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

Anke Lüdeling*
Humboldt-Universität zu Berlin, Germany
anke.luedeling@rz.hu-berlin.de

Stefan Evert
Universität Stuttgart, Germany
evert@ims.uni-stuttgart.de

1 Evidence for morphological productivity

This paper is concerned with types of 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 illustrate our claim by looking at a morphological element that has not received much attention in morphological descriptions yet: non-medical *-itis*. Before we come to different aspects of morphological productivity, we briefly describe the properties of non-medical *-itis*.

The German morphological element *-itis*[1] 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'. *-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'. Non-medical *-itis* is also always used bound. It combines mostly with neoclassical elements but (in recent years, see below) more and more also with native elements and names, cf *Fresseritis* "eating too much" or *Wehneritis* "being too much like Wehner (a German politician in the 1960s and 1970s)".[2]

*We would like to thank Alexander Geyken and Gerald Neumann who provided the *Textbasis* data on which this study is based.

[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 L ̈udeling et al. (2002).

[2]All examples are taken from the full 980 million word corpus (hence *Textbasis*) collected by the Berlin Brandenburgische Akademie der Wissenschaften, http://www.dwds.de/pages/pages_textba/dwds_textba.htm. The corpus is an opportunistic collection of newspaper data, literature, informative texts, scientific texts and spoken language from the 20th century. The problems involved in using an opportunistic corpus of this sort are discussed below.

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)'. *-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 CONCEPT' where 'CONCEPT' is some activity related to the meaning of the non-head.

2 Categorial rules and productivity

Morphological rules in generative theory of any flavour are typically described in categorial terms. A rule for medical *-itis* could look like

$$N \leftarrow \textit{Formative}_{\text{neoclassical}}[[\text{body-part}]] + \textit{-itis} \tag{1}$$

(translate this to your favourite formalism). Such rules contain categories that are intensionally definable. Therefore, in a production system based on such rules, all rules are 100% productive or, in other words, each element that fits the intensional description of such a category can be inserted here. Bauer (2001) calls this the *availability* of a rule. Inside such rule systems, which describe the linguistic competence of a speaker, it is formally impossible to express the notion that such rules have different degrees of productivity. Evidence for the adequateness of such a rule can only come from the linguistic intuition of a speaker.

The notion that some morphological rules form new words more easily and more frequently than others is commented on by most descriptive and formal morphologists (Aronoff (1976), Plag (1999), Bauer (2001)). Many attempts to measure or express such 'degrees of productivity' have been made. Corpus data can provide evidence for a quantitative measure of productivity, but it is important to note that any computations have to be based on a fine-grained linguistic analysis of manually cleaned-up data (Lüdeling and Evert, 2003).

Intuitively, the degree of productivity of a morphological process depends on how frequently a new word is formed by the process (Baayen, 1992, 2001). Therefore, a natural quantitative approach is to count the number of different word types (formed by the process of interest) found in the observed data up to a given time t, the vocabulary size $V(t)$. The top left panel of Figure 1 plots $V(t)$ against t (called a vocabulary growth curve) for non-medical *-itis* nouns in the *Textbasis*. The slope of this graph represents the rate at which new *-itis* types appear in the corpus.

Especially when a large time span is covered by the data, it is tempting to interpret the shape of the vocabulary growth curve as a measure of how productivity changes over

time. In our example, a steep rise shows that many new *-itis* types were introduced in the 1990's, indicating that the process suddenly became much more productive. This conclusion is not justified, though, because other factors have an influence on the shape of the vocabulary growth curve as well. As the top right panel of Figure 1 shows, almost all instances of non-medical *-itis* in the *Textbasis* are from the last decade of the century. Therefore, the large number of new *-itis* types in this period may simple be a consequence of the large number of *-itis* tokens. It is not necessarily the case that it has become easier to coin new *-itis* words.

The bottom left panel of Figure 1 plots the vocabulary size V against the number of observed tokens N, independent of the elapsed time. The slope of this graph, which represents the likelihood that the next *-itis* token is a previously unseen type, decreases over time. Taken at face value, this would suggest that the process was more productive at the beginning of the century than in the 1990's. Such contradictory results show that a more sophisticated analysis is needed, which makes a clear distinction between *synchronic* productivity (at a given time t) and *diachronic* productivity (changes in synchronic productivity between times t_1 and t_2).

3 Synchronic and diachronic productivity

Baayen (2001) describes statistical models of vocabulary growth, which interpret the set of N tokens as a homogeneous random sample from a population of S types[3] with (unknown) occurrence probabilities. In this approach, productivity is inherently a property of the population: the more skewed the distribution of probability parameters is, with a large number of low-probability types, the higher the degree of productivity.[4] The observed data are used to make inferences about the population probabilities, and hence about the degree of productivity of a process.

Synchronic productivity captures the behaviour of a single speaker or a community of speakers at a fixed point in time. Since it is reasonable to assume that the occurrence probabilities are constant, we can apply the random sample model described above. Note that vocabulary growth curves have a different interpretation in this model. In contrast to the top left panel of Figure 1, which depicts the appearance of new types over time, *synchronic growth curves* as shown in the bottom left panel measure the growth of the observed vocabulary as more instances of the targeted process are encountered in text from the same time and group of speakers. Ideally, a given text sample (such as newspaper volume) should be processed in random rather than chronological order.

As an example, we determine synchronic productivity for the data on non-medical *-itis* from the 1990's ($N = 254$ tokens with $V = 83$ different types), making the implicit

[3] S is called the population size and may be infinite.

[4] Baayen (2001) uses the term LNRE distributions, for Large Number of Rare Events.

assumption that the productivity of the process – even the occurrence probability of each type – is constant during this period. The data are analysed with the help of a population model, i.e. a model for the distribution of occurrence probabilities in the population (Baayen, 2001, Ch. 3). The goodness-of-fit of such a model is determined by comparing the vocabulary growth curve or frequency spectrum[5] with the predictions of the model, as shown in the bottom right panel of Figure 1. When the model is consistent with the observed data, its parameters (or other values computed from the parameters) can be interpreted as a quantitative measure of synchronic productivity.

The statistical model for synchronic vocabulary growth cannot be applied to the diachronic case because it assumes a homogeneous random sample.[6] Therefore, a statistical measure of *diachronic productivity* has to determine and compare the synchronic productivity of a morphological process at two points in time, t_1 and t_2. Unfortunately, the data on non-medical *-itis* from earlier decades (with $N = 16$ tokens and $V = 15$ types) is insufficient for a statistical analysis and comparison with the 1990's. However, from the diachronic vocabulary growth curve in the top left panel of Figure 1 we can at least conclude that non-medical *-itis* has existed before the 1990's. A new type is encountered every few years, starting from the first occurrence in 1915.[7]

To summarize: in order to provide a unified account of morphological productivity, we need (at least) three different kinds of linguistic evidence (a) the intuition of a native speaker to formulate qualitative rules, (b) the distribution (type-token ratios) of the types produced by that rule to model synchronic productivity, and (c) distributions of the types produced by the rule at different points in time that can be compared to model diachronic productivity.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar*. The MIT Press, Cambridge, MA.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij and J. van Marle, eds., *Yearbook of Morphology 1991*, pp. 109 – 150. Foris, Dordrecht.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

[5] The frequency spectrum lists, for each frequency rank m, the number of different types V_m that occur exactly m times in the sample.

[6] It is mathematically possible to treat the parameters of an LNRE model as functions of time (cf. Baayen, 2001, Sec. 5.2.2) in order to account for changes in the population probabilities. However, this approach is unsatisfactory: the model parameters would describe the cumulative frequency distribution (i.e. a mixture from different periods) rather than the productivity at the current time.

[7] In order to provide a rigorous statistical analysis of diachronic productivity, a sufficient amount of corpus data has to be available from two or more time points; the samples should be comparable in size and composition.

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press, Cambridge.

Lüdeling, A. and S. Evert (2003). Linguistic experience and productivity: corpus evidence for fine-grained distinctions. In D. Archer, P. Rayson, A. Wilson, and T. McEnery, eds., *Proceedings of the Corpus Linguistics 2003 conference*, pp. 475–483. UCREL.

Lüdeling, A., T. Schmid, and S. Kiokpasoglou (2002). Neoclassical word formation in German. *Yearbook of Morphology 2001*.

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

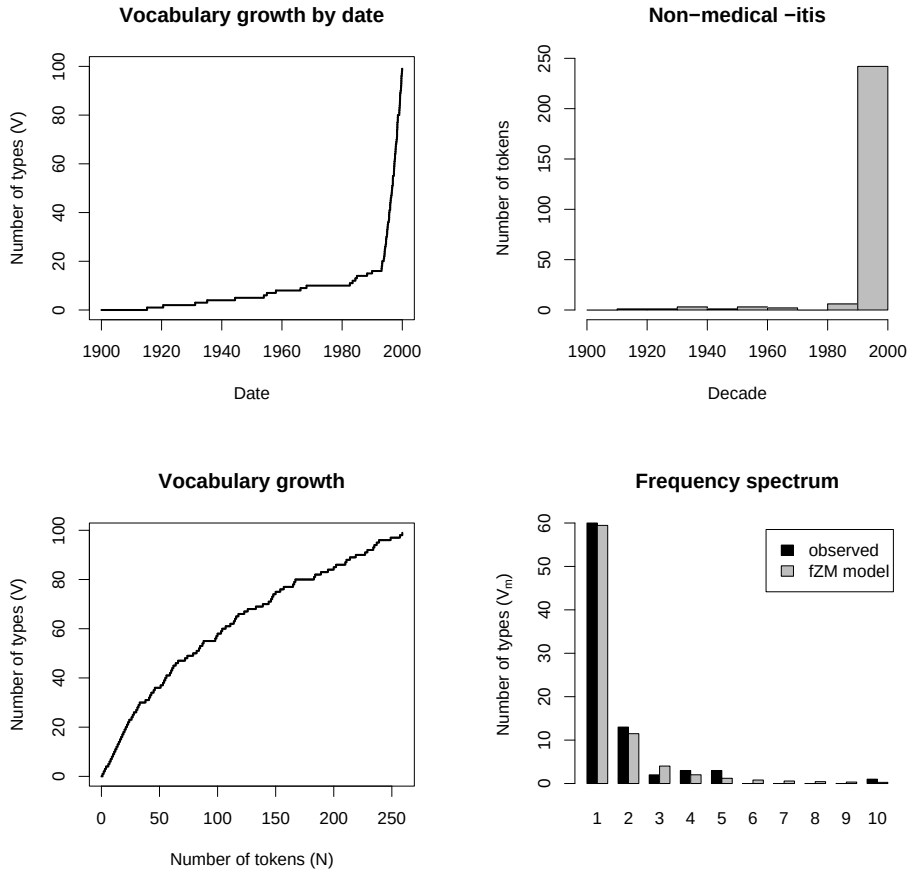

Figure 1: Top row: Non-medical *-itis* in the *Textbasis* (left: vocabulary growth; right: distribution of *-itis* tokens across decades). Bottom left: Vocabulary growth in token units. Bottom right: Frequency spectrum of non-medical *-itis* in the 1990's and prediction of a population model based on the Zipf-Mandelbrot law.