and non-medical -*itis* have become much more productive in the 1990's. There is also a startling jump in the left graph, where more than 100 new medical -*itis* words suddenly appear in the data. A closer inspection reveals that *Textbasis* comprises a substantial part of the 1906 edition of the German Brockhaus Encyclopedia, including definitions of a large number of medical terms. At first, one may be inclined to dismiss this as a quirk in the composition of the corpus and discard the dictionary data. The situation reveals a fundamental problem of the vocabulary growth approach to productivity, though. Obviously, all -*itis* words listed in the encyclopedia must have been in use at the time of publication and deserve to be included in $V(t)$, at least when the latter is given a strict interpretation as the number of different words formed up to the time t. In a corpus of 'ordinary' text, on the other hand, many of these medical terms would have been encountered at a much later time, or perhaps not at all, giving a smooth growth curve similar to that of non-medical -*itis*. This shows that one cannot know whether a 'new' word was actually formed at time t or whether it had already been established in the language and just happened not to occur in the corpus data from the preceding time period (in this case, it would have occurred if more or different text had been sampled from this time period).[8]

As a consequence of the stochastic nature of growth curves, the number of new types encountered in a given time period depends crucially on the amount of text sampled. Figure 2 shows the number of instances of -*itis* nouns in each five-year period (left panel: medical use; right panel: non-medical use). Almost all tokens occur in the last decade of the century (with the exception of the medical -*itis* nouns in the Brockhaus Encyclopedia). The large number of new types found during this period may simply be a correlate of the large number of tokens and need not imply a change in the degree of productivity.[9]

Vocabulary growth curves as shown in Figure 1 mix up two different effects: (i) how new types are encountered when more text is sampled (synchronic vocabulary growth, cf. Section 3.2), and (ii) how easily new types are formed by speakers of the language (changes in the degree of productivity, which may lead to diachronic vocabulary growth when complex words are formed that were previously impossible or at least highly unusual). In order to obtain meaningful results from a statistical analysis, it is necessary to separate these two effects. We suggest to employ the following procedure: First, determine the *synchronic* productivity of the process at a given point in time (Section 3.2), using a statistical model that takes the stochastic nature of (synchronic) vocabulary growth into account. The resulting measure of productivity must be independent of the amount of text sampled. Second, study the *diachronic* aspect of productivity by comparing the degree of synchronic productivity at two (or more) different points in time (Section 3.3). In or-

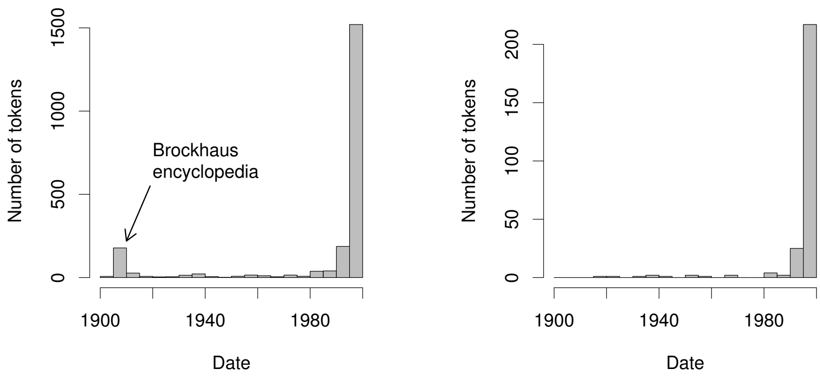

Figure 2: Number of instances found in *Textbasis* for five-year periods in the 20th century (left: medical -*itis*, right: non-medical -*itis*)

der to make this comparison possible, the source corpus must satisfy certain criteria, which are also summarized there.

3.2 Synchronic productivity

Synchronic productivity captures the behaviour of a single speaker or a community of speakers at a fixed point in time. The standard models interpret the observed corpus data as a random sample from the potential output of the speaker(s). More precisely, the relevant -*itis* tokens (either medical or non-medical) in the observed data are treated as a random subset of the -*itis* token in the speakers' output; all other tokens are discarded. In order to obtain a fully synchronic measure, the time span covered by the corpus should be as short as possible. However, a sufficient amount of data (both a sufficient number of tokens and a sufficient number of different types) is necessary for the statistical analysis. Otherwise, the inherent uncertainty of statistical estimates (such as the ones introduced in this section) would become too high to allow a meaningful interpretation. The following examples are based on *Textbasis* data from the years 1990–1999, although a shorter time span would be desirable (cf. Section 4).

Vocabulary growth curves, albeit of a different kind, provide an intuitive visual approach to synchronic productivity (Baayen 2003: 236–242). Here, vocabulary growth is measured in text time, i.e. with respect to the number of -*itis* tokens encountered as an increasing amount of text is sampled. Figure 3 displays synchronic vocabulary growth curves for -*itis* nouns (left panel: medical use, right panel: non-medical use). Note that both graphs are drawn to the same relative scale, with 10 units on the x-axis corresponding to 3 units on the y-axis. However, the sample size N is vastly different for the two processes ($N = 1707$ for medical vs. $N = 242$ for non-medical -*itis*). For direct comparison, the growth curve of non-medical -*itis* is shown as a thin dotted line in the left panel, and that of medical -*itis* is shown as a thin dotted line in the right panel.

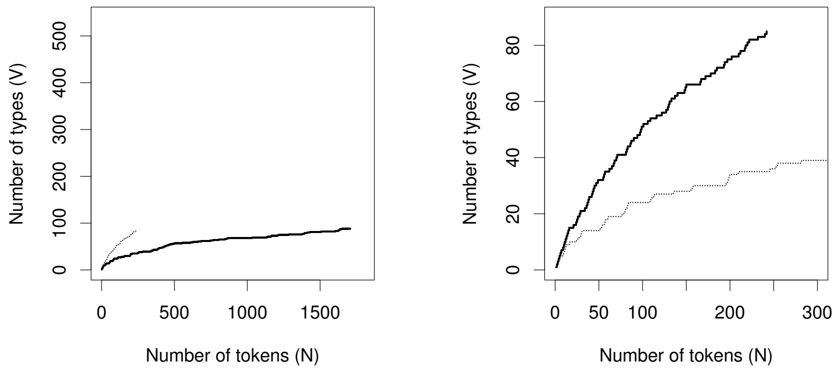

Figure 3: Synchronic vocabulary growth curves for -*itis* in the 1990's, showing the number of different types among the first N instances of -*itis* words in the corpus (left: medical -*itis*, right: non-medical -*itis*)

The slope of a vocabulary growth curve, which can be interpreted as the probability that the next -*itis* token will be a previously unseen one, provides a natural measure of productivity. It is sometimes referred to as the category-conditioned degree of productivity $\mathcal{P}$ (Baayen 2003: 240). Obviously, the jagged growth curves would need to be smoothed in some way before their slope can be computed. These irregularities are a stochastic effect of sampling, depending on the particular order in which the tokens are arranged in the sample. Under the random sample model, the precise arrangement is irrelevant: all re-orderings of the sample are equally likely. An 'average' value for the growth rate $\mathcal{P}$ is thus obtained by averaging over all possible re-orderings. It can easily be estimated from the sample size N and the number V_1 of *hapax legomena* (word types that occur just once in the sample): $\mathcal{P} \approx V_1/N$ (Baayen 2001: 50).

From the growth curves in Figure 3, we obtain $\mathcal{P} \approx .0217$ for medical

-*itis* and $\mathcal{P} \approx .248$ for non-medical -*itis*. On this scale, the productivity of non-medical -*itis* seems to exceed that of medical -*itis* by more than a factor of eleven. Such a 'naive' interpretation of $\mathcal{P}$ is problematic, though, mostly because the growth rate depends critically on the sample size. When samples of identical size $N = 200$ are compared for the two processes (cf. the right panel of Figure 3), the difference in the degree of productivity is less striking: $\mathcal{P} \approx .075$ vs. $\mathcal{P} \approx .265$, a factor of 3.5 only. This example shows that despite its intuitive interpretation, it is difficult – if not impossible – to use $\mathcal{P}$ as a measure for the degree of productivity of a word formation process. $\mathcal{P}$ is much more an extrapolation of the observed sample than an absolute (i.e. size-independent) measure of productivity.

The measure $\mathcal{P}$ focuses entirely on the number of hapax legomena in the sample. Intuitively, this approach makes sense: after all, the hallmark of a productive process are nonce formations, created as they are needed in a specific situation to express a certain concept. Such a need may arise again on a similar occasion, though, so that the same word will once more be productively formed by the same or a different speaker. When a sufficient amount of text is sampled, many types will be seen more than once even for a highly productive process. It is therefore necessary to look at all low-frequency types, not just the hapax legomena. Figure 4 shows the number V_m of -*itis* types that occur exactly m times in the sample, for $m = 1 \dots 10$ (left panel: medical use, right panel: non-medical use). Such a bar graph (or a corresponding table of m and V_m) is referred to as the frequency spectrum (Baayen 2001: 8) of a word formation process with respect to the observed corpus.

Although productively formed types may occur more than once, they will in general be less frequent than well-established words. This reasoning implies that a productive process should be characterized by a frequency spectrum that is skewed towards the lower end. The stronger the skew, the more productive the process is. The frequency spectra in Figure 4 confirm the impression given by the growth curves, with the spectrum of non-medical -*itis* being dominated by hapax and dis legomena (types occurring twice).

Baayen (2001: Chapter 3) describes statistical models that abstract away from the stochastic irregularities of a sample-based frequency spectrum and estimate how much the full output of the speaker(s) is skewed towards low-frequency words (cf. the remarks at the beginning of this section). He refers to them as LNRE models, where LNRE stands for "large number of rare events" (after Khmaladze 1987). It is not obvious which one of several possible LNRE models should be used. These models differ in their flexibility and accuracy, but also in their computational complexity. None of them has

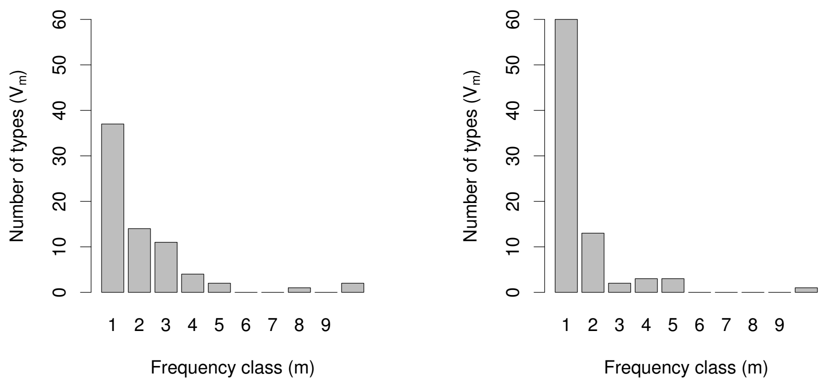

Figure 4: Frequency spectrum for -*itis* nouns in the 1990's, showing the number V_m of types that occur exactly m times in the sample (left: medical -*itis*, right: non-medical -*itis*)

a theoretical foundation rooted in the theory of morphological productivity. Therefore, a multivariate goodness-of-fit test is applied to find out how well the predictions of the model agree with the observed spectrum (Baayen 2001: 118–122). It is only appropriate to draw further inferences from an LNRE model when it has been confirmed by the goodness-of-fit test as a plausible explanation for the observed data.

For the experiments reported in this paper, we used the *finite Zipf-Mandelbrot* (fZM) LNRE model introduced by Evert (2004), which is based on the Zipf-Mandelbrot law (Zipf 1949; Mandelbrot 1962). The fZM model is both computationally efficient and flexible, and it is reported to achieve better goodness-of-fit than many other LNRE models (Evert 2004: 420–421). Figure 5 compares the observed frequency spectra of medical and non-medical -*itis* with the predictions of the fZM models. The multivariate goodness-of-fit test shows an acceptable fit for medical -*itis* ($\chi^2 = 22.59$, df $= 13$, $p = .047$) and an excellent fit for non-medical -*itis* ($\chi^2 = 13.91$, df $= 13$, $p = .380$).

The overall shape of the frequency spectrum predicted by the fZM model is mainly determined by the model parameter α. Its values range from $\alpha = 0$ (indicating a balanced spectrum, where the number of hapax legomena is not much larger than the number of types in higher frequency ranks) to $\alpha = 1$ (indicating a highly skewed spectrum that is entirely dominated by the hapax legomena).[10] We can thus tentatively use α as a quantitative measure for the degree of productivity. When α is close to 0, the morphological process in

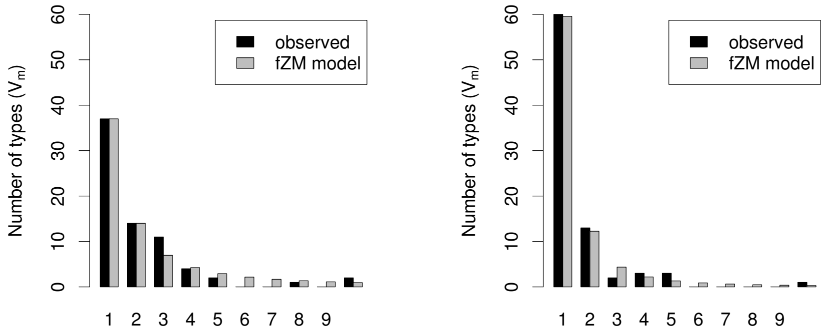

Figure 5: Frequency spectrum for *-itis* nouns in the 1990's together with the predictions of the finite Zipf-Mandelbrot LNRE model (left: medical *-itis*, right: non-medical *-itis*)

question may not be productive at all, when $\alpha \approx 0.5$, it is moderately productive, and when α is close to 1, it has a very high degree of productivity. For medical *-itis*, the shape parameter is $\alpha \approx 0.565$; for non-medical *-itis*, it is $\alpha \approx 1$, indicating that the latter is indeed much more productive. The finite Zipf-Mandelbrot model also provides an estimate for the total number of complex *-itis* types that can be formed by the two processes, which is $S \approx 183$ for medical use and $S \approx 435$ for non-medical use (see Evert 2004: 417–418).

Such estimates must not be taken all too literal, though, because the fZM model and similar LNRE models gloss over many of the complexities of word frequency distributions, concentrating on the more 'regular' lower end of the frequency spectrum (cf. Baayen 2001: chapter 4).[11] Moreover, the relatively small size of our samples implies that many different classes of LNRE models (beside the fZM model used in our experiments) will be consistent with the observed data (as measured by their goodness-of-fit), some of which may predict a much larger or even infinite value for S. One way of testing the plausibility of our estimates is to compare the value $S \approx 183$ with the number of established *-itis* terms in medical jargon. Manual counts on randomly selected pages from a German medical dictionary (Ahlheim and Lichtenstern 1968) indicate a minimum of 220 such terms (and probably even more than 300 terms).[12] One possible explanation for the substantial underestimation of S by the fZM model is the composition of the *Textbasis*, which contains little technical writing from the domain of medicine. Therefore, statistical models applied to these data estimate the number of *-itis* nouns that are used in general language rather than the possibly much greater number available to a medical expert.

Despite these reservations, the estimated values of S are useful as an intuitive and readily interpretable way of comparing the productivity of different processes in our experiments. The comparison is valid because both processes are analyzed with the same class of statistical models (namely, the fZM model), so that differences in the estimated parameters reflect actual differences between the frequency distributions of the two processes (rather than resulting from the assumptions underlying different statistical models).

3.3 Diachronic productivity

Our approach to diachronic productivity, changes in the readiness with which a morphological process forms new words, is based on the measures of synchronic productivity developed in section 3.2. We compute the degree of synchronic productivity for a given process at two points in time, t_1 and t_2. By comparing e.g. the shape parameters $\alpha(t_1)$ and $\alpha(t_2)$ (or the estimated total number of types, $S(t_1)$ and $S(t_2)$) we can detect an increase or decrease in productivity. For a precise description of diachronic trends, it would be necessary to consider further points in time, $t_1, \ldots, t_n$, and formulate a mathematical model $t \mapsto \alpha(t)$ for the development of the shape parameter. This model could take the form of a logistic function, for instance, which is often used in research on language change (see e.g. Zuraw 2003: 148–149).

In order to make this comparison, we need text samples from the time points t_1 and t_2, or short time spans containing those points. The statistical models ensure that we need not worry about different sample sizes. However, some requirements remain, which unfortunately are not met by the *Textbasis* corpus. First, we have already pointed out at the beginning of Section 3.2 that a certain minimal amount of data is needed in order to carry out a meaningful statistical analysis, both with respect to the number of tokens and the number of types. This means that even a corpus containing millions of words may not be large enough when words formed by the process of interest are rare in the language. Moreover, a process with a low degree of productivity might require even larger samples in order to have a sufficient number of different types. In *Textbasis*, a sufficient number of *-itis* tokens are only found for the years from 1993, where several hundred million words of newspaper text are included in the corpus. During the earlier decades, there are only isolated instances of *-itis* words, both medical and non-medical – far too little data for the application of an LNRE model (e.g., there are only $N = 16$ instances of non-medical -*itis* in *Textbasis* before the year 1990).

Second, the text samples from t_1 and t_2 must have similar composition (with respect to modality, text type, domain, etc.) in order to allow a direct comparison of the productivity measures. For instance, it is quite plausible to assume that non-medical -*itis* is more productive in fashionable journalistic writing than in literary or scientific texts. Even if we had a sufficient amount of data in *Textbasis* for the earlier decades, the prevalence of newspaper text in the 1990's might be responsible for a significantly higher degree of productivity. Finally, the individual text samples must be taken from a short time span in order to measure short-term developments. While it is not clear yet whether a morphological process can become productive (or unproductive) within a few years, such rapid changes are commonplace at the level of individual types. Figure 6 illustrates this claim with the example of non-medical -*itis*. The bar graph shows the relative frequencies of the four most frequent word types in the years 1993–1999. While *Fusionitis* "too many mergers" rapidly becomes popular towards the end of the century, *Subventionitis* "too much subsidizing" has its heyday in the years 1994–1995, and seems to fall out of use afterwards.

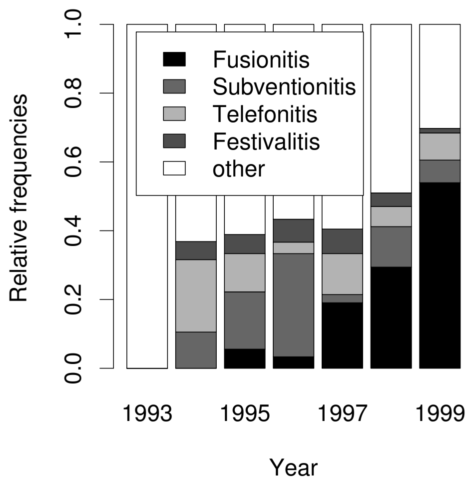

Figure 6: The relative frequencies of the four most frequent non-medical -*itis* words in the years 1993–1999.

With all its limitations, the *Textbasis* corpus still has great value for the *qualitative* description of productivity, showing that non-medical -*itis* has existed before the 1990's. A new type is encountered every few years, starting with the first occurrence of *Spionitis* "excessive fear of spies" in 1915.

4 Conclusion

To sum up, we have discussed the productivity of two morphological processes with different qualitative properties, categorial or rule-based medical -*itis* and similarity-based non-medical -*itis*. Since qualitative evidence alone is not sufficient to explain productivity, we have also used quantitative evidence from a German text corpus.

We have argued that a theoretical distinction between rule-based and similarity-based processes should be reflected in their quantitative behaviour: rule-based processes should be more productive, lead to frequency distributions that can accurately be described by statistical LNRE models, and their degree of productivity should not change over time. We have then shown that the quantitative properties of the two processes in question do not confirm our hypotheses. Although this surprising result may well be due to the nature of our data, one might also come to the (at this point very tentative) conclusion that morphological theory does not need to make a distinction between rule-based and similarity-based processes.

Acknowledgements

We would like to thank Alexander Geyken and Gerald Neumann who provided the *Textbasis* data on which this study is based. We are grateful to the audience at the Tübingen conference on Linguistic Evidence, and especially to Peter Bosch, Harald Baayen and an anonymous reviewer, for much helpful advice.

Notes

1 Note that we focus on German -*itis* which differs in some respects from English -*itis*. For a discussion of the morphological status of -*itis* see Lüdeling et al. (2002).

2 We will speak of rules and use a simple rule-based model for the sake of simplicity here but our arguments carry over to constraint-based systems.

3 In an IA model there is no need to distinguish between derivation and compounding. (We restrict ourselves to concatenative processes here.) This means that not only every stem is associated with its word formation stem forms (as assumed in Eisenberg 1998; Fuhrhop 1998) but also every bound entry (see Lüdeling and Fitschen 2002; Fitschen 2004: for a discussion).

4 The idea that only irregular words are stored in a lexicon while all regular words can be derived via rules is, of course, older than generative linguistics (see for example Bloomfield 1933). In psycholinguistics the question of what needs to be stored has been discussed for a long time, resulting in models like that of Pinker (1999). We cannot go into the psycholinguistic debate on the storage of complex items. We only want to say here that there has been recent evidence that even regularly inflected words seem to be stored in the mental lexicon. Baayen et al. (1997).

5 Paul assumes that words are combined into groups according to phonological or semantic similarity: "[...] attrahieren sich die einzelnen Wörter in der Seele, und es entstehen dadurch eine Menge größerer oder kleinerer Gruppen. Die gegenseitige Attraktion beruht immer auf einer partiellen Übereinstimmung des Lautes oder der Bedeutung oder des Lautes und der Bedeutung zugleich." (Paul 1920: 106)

6 For more information about the Textbasis, see `http://www.dwds.de/pages/pages_textba/dwds_textba.htm`.

7 We ignore the fact that the publication date is not necessarily the date of production.

8 Still, the linguistic experience of a particular speaker may in fact show a development just as it happens to be documented in *Textbasis*.

9 One might speculate whether the larger number of tokens observed in the last decade of the century is connected to intensity of use, which is a different aspect of morphological productivity. A more likely explanation is found in the opportunistic nature of *Textbasis*. Since the early 1990's, entire volumes of newspapers have become conveniently available in machine-readable form. *Textbasis* includes a large amount of such newspaper text, which skews the data in two ways: (i) there is much more text from the 1990's than from earlier decades, and (ii) this text is dominated by journalistic writing. All instances of non-medical *-itis* in the 1990's are from newspaper sources, with the single exception of *Aufschieberitis* (from Kellner 1998).

10 It has to be noted at this point that the finite Zipf-Mandelbrot model, like most other LNRE models, is only suitable for productive processes with a skewed

frequency spectrum. It will not achieve a satisfactory goodness-of-fit for a completely unproductive process.

11 As an example, Grigorij Martynenko estimated from the Brown corpus (Kučera and Francis 1967) that the total vocabulary of American English comprises only $S = 112,500$ words (Martynenko 2000: Table 3).

12 A random selection of 77 out of 1277 half-page columns from Ahlheim and Lichtenstern (1968) were inspected manually for *-itis* headwords, which were found in 17 columns. This gives a maxium-likelihood estimate of 362 *-itis* terms in the dictionary, with a two-sided 95% confidence interval ranging from 226 to 525 terms (hypergeometric test). Note that a further subclassification of *-itis* terms is often expressed by combination with Latinate words or phrases, e.g. *Dermatitis ab acribus* 'dermatitis caused by chemical irritants'. Since this highly productive process is different from the affixation of *-itis*, the subclassified terms were not included in the counts.

References

Ahlheim, Karl-Heinz and Hermann Lichtenstern, (eds.)
1968 *DUDEN Wörterbuch medizinischer Fachausdrücke*. Bibliographisches Institut and Georg Thieme Verlag, Mannheim, Stuttgart.

Aronoff, Mark
1976 *Word Formation in Generative Grammar*. The MIT Press, Cambridge, MA.

Baayen, R. Harald
1989 *A Corpus-Based Approach to Morphological Productivity*. Ph.D. thesis, Vrije Universiteit de Amsterdam.
1992 Quantitative aspects of morphological productivity. In Geert Booij and Jaap van Marle, (eds.), *Yearbook of Morphology 1991*, pp. 109 – 150. Foris, Dordrecht.
1993a On frequency, transparency and productivity. *Yearbook of Morphology 1992*, pp. 181–208.
1993b Statistical models for word frequency distributions. *Computers and the Humanities*, 26: 347 – 363.
1994 Productivity in language production. In Dominiek Sandra and Marcus Taft, (eds.), *Morphological Structure, Lexical Representation and Lexical Access*, Special Issue of Language and Cognitive Processes, pp. 447 – 496.
2001 *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

2003 Probabilistic approaches to morphology. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, (eds.), *Probabilistic Linguistics*, chapter 7, pp. 229–287. MIT Press, Cambridge.

Baayen, R. Harald, T. Dijkstra, and Robert Schreuder
1997 Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 36: 94–117.

Baayen, R. Harald and Rochelle Lieber
1991 Productivity and English derivation: a corpus-based study. *Linguistics*, 29: 801 – 843.

Bauer, Laurie
2001 *Morphological Productivity*. Cambridge University Press, Cambridge.

Becker, Thomas
1990 *Analogie und morphologische Theorie*. Fink, München.

Bloomfield, Leonard
1933 *Language*. Holt, Rinehart and Winston, New York.

Booij, Geert
1977 *Dutch Morphology. A Study of Word Formation in Generative Grammar*. de Ridder, Lisse.

Eisenberg, Peter
1998 *Grundriß der deutschen Grammatik. Band 1: Das Wort*. J.B. Metzler, Stuttgart.

Evert, Stefan
2004 A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pp. 411–422. Louvain-la-Neuve, Belgium.

Evert, Stefan and Anke Lüdeling
2001 Measuring morphological productivity: Is automatic preprocessing sufficient? In *Corpus Linguistics 2001*. Lancaster.

Fitschen, Arne
2004 *Lexikon als komplexes System*. Ph.D. thesis, Universität Stuttgart.

Fuhrhop, Nanna
1998 *Grenzfälle morphologischer Einheiten*. Stauffenburg-Verlag, Tübingen.

Kellner, Hedwig
1998 *Das geheime Wissen der Personalchefs*. Eichborn, Frankfurt a. M.

Khmaladze, E. V.
1987 The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Kornai, András
2002 How many words are there? *Glottometrics*, 4: 61–86.

Kučera, H. and W. N. Francis, (eds.)
1967 *Computational Analysis of Present-Day American English*. Brown University Press, Providence.

Lüdeling, Anke and Arne Fitschen
2002 An integrated lexicon for the analysis of complex words. In *Proceedings of EURALEX 2002*, pp. 145 – 152. Copenhagen.

Lüdeling, Anke, Tanja Schmid, and Sawwas Kiokpasoglou
2002 Neoclassical word formation in German. *Yearbook of Morphology 2001*, pp. 253–283.

Mandelbrot, Benoit
1962 On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, (ed.), *Structure of Language and its Mathematical Aspects*, pp. 190–219. American Mathematical Society, Providence, RI.

Martynenko, Grigorij
2000 Statistical consistency of keywords dictionary parameters. Technical report, Department of Computational Linguistics, St. Petersburg State University. Available from `citeseer.ist.psu.edu/407553.html`.

Nishimoto, Eiji
2004 *A Corpus-Based Delimitation of New Words: Cross-Segment Comparison and Morphological Productivity*. Ph.D. thesis, Graduate Faculty in Linguistics, The City University of New York.

Paul, Hermann
1920 *Deutsche Grammatik. Band V. Teil IV: Wortbildungslehre*. Verlag von Max Niemeyer, Halle a.S.

Pinker, Steven
1999 *Words and Rules. The Ingredients of Language*. Basic Books (Perseus Book Group), New York.

Plag, Ingo
1999 *Morphological Productivity. Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

Zipf, George Kingsley
1949 *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

Zuraw, Kie
2003 Probability in language change. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, (eds.), *Probabilistic Linguistics*, chapter 5, pp. 139–176. MIT Press, Cambridge.