On Measuring Morphological Productivity[1]

Anke Lüdeling, Stefan Evert, Ulrich Heid • Institut für Maschinelle Sprachverarbeitung • Universität Stuttgart

Abstract

In this paper we discuss a number of pre-requisites for quantitative approaches to morphological productivity. We compare the degree of procutivity of three German adjective derivations computed with the productivity measure introduced in [2]. We show that this measure does not yield the expected results unless the data is pre-processed according to a very good understanding of the morphological process in question.

1 Introduction: Context and Objectives

In the development of computational systems for the analysis of unrestricted text, the productivity of morphological processes[2] is problematic because it is impossible simply to list all the words that can occur in the input. While unproductive patterns can be listed in the lexicon, productive processes must be described by rules. The productive processes also differ in the degree of their productivity. New formations of a highly productive process occur more frequently in the input than formations of a marginally productive one. Hence it is important to compare the degrees of productivity of different processes in order to be able to decide for which ones rules should be provided first if resources are limited.

In order to do this, more than an intuitive notion of productivity is required. Most literature on morphological productivity focuses on the qualitative aspects of specific word formation processes, giving a description of the syntactic, semantic, morphological, phonological, or other restrictions that a certain affix, for example, imposes on its bases. While such analyses are essential for the formulation of rules, they do not answer our initial question, namely, to what degree a process is productive. In [1, 2] Baayen proposed a measure for the quantitative analysis of productivity, which was successively refined during the following years, leading to the sophisticated statistical analyses for lexical statistics presented in [3]. Baayen's measures have been used in many publications for calculating morphological productivity (see [4, 7] for instance).

This paper is about the prerequisite conditions for using the productivity measure described in [2]. As these conditions pertain mostly to linguistic aspects and corpus preprocessing, they also apply to the more sophisticated mathematical models in [3].

In a case study, we analyzed German adjective formations in *–bar*, *–sam* and *–ös* extracted from a 36 million word corpus consisting of two years of the newspaper *Suttgarter Zeitung* (StZ, 1992/93). The intuition is that the derivational suffix *–bar* is productive, while *–sam* and *–ös* are completely unproductive. However, the naive application of Baayen's measure indicates that all three processes are productive, with *–ös* having the highest degree of productivity. A closer analysis of the data used in the measurements reveals the need for a far more detailed morphological specification of the material to be counted, and for high-quality corpus preprocessing. To our knowledge, these issues have not been addressed in the literature so far.

In Section 2 we will describe how Baayen's measure is applied. In Section 3 we will discuss various reasons for the counterintuitive results observed, and apply the productivity measure to manually cleaned-up data. Finally, we will discuss the results of our experiments in Section 4.

2 Baayen's productivity measure

Simplifying somewhat, we can summarize Baayen's approach to productivity measurement as follows:

1. The number N of occurrences of a given word formation process in the corpus is counted.
2. From the list of occurrences, an inventory of types can be derived; for each of the V different types, the number of occurrences is noted. We are particularly interested in the number n_1 of types which occur only once (hapax legomena): they could be unintentionally coined and thus evidence of productivity.
3. From the number n_1 of hapax legomena and N, the number of total occurrences of instances of the process, the so-called productivity index $P = n_1/N$ is

[1]This paper is the result of research carried out in the DeKo project (*Derivations- und Kompositionsmorphologie*), a project in the framework of the *Forschungsschwerpunktprogramm des Landes Baden-Württemberg*. We would like to thank Peter Bosch, Arne Fitschen, Bernd Möbius, Bettina Säuberlich, and Tanja Schmid for their comments on earlier versions of this paper.

[2]"[...] the possibility for language users to coin, unintentionally, a number of formations which are in principle uncountable" ([8], translated by [2], p. 109).

computed. It is assumed to be smaller for unproductive processes, and bigger for productive ones.

According to Baayen's statistical model, the productivity index P corresponds to the rate at which new types are expected when more tokens are sampled (cf. [3], Sec. 2.3). It is therefore dependent on N. To assess the degree of productivity of a word formation process, one wants to follow the growth of V for corpora of increasing size: For unproductive (or marginally productive) processes, increasing corpus size does not add new types (or very few ones) after a certain saturation point, and hence a plot of V (number of types) against N (number of tokens) shows close to no growth after that point (see the left panel in **Figure 1**). For productive processes, V continues to increase with N as new hapaxes are encountered, accounting for the fact that people have "unintentionally [coined] formations which are in principle uncountable" (right panel in Figure 1).

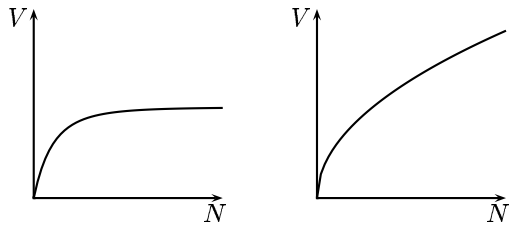

Figure 1: Typical plots of V against N

raw	N	n_1	P
–ös	4383	70	0.0160
–sam	22667	78	0.0034
–bar	37783	324	0.0086

Table 1: Growth rates P for the complete StZ corpus (raw data)

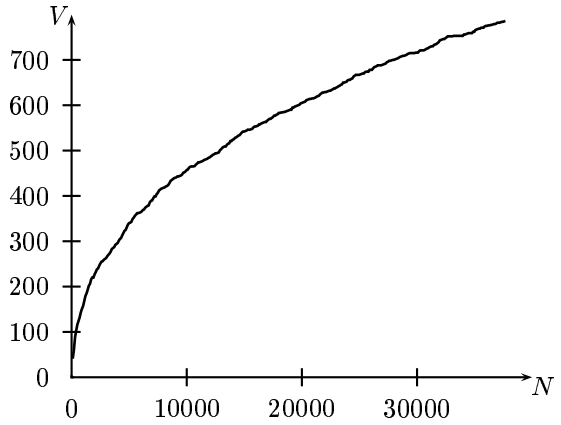

Figure 2: Relation between token count N and type count V (*–bar* derivation, raw data)

3 Applying the productivity measure

In order to compare the degree of productivity of adjective-forming suffixes in German we used Baayen's formula to compute the productivity index P for the adjective forming suffixes *–ös*, *–sam* and *–bar*. *–ös* roughly translates to *–ous*, *–sam* to *–ful*, and *–bar* to *–able*. The literature ([5]) as well as our intuitions suggest that *–bar* is highly productive while *–ös* and *–sam* are at best marginally productive.

We extracted[3] adjectives in *–ös*, *–sam* and *–bar* from the StZ corpus: **Table 1** gives the number N of occurrences, the number n_1 of hapax forms, and the calculated value of P. Plots of V against N are shown in **Figures 2, 3, and 4**. These results suggest – contrary to our intuition – that all three formation types are productive and that adjectives in *–ös* are more productive than adjectives in *–bar*.

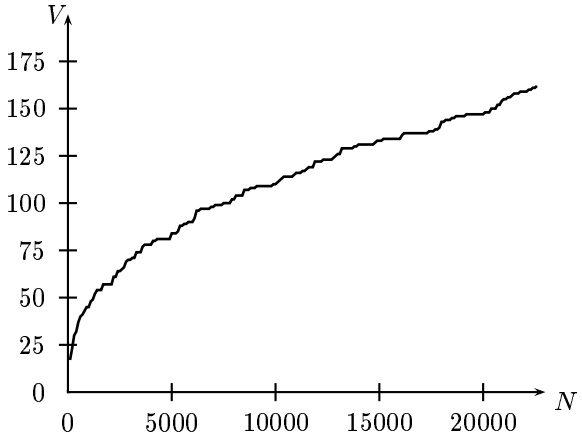

Figure 3: Relation between token count N and type count V (*–sam* derivation, raw data)

3.1 Problems

There are two kinds of factors that lead to the surprisingly high productivity index of *–ös* and *–sam*: corpus preprocessing and linguistic factors. Both kinds of problems

[3]The extraction was done with standard corpus tools (the CWB corpus workbench, cf. [6]).

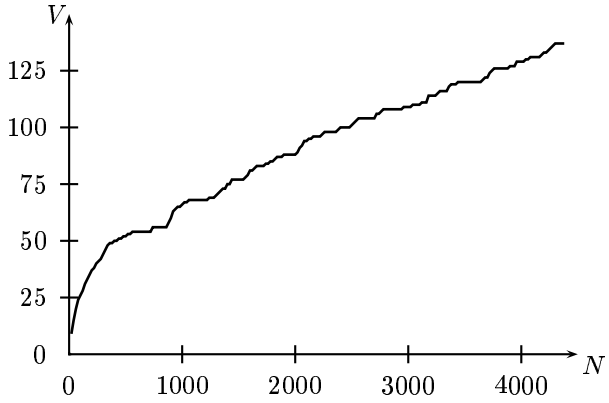

Figure 4: Relation between token count N and type count V (*–ös* derivation, raw data)

can be illustrated by the data in **Table 2**. Table 2 shows the so-called grouped frequency distribution of *–ös*-adjectives. For a grouped frequency distribution the different types are ordered by their frequency r and types with the same frequency are grouped together; n_r is the number of types which occur exactly r times in the sample. The top part of Table 2 shows that there is one type which occurs 788 times in our corpus (*bös* "evil"), one type that occurs 760 times (*religiös* "religious") etc. At the bottom of the table the low frequency types are listed. These typically occur more than once. There are, for example, two types that occur 8 times (*ambitiös* "ambitious", *Rös*, an author identification code, see below), 15 types that occur twice and 47 hapax legomena.

n_r	r	
1	788	bös
1	760	religiös
1	322	seriös
1	171	mysteriös
1	103	skandalös
1	101	luxuriös
1	98	nervös
...	...	...
2	8	ambitiös, Rös
4	6	inzestuös, minuziös, ingeniös, jös
4	5	maliziös, pseudoreligiös, libidinös, geistig-religiös
1	4	kapriziös
3	3	multireligiös, ultrareligiös, freireligiös
15	2	pathetisch-pompös, hypernervös, tumultuös, ethnisch-religiös, ...
47	1	gestelzt-seriös, hymnenhaft-melodiös, parareligiös, mystisch-mysteriös, übernervös, ...

Table 2: Grouped frequency distribution for *–ös*

As stated above, the first set of problems arises in corpus pre-processing. Baayen does not mention these problems and seems to assume perfectly prepared corpora. In reality, text corpora contain a considerable number of errors. Since these often occur only once and thus add to the number of hapax legomena they can have considerable effects on the result. Corpus preprocessing problems include the following:

1. Mistagged items (*Rös* and *jös* in the *–ös* example are not adjectives but author identification codes, left over, in our version of the StZ, from the original typesetting material of the newspaper; further examples are names such as *Erdös*).

2. Typographic errors (**offebar* instead of *offenbar* "apparently") and tokenizing errors (*schnell/langsam* "fast/slow" should be two tokens).

3. Corpus composition and repetitivity of texts. Often sentences and even whole articles are repeated in corpora.

Certainly, these problems have to be taken into account. However, even an error-free corpus does not allow the direct application of Baayen's methods. Without a very good understanding of the morphological process in question it is not clear how to interpret the computed values of the productivity index P. Instances of morphological processes other than the targeted process are commonly found in the sample data.

1. Forms that only accidentally end in the same affix: *bös* "evil", for example, is not an *–ös* derivation but a stem.

2. Derivation vs. compounding: *Religiös* "religious" is an instance of *–ös*-derivation, but *geistig-religiös* "spiritual-religious", or the non-hyphenated *pseudoreligiös* "pseudo-religious" are produced by adjective-adjective compounding or *pseudo–* prefixation. Hence they should not be counted as hapax legomena when one considers *–ös*-derivation.

3. Creativity vs. word formation: our data contain *unterhaltsam* "entertaining", but also *alleinunterhaltsam* (as a hapax legomenon). The latter is clearly neither an example of *–sam*-derivation nor of compounding (the analysis is not [*allein+unterhaltsam*]): it may appropriately be analyzed as created along the model of *unterhaltsam*, from *Alleinunterhalter* "solo entertainer", and should not be counted into n_1.

4. Nature of the bases: There is a need for guidelines for the handling of complex bases. Some negated adjectives of the *un–V–bar* type have no *V–bar* counterparts, others have. The former may be instances of derivation in *–bar* (and hence add to the V and

n_1 counts), but not the latter, which belong to a single group: *unübertragbar* and *nichtübertragbar* "non-transferable" are instances of prefixation applied to the same *–bar*-adjective (*übertragbar* "transferable"). Particle verbs have to be counted as separate bases. *Absehbar* "predictable" cannot be analyzed as a compound from *sehbar* "visible" and *ab*. It is sometimes difficult to distinguish between compounding and derivation from particle verbs.

These linguistic problems cannot be handled automatically but have to be dealt with semi-automatically or manually. For each problem type, a set of guidelines is needed.

3.2 Refined measurements

We cleaned up the word-lists for *–ös*, *–sam* and *–bar* adjectives, doing as much of the work as possible automatically and then manually checking the rest. We counted compounds and *un–* derivations as instances of their head word and spelling mistakes as instances of the correct word. This reduced the number of hapax legomena considerably. Words formed by other processes than the targeted process were eliminated from the sample.

Then we re-computed the productivity rates. **Table 3** shows the values of N, n_1 and P for the cleaned-up data. The solid lines in **Figures 5, 6 and 7** show the relation between N and V after manual cleanup (for comparison, the curves from Figures 2, 3, and 4 are repeated as dashed lines). The results now correspond to our intuitions: *–bar* derivation is productive while *–ös* and *–sam* are not.

cleaned-up	N	n_1	P
-ös	3404	5	0.0015
-sam	22654	5	0.0002
-bar	35562	189	0.0053

Table 3: Growth rates P for the complete StZ corpus (cleaned up)

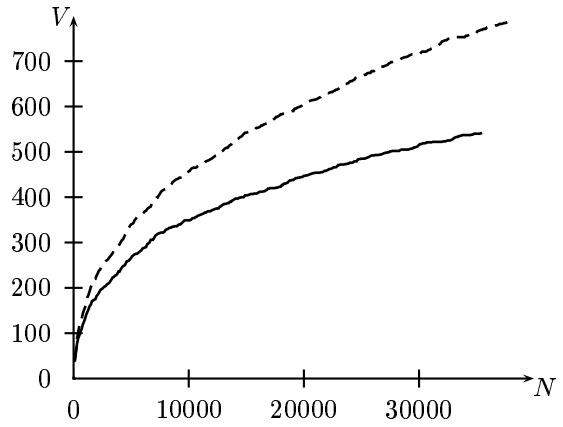

Figure 5: Relation between token count N and type count V (*–bar* derivation, cleaned up)

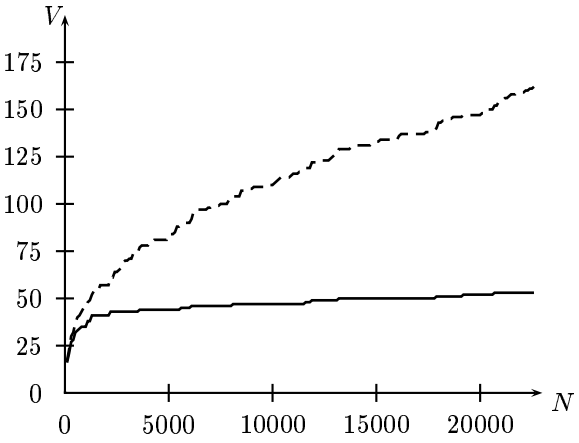

Figure 6: Relation between token count N and type count V (*–sam* derivation, cleaned up)

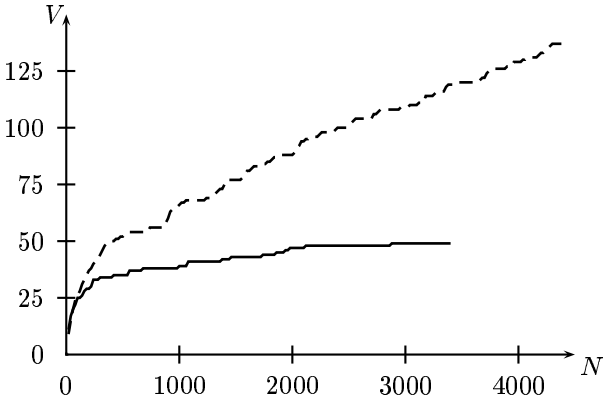

Figure 7: Relation between token count N and type count V (*–ös* derivation, cleaned up)

4 Conclusions

Our case study leads to the following conclusions:

- Baayen's productivity measure produces usable and linguistically significant results.
- However, it can be applied only after thorough preprocessing of the targeted data:
  - Morphological pre-analysis: the measure must be applied to a clearly defined word formation process. This definition must include a specification of the allowed bases and a clear separation from other processes.

– Corpus-linguistic preprocessing: the material must be of high quality with respect to tokenizing and tagging.

The morphological preanalysis is not automatic, since no sufficiently sophisticated morphology system is available. Productivity measurements can be employed for linguistic as well as engineering purposes only when they are accompanied by a description of the criteria applied in preprocessing.

Unfortunately, productivity measurement did not prove to be error tolerant, but the procedures needed on top of the mere statistics to provide significant and interpretable figures are becoming clearer.

References

[1] R. Harald Baayen. *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. PhD thesis, Vrije Universiteit Amsterdam, 1989.

[2] R. Harald Baayen. Quantitative aspects of morphological productivity. In Geert Booij and J. van Marle, editors, *Yearbook of Morphology 1991*, pages 109 – 150. Foris, Dordrecht, 1992.

[3] R. Harald Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, 2000.

[4] R. Harald Baayen and A. Renouf. Chronicling The Times: productive lexical innovations in an English newspaper. *Language*, 72:69 – 96, 1996.

[5] Wolfgang Fleischer and Irmhild Barz. *Wortbildung der deutschen Gegenwartssprache*. Max Niemeyer Verlag, Tübingen, 1992.

[6] Esther König, Oliver Christ, Bruno M. Schulze, and Anja Hofmann. *CQP User's Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1999. `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`.

[7] Bernd Möbius. Word and syllable models for German text-to-speech synthesis. In *Proceedings of the Third ESCA Workshop on Speech Synthesis*, pages 59–64, Jenolan Caves, Australia, 1998. ESCA.

[8] H. Schultink. Produktiviteit als morphologisch fenomeen. *Forum der Letteren*, pages 110 – 125, 1961.