Need and Competition: Deconstructing Quantitative Productivity

1 Need and Competition in word formation

Morphological theory is concerned not only with the structure of existing complex words but also with the causes and mechanisms that determine the 'availability' (Bauer, 2001) of morphological processes to form new complex words. While in some approaches each rule is either fully available or not all (qualitative productivity; Dressler (2003)), many researchers in morphology focus on the quantitative potential of a morphological process (Bauer calls this the profitability of a rule; we will speak of quantitative productivity). There are different aspects of this potential, e.g., how many complex words have been produced by a given process $\mathfrak{P}$, how likely it is that $\mathfrak{P}$ will produce more words in the future, how the potential of a rule changes over time etc. A number of different measures and procedures to calculate these have been proposed (Baayen, 2001; Nishimoto, 2004; Meibauer *et al.*, 2004; Lüdeling and Evert, 2005). All of these measures use frequency data from corpora, mostly focusing on the distribution of the low-frequency words formed by a given process. Here, we argue that the statistical models that have been used so far do not compute what we intuitively understand as productivity, but rather confound three very different but closely intertwined quantities: (a) the *need* to express a certain concept – an extralinguistic property of communication; (b) the effect of *competition* from other processes that could express the same need and (c) quantitative *productivity* – the inherent facility with which a given word-formation process can fill the target need.

In order to illustrate our theoretical criticism of the traditional approaches to quantitative productivity, we use data that are particularly well-suited to our argument, as they pertain to a set of morphological processes whose semantic, connotational and morpho-syntactic properties are highly similar, i.e., they satisfy the same "need" and are thus in close competition. We turn now to a description of these data.

2 The 'too much' data

We will illustrate our point by looking at a number of German word formation processes that have a "too much" reading. These are *-itis*, *wahn*, *hysterie*, *zwang*, *sucht*, *besessenheit* and *obsession* in their use as heads of compounds.[1]. All of these morphological elements have several readings, one of which stems from medical terminology. In addition – probably by some kind of transfer – these elements have readings in everyday language where they mean something like "somebody does too much of *X*" or "somebody likes *X* too much" with an ironic connotation.[2] Examples that illustrate this reading are given in (1) and (2) below.

(1) Die Förderung der Dialekte und Regionalsprachen (ob als Erst- oder Zweit-Idiom) schützt vor Anglizismeritis! "The promotion of dialects and regional languages (whether as first or or as second languages) prevents the inflated use of anglicisms."

(2) Ich verstehe deinen Standpunkt bis zu einem gewissen Grad, da ich dem allgemeinen Anglizismenwahn (auch ?) sehr kritisch gegenüber stehe. "I understand your standpoint up to a certain degree, because I'm also very critical about the general inflated use of anglicisms."

We are interested only in this non-medical reading, which we call the TOOMUCH reading. We extracted all complex words ending in *-itis* etc. from deWaC, a 1.7 billion word corpus of German Web pages.[3]

[1] Note our assumption that neoclassical elements such as *-itis* participate in compounding, see Lüdeling *et al.* (2002)

[2] The medical reading still somehow resonates in this non-medical use, which can be seen from the fact that these words often collocate with expressions like *severe*, *suffer from*, etc.

[3] The deWaC corpus has been collected in 2005 by the WaCky initiative, see `http://wacky.sslmit.unibo.it`. All examples in this abstract are taken from deWaC and have not been edited in any way (so they may contain orthographic mistakes). In addition, instances of TOOMUCH nouns as well as their translations have been underlined in the examples.

Then we manually identified the TOOMUCH readings and extracted the non-heads of these compounds.[4] A qualitative analysis of the non-heads shows no clear categorial constraints (although there seem to be certain preferences): each process combines with verb-stems, nouns, and proper nouns. We will come back to these examples in Section 5.

3 Productivity

Many quantitative measures for the degree of productivity of a word-formation process have been proposed in the literature. Following Baayen (1989) and subsequent work, most measurements focus on mathematical models based on type-token statistics for each process $\mathfrak{P}$. While there are substantial differences in the statistical distributions underlying the models (see Baayen, 2001; Evert, 2004), the general principle behind them is quite intuitive: processes which produce a large number of rare types are considered highly productive, and processes for which one does do not find any rare types are assumed to be unproductive. Such measures of productivity are essentially based on surface counts. The implicit question they address is thus the following: "How likely is it that we will encounter a new type formed by $\mathfrak{P}$ after sampling a certain amount of data?"

Quantitative notions of productivity based on frequency data, however, do not measure what linguists are most likely to be interested in, namely an inherently linguistic quality of $\mathfrak{P}$ that one might call the 'cost' of using the process. The question is thus "how easy is it to use $\mathfrak{P}$?" rather than "how often do we use $\mathfrak{P}$?" The remainder of this abstract is concerned with two factors that might play a role in determining the frequency distribution of a morphological process, and that confound traditional type-token productivity calculations.

4 Need

Corpus counts are influenced not only by the inherently linguistic properties of morphological processes but also by something extra-linguistic: the situational or communicative need to express a certain concept. This has, of course, been noticed long ago and formulated time and again. Compare Hermann Paul:

> Die Möglichkeit zur Bildung von Zuss. aus zwei Substantiven ist unbegrenzt. Ob solche aber wirklich gebildet werden, hängt natürlich vom Bedürfnis ab. (Paul, 1920, 15) "The possibility to form noun-noun compounds is unlimited. Whether they are actually formed, however, depends on the need." (our translation)

or Laurie Bauer:

> Words are only formed as and when there is a need for them [...] (Bauer, 2001, 143).

Using our TOOMUCH data from the deWaC corpus, we can compute the frequency distribution of the TOOMUCH need across different concepts. For example, the frequency of the need to say that somebody talks or thinks too much about football can be estimated by adding up the occurrences of *Fußballitis*, *Fußballwahn*, *Fußballhysterie*, etc.[5] Interestingly, as Figure 1 clearly demonstrates (based on a small preliminary sample of the TOOMUCH data), the extra-linguistic *need* itself has the same typical Zipfian distribution that we came to expect from productive linguistic processes. On the one hand, it is intuitively plausible that many needs behave like productive processes due to purely extra-linguistic factors (e.g., the TOOMUCH need is productive because people feel stressed about many different things). On the other hand, this implies that the Zipfian frequency distribution of an apparently productive process $\mathfrak{P}$ may simply reflect the distribution of the need satisfied by $\mathfrak{P}$, rather than the inherent productivity of $\mathfrak{P}$ itself.

[4] For instance, from both *Anglizismeritis* and *Anglizismenwahn* we extracted the non-head *anglizismus* "anglicism". In the following, we will sometimes also refer to the non-head as the 'lexical root' to which a word-formation process applies.

[5] We make the simplifying assumption that there is a perfect mapping between concepts and the lexical roots.

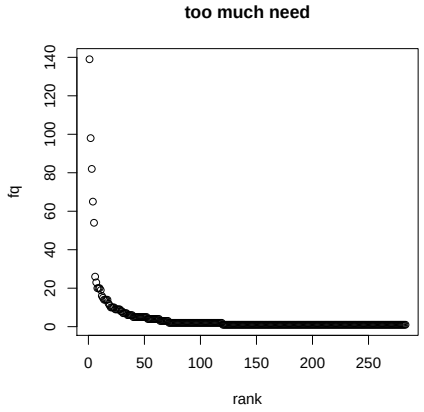

Figure 1: Rank-frequency plot of concepts undergoing a TOOMUCH process in the deWaC corpus.

5 Competition

When there is only a single morphological (or syntactic) process that satisfies a given need, we cannot draw any conclusions about the inherent productivity of this process because what we observe in the corpus data is merely the distribution of the need itself. However, a need can almost always be satisfied in more than one way. For example, consider the need to express one's opinion that somebody (X) is doing too much of something (Y). This need could be expressed morphologically by one of the TOOMUCH processes (and probably other morphological processes as well), but also by syntactic means. Here we focus on noun-forming morphological processes, since they can be inserted into the same syntactic environment and thus do not require changes in sentence planning. Except in cases where categorial constraints block all but one of the processes (these are the domain of qualitative productivity studies), the potential need satisfiers compete with one another. It should be noted here that we do not claim that all TOOMUCH processes compete in all possible contexts. But in contrast to Plag (1999), who argues that qualitative constraints prevent competition, we assume that there are many contexts in which the speaker has a choice. The following sentences (taken from the deWaC corpus) illustrate that most of the TOOMUCH processes can apply to the lexical root/concept *Fußball* "football" in very similar contexts:

(3) Das besondere an dem Buch ist, das man endlich versteht, was Fussballbesessenheit ausmacht [...] "What is special about this book is that you finally understand what football obsession is all about [...]"

(4) Rehhagels EM-Erfolg löst in Griechenland Fußball-Hysterie aus. "Rehhagel's success in the European Championship triggers football hysteria in Greece."

(5) Es tut bestimmt ganz gut, von Zeit zu Zeit auch einmal aus dem "Fußball-Wahn" [...] auszusteigen. "It is probably good for you to take a break from football-mania [...] every once in a while."

(6) In jungen Jahren ein Brillenkind, das wusste, wie der König von Tonga hieß, auf Bolzplätzen aber versagte, ist er heute seiner Fußballobsession vollständig verfallen. "After a childhood where he was the kid with glasses, who could name the king of Tonga but was a complete failure on the football pitch, he is now totally obsessed with football."

(7) Egal, diese blöde Fußballsucht macht doch eh nur Probleme. Streit mit der Freundin. Zeitverschwendung. "Anyway, this stupid football addiction only causes problems. Quarrels with my girl-friend. Waste of time."

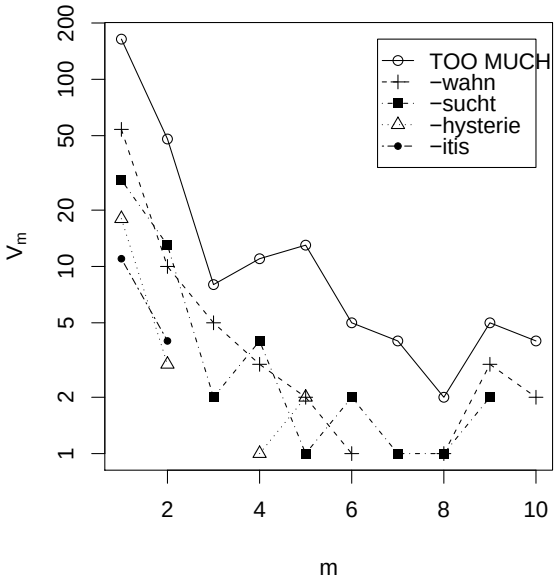

Figure 2: Frequency spectra of the TOOMUCH need and the processes satisfying it in the deWaC corpus (logarithmic scale).

While the issue of competition between morphological processes is often mentioned in the literature,[6] calculations of quantitative productivity typically look at single unrelated processes. When the productivity rates of different processes are compared, need or competition are not taken into account.

Assuming that all the processes that can satisfy a given need have comparable phonological, morpho-syntactic and semantic constraints (as is the case for our TOOMUCH data), there are still a large number of (probabilistic) factors that can make one of the processes more likely to win over its competitors when the need occurs. Some of these factors will depend on the lexical root that realizes a certain concept and will thus be relatively stable (e.g., non-categorial factors relating to the phonological shape or specific semantic structure of the root). As a result, one of the processes may be used much more often with specific roots, independent of the context. Other factors will be heavily context-dependent (e.g., those reflecting stylistic, discourse-related and collocational constraints), so that even for the same root different processes will win the race at different times.

In order to understand and model these effects, we have to look not only at the absolute frequencies of words formed by a certain process, but also at the relative distributions of the different processes that express the same need, which of them combine with a certain root/concept, and in what contexts such combinations occur. We are not aware of any model of quantitative productivity that takes competition into account, and we suspect that such a model would be of staggering mathematical complexity.

Indeed, if our current analysis is correct, it raises the question whether quantitative productivity can be defined as an inherent linguistic property of morphological processes at all (e.g., processes with higher productivity have a lower activation threshold or 'cost', which makes them more likely to win against competitors, all other factors being equal), or whether productivity is an epiphenomenon that can be completely reduced to need and the categorial factors that determine which of the rivalling processes wins the competition under what circumstances.

6 Ongoing and planned work

We are currently in the process of manually analyzing a sizable sample of the TOOMUCH data. We can then use these data to illustrate the relationship between the frequency distribution of the overall need, the frequency distributions of the different TOOMUCH processes, and the distributions of the TOOMUCH

need across processes for each concept. Preliminary results on the sample we have analyzed at this stage are presented in Figure 2. This first analysis already reveals interesting patterns for the most frequent processes, viz. *sucht* and *wahn* compounding, where the former appears to be a rather direct reflection of the need, whereas the latter has a smoother and more pronouncedly Zipfian shape. We expect that by comparing the distribution of the need with those of the individual processes on larger amounts of data we can achieve a better understanding of how competition and, possibly, inherent productivity modulate the expression of a need through a certain morphological process. Moreover, the data we have collected should allow us to run preliminary experiments that compare the relative distributions under different conditions, as a first attempt to discover the contextual factors governing the competition between word-formation processes.

References

Baayen, R. Harald (1989). *A Corpus-Based Approach to Morphological Productivity*. Ph.D. thesis, Vrije Universiteit de Amsterdam.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

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

Dressler, Wolfgang U. (2003). Degrees of grammatical productivity in inflectional morphology. *Rivista di Linguistica*, **15**, 31–62.

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

Lüdeling, Anke and Evert, Stefan (2005). The emergence of productive non-medical *-itis*: Corpus evidence and qualitative analysis. In S. Kepser and M. Reis (eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Mouton de Gruyter.

Lüdeling, Anke; Schmid, Tanja; Kiokpasoglou, Sawwas (2002). Neoclassical word formation in German. *Yearbook of Morphology 2001*.

Meibauer, Jörg; Guttropf, Anja; Scherer, Carmen (2004). Dynamic aspects of German *-er*-nominals: a probe into the interrelation of language change and language acquisition. *Linguistics*, **42**, 155–193.

Nishimoto, Eiji (2004). *A corpus-based delimitation of new words: cross-segment comparison and morphological productivity*. Ph.D. thesis, City University of New York.

Paul, Hermann (1920). *Deutsche Grammatik. Band V: Wortbildungslehre*. Max Niemeyer Verlag, Halle a.S.

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

[6] Syntactic alternatives – which are also part of the competition, of course – are usually ignored, though.