Measuring morphological productivity: Is automatic preprocessing sufficient?

Stefan Evert & Anke Lüdeling,
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany.
e-mail: {evert, anke}@ims.uni-stuttgart.de

1. Morphological productivity

In this paper we want to focus on a small facet of morphological productivity: on quantitative measures and their applicability to "real life" corpus data.[1] We will argue that – at least for German – there are at present no morphological systems available that can automatically preprocess the data to a quality necessary to apply statistical models for the calculation of productivity rates.[2]

Before coming to the quantitative aspects we want to clarify the notion *morphological productivity*. Morphological productivity has long been a topic in theoretical morphology (see for example Schultink 1961, Aronoff 1976, van Marle 1985, and Plag 1999). It has been defined in many ways. We choose a definition by Schultink (1961, p. 113) which contains three aspects that are important to us:

> We see productivity as a morphological phenomenon as the possibility for language users to coin unintentionally an in principle unlimited number of new formations, by using the morphological procedure that lies behind the form-meaning correspondence of some known words.[3]

The three important aspects are *unintentionality*, *unlimitedness*, and *regularity*. They are all interdependent. The first aspect – unintentionality – helps us to distinguish between productivity (which is a linguistic rule-based notion) and creativity (which is a general cognitive ability and cannot be captured within morphology alone): Words formed by productive processes are often not recognized or noticed as new words (this is true for speaker and hearer) while words formed by other (creative) processes are carefully produced and are perceived as new words. The second aspect is unlimitedness – if productive word formation patterns are in principle unlimited, it is not possible to give a finite list of words (some implications of this are discussed below). Both unlimitedness and unintentionality require that the words formed by a given process are morphosyntactically and semantically regular.

Theoretical and descriptive works on word formation mostly focus on what Baayen (1992) calls the *qualitative* aspect of productivity: the morphological, phonolocical, syntactic, semantic and other restrictions of a specific word formation process are studied. The adjective-forming suffix *–bar* (roughly comparable to English *–able*), for example, takes as bases transitive activity verbs (*lesbar* "readable" from transitive *lesen* "to read", but not **schlafbar* from intransitive *schlafen* "to sleep" or **wissbar* from stative *wissen* "to know"), the noun-forming circumfix *Ge- -e* does not take prefix verbs (*Geschimpfe* "continued or iterative scolding" from simplex *schimpfen* "to scold" but not **Gebeschimpfe* from prefixed *beschimpfen* "to insult"). The goal is an intensional description of the possible bases for a given word formation process. Next to qualitative approaches to productivity, *quantitative* approaches are suggested (Baayen 1992, Baayen and Lieber 1991, Baayen 2001). These aim at calculating the probability of finding a new word formed by a given morphological process in a text once a given amount of text is sampled (see Section 2).

What is the relevance of quantitative approaches to productivity? What can quantitative approaches tell us that qualitative approaches cannot? At first glance the qualitative view of productivity differs fundamentally from the quantitative view: A pattern X can be fully productive in the qualitative analysis – it applies to all possible bases – but unproductive in the quantitative analysis, if there is a finite number of bases and all the words that pattern X can possibly form from these bases have been sampled. The probability of encountering new words formed by that word formation process is then 0. But qualitative and quantitative approaches really complement each other. First of all, quantitative approaches cannot be used without careful linguistic interpretation, as we will see below. From the

---

[1] We illustrate our points with German data from the StZ corpus, which contains roughly two years of the newspaper *Stuttgarter Zeitung* (1991 – 1993, 36 million tokens). We suspect that the problems that arise when corpus data is used for quantitative approaches to productivity are similar for other languages.

[2] We will not deal with the statistical models or the mathematics behind them in this paper. Excellent work on morphological productivity has been done by Harald Baayen in a series of articles over the last 10 years (starting with Baayen 1992). He gives a detailed overview over the LNRE models that can be used to calculate productivity in Baayen (2001).

[3] Our translation of "Onder produktiviteit als morfologisch fenomeen verstaan we dan de voor taalgebruikers bestaande mogelijkheid door middel van het morfologisch procédé dat aan de vorm-betekeniscorrespondentie van sommige hun bekende woorden ten grondslag ligt, onopzettelig een in principe niet telbaar aantal nieuwe formaties te vormen."

opposite perspective, quantitative studies help us to find out more about the nature of word formation processes: Baayen and Neijt (1997) show, for example, that the words formed by some word formation patterns really belong to two different distributions: the lexicalized words behave statistically different from the productive words. This insight is made possible by studying the shape of the distribution and analysing its anomalies. See also Baayen and Lieber (1991) for an overview of how linguistic knowledge and quantitative results profit from each other.

In addition to being a valuable facet of the linguistic study of productivity, quantitative approaches also make a significant contribution to computational morphology. Many applications (for example machine translation, dialogue systems, text-to-speech systems) have to deal with unseen text. This involves not only parsing unseen sentences but also analysing new words. Because of productivity it is not possible to have a finite lexicon containing all the words of a language. Quantitative productivity studies help us determine which patterns can be listed and for which patterns rules have to be formulated.

In Section 2 we will introduce a few statistical notions necessary for the understanding of productivity measures and sketch how statistical models dealing with productivity work. These methods depend on corpus data. But using corpus data is quite problematic, as discussed in Section 3. The data have to be thoroughly preprocessed before quantitative measures can be applied. We will investigate how automatic preprocessing compares with manual preprocessing in Section 4.

2. Vocabulary growth curves

A minimal understanding of the statistical ideas is essential as a background for the following sections. We do not have the space to introduce the statistical assumptions and models in detail and have therefore tried to define the necessary notions rather intuitively. For a formal introduction to LNRE models refer to Baayen (2001).

For a quantitative approach to productivity, we need a mathematical definition of the *degree of productivity*, based on observable quantities. Following Schultink's definition of morphological productivity, we define the *vocabulary* of a morphological process as the number of *types* (i.e different lemmata) that the process can potentially generate. A productive pattern is, in theory, characterised by an *infinite* vocabulary (cf. the notion of *unlimitedness* in Schultink's definition), whereas a totally unproductive pattern is expected to have a *finite*, and often quite small, vocabulary.

In order to estimate the *vocabulary size S* of pattern X from a corpus, we look at the subset of the corpus consisting of all tokens generated by X (e.g. all tokens with the suffix *–bar*), in the order in which they appear in the corpus. Let *V* be the number of different types among the first *N* tokens in the subset. Plotting *V* against *N*, we obtain *vocabulary growth curves* as shown in Figure 1. If we had an infinite amount of data, we would eventually sample the entire vocabulary of X, and *V* would converge to *S*. The left panel of Figure 1 shows the typical shape of the vocabulary growth curve of an unproductive process. After enough data has been sampled, the curve flattens out and converges to a constant value, the full vocabulary size *S*. In the case of an infinite vocabulary, *V* would continue to grow indefinitely (see the right panel in Figure 1).

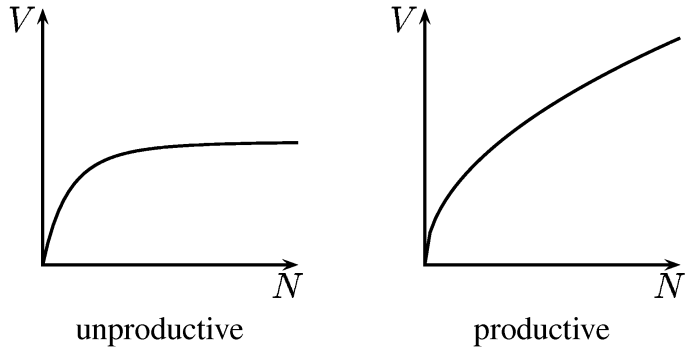

Figure 1: Typical shapes of vocabulary growth curves (idealised).

Baayen (1992) uses the slope *P* of the vocabulary growth curve after the entire corpus has been sampled as a measure of productivity (*P* is referred to as the *productivity index* of pattern X). Intuitively speaking, *P* is the probability that the next token generated by X is a previously unseen type,

i.e. a new complex form. Under certain simplifying assumptions we find that $P = \frac{n_1}{N}$, where n_1 is the number of hapax legomena (words occurring only once, cf. Baayen 2001, Chapter 2).

Unfortunately, *P* cannot be used to compare the productivity of affixes with substantially differing sample sizes. If we look at the left panel in Figure 1, we see that for the full sample, *P=0* as we would expect from a completely unproductive process. However, had we looked at the first 15 or 20% of the curve (e.g. because we are working on a small corpus, or sampling a rare pattern), we would have observed a much larger value of *P*. Larger, perhaps, than that of the productive process in the right panel (for the full sample). Since we do not know where in the curve we are, we cannot simply compare *P* values for processes of different sizes.

In order to compare different processes, we therefore need to be able to extrapolate the value of *P* to larger sample sizes *N*, or, equivalently, extrapolate the shape of the vocabulary growth curve. We cannot rely on standard statistical models for this purpose because the type frequency distributions of productive patterns are so-called LNRE distributions (for "large number of rare events"), in which low-frequency types (including hapax legomena, but also types occurring two, three, etc. times) account for a major part of the vocabulary. Ordinary statistical techniques are based on assumptions that do not hold for LNRE distributions (the most prominent one being the law of large numbers). Hence, Baayen (2001) introduces specialised models, for which parameter estimates are obtained from the counts of low-frequency types in the corpus (the *frequency spectrum*). These models can then be used to extrapolate the vocabulary growth curves to larger values of *N*. Baayen's comparison of the degrees of productivity of different patterns is based on the predicted vocabulary sizes.

Any errors in the type counts have a direct influence on the model parameters, and thus on the predicted vocabulary size. It is therefore essential to correct the input data for the errors described in Section 3.[4]

3. Problems with corpus data

Since we want to compute the probability *P* of a new complex word formed by a given word formation pattern *in a given text type* (for example newspaper data or technical text) and because the productivity of word formation patterns is highly dependent on text type (neoclassical words, for example are much more likely in scientific texts than in everyday language, see also Baayen 1994) it is necessary to sample a corpus of that text type. Dictionary data cannot be used for our purpose for two reasons: (i) Dictionaries often contain obsolete words but typically do not contain regular newly formed words. (ii) The advanced statistical models introduced in Section 2 are based on vocabulary growth curves that can only be computed from corpus data.

If we want to apply the measures described in Section 2 to real corpus data we encounter a number of problems. Lüdeling, Evert, and Heid (2000) describe these problems and show that the data have to be preprocessed before they can be used. In this section we want to summarise the problematic factors briefly, using two adjective forming suffixes (*–sam* and *–bar*) for illustration. The intuition is that *–sam* is unproductive while *–bar* is productive. If we simply extract all adjectives ending in the letters *sam* and *bar* from the STZ corpus and calculate *P* from the frequency spectra we get very similar graphs that suggest that both processes are productive (the solid lines in Figure 2). A closer look at the data reveals that the data have to be corrected for the following factors:

- incorrect data
    - mistagged items
    - misspelt items
    - corpus structure: repeated articles and sections influence the frequency distribution
- linguistic factors: many complex words that look like they are formed by the word formation process in question are really formed by other processes:
    - compounds: these make up the largest portion of the hapax legomena of most word formation patterns. We have to distinguish between complex words where the

[4] Baayen (2001) also describes more sophisticated models (*adjusted* and *mixture distributions*) which are estimated from the (known) shape of a pattern's vocabulary growth curve. Just as it is the case for the simpler models, errors in the type counts have a significant influence on the shape of the vocabulary growth curve and hence on the predicted vocabulary size.

compounding happens before the derivation (with the structure ((stem+stem)+affix)) and cases in which something attaches to an already affixed word (stem+(stem+affix)). The former cases have to be counted as new types while the latter cases have to be counted as instances of the affixed word. We will discuss an example in Section 4.

- other complex bases: likewise we have to distinguish between cases in which the affix in question attaches before other affixes or after them. For *–bar* adjectives that also contain a prefix, for example, we have to decide whether the prefix is attached to the *–bar* adjective (like the negation prefix *un-* which attaches to *verzichtbar* in *unverzichtbar* "indispensable" and thus should not be counted as a new type for the *-bar* pattern) or to the verb which is the base for *–bar* adjectivisation (*fahrbar* "ridable" is formed from the simplex verb *fahren* "to ride" while *befahrbar* "passable" is formed from the prefix verb *befahren* "to pass over, to drive on" – in this case we have two new types).
- words that "accidentally" end in the same letters: *Balsam* "balm", for example, does not contain the suffix *–sam.*
- words formed by creative rather than by productive processes: *kinobetriebsam* is a creative formation merging *betriebsam* "busy" and *Kinobetrieb* "movie theatre business".

As we emphasised in Section 2, the correct application of the guidelines is so important because many of the problematic cases are hapax legomena and thus have a direct influence on the productivity index *P*.

In Lüdeling, Evert, and Heid (2000) we manually corrected the data for *–sam* and *–bar* according to these guidelines. The dotted lines in Figure 2 show the vocabulary growth curves for the corrected data. They now conform to intuition: *-sam* is clearly unproductive while the curve for *-bar* still shows a productive pattern.

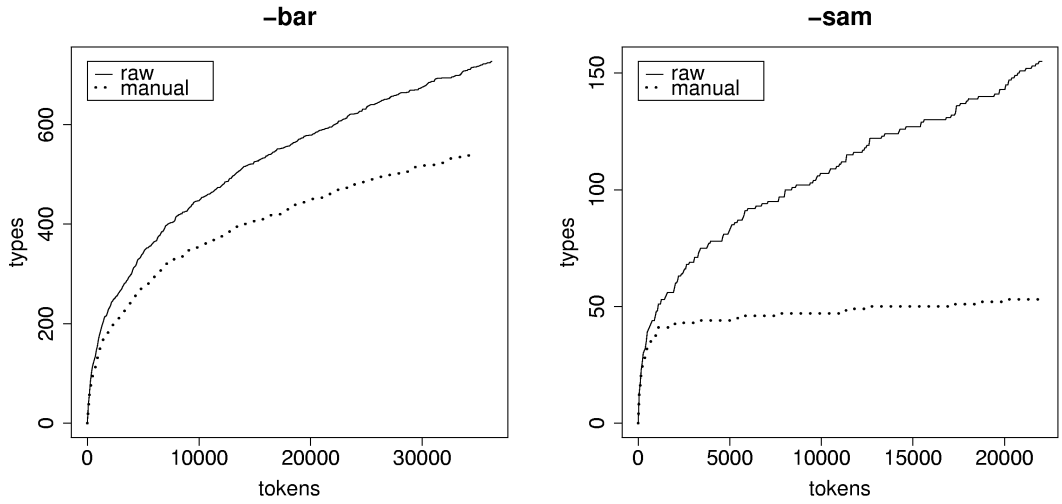

Figure 2: Raw and manually corrected vocabulary growth curves for *–sam* and *–bar*.

4. Automatic preprocessing vs. manual preprocessing

It is not feasible to correct the input for all word formation processes manually since some word formation patterns are very large. It would, therefore, be desirable to preprocess the input automatically. In this section we want to explore how automatic preprocessing compares to manual correction. We will find that automatic preprocessing – at least with the tools that are available to us – is not yet a usable alternative to manual preprocessing.

Before we describe the programs we used we want to recapitulate the requirements for morphological preprocessing: in order to perform like a human corrector, a morphology program would have to be able to (a) find spelling and tagging mistakes, (b) make a morpheme analysis, and (c) provide the

correct hierarchical structure for complex words consisting of more than two morphemes. At least for German, there are no programs available that meet these requirements for unseen words.[5]

Because no fully fledged morphology analyser is available, we wanted to find out how much preprocessing can be done automatically with the available tools. We chose:

- MORPHY, a freely available German morphological analyser (see Lezius, Rapp, and Wettler 1998). It was developed mainly for the analysis and lemmatisation of inflected forms, as well as context-sensitive part-of-speech tagging. In our experiment we used the built-in lexicon of MORPHY together with a simple heuristic compound analysis.
- DMOR, a two-level morphology system developed at the IMS Stuttgart (cf. Schiller 1996). Like MORPHY, it is intended mainly for generating and analysing inflected forms, but it includes compound analysis as well: compounds are analysed as sequences of stems and linking elements (somewhat restricted by simple heuristic patterns). The DMOR lexicon contains 65,000 stems (mostly simplexes, but some complex words are listed).

Neither program analyses derivational morphology. The built-in stem lexicon of DMOR is considerably larger than that of MORPHY.

We manually corrected the output for a number of word formation patterns and compared the vocabulary growth curves of uncorrected (raw), manually corrected and automatically corrected data. Figure 3 shows the manually and automatically corrected plots for *–bar* and *–sam*. The automatically corrected curves lie between the uncorrected curve and the manually corrected curve in both cases. For *–bar*, the results produced by MORPHY (dashed lines) are hardly different from the uncorrected curve; DMOR (broken lines) performs somewhat better. For *–sam*, the automatically corrected curves are much closer to the manually corrected curve. Qualitatively, the autocorrected curves are similar to the manual result and hint at an unproductive process (although not as clearly as the manually corrected data). Why does automatic correction work so much better for *-sam* than for *-bar*? This is due to the fact that almost all of the hapax legomena in the *–sam* spectrum are compounds involving only very few head types. The compound analysis components of DMOR and MORPHY seem to be able to deal with these cases quite well. The hapax legomena in the *–bar* spectrum are much more diverse: *un-* prefixations, many spelling errors and very few compounds – since MORPHY and DMOR do not have derivation analysis components, they cannot correct these cases.

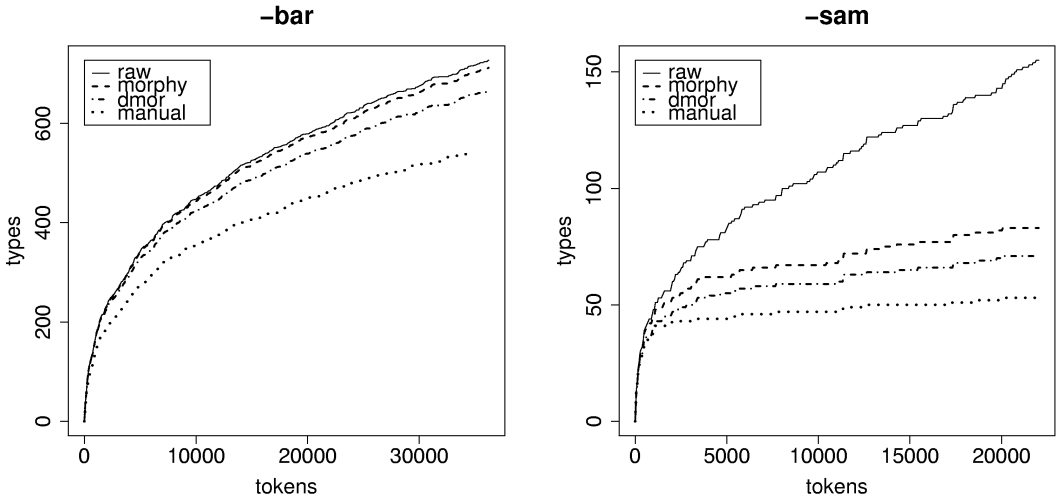

Figure 3: Vocabulary growth curves for *–bar* and *–sam*.

The next process we look at is the diminutive suffix *–chen* which combines with nouns (and sometimes adjectives) to form nouns. Here again we have a pattern where the automatically corrected curves are

[5] There are a number of morphology programs which analyse complex words
http://services.canoo.com/MorphologyBrowser.html
http://wortschatz.uni-leipzig.de/
http://www.linguistik.uni-erlangen.de/cgi-bin/orlorenz/dmm.cgi
As far as we can see from the available versions, they all work with finite lexicons. There are no downloadable tools or word lists. None of them provides hierarchical structure.

somewhere between the uncorrected curve and the manual curve, with DMOR (with its larger lexicon) giving considerably better results than MORPHY (Figure 4). Note, however, that the manually corrected curve has become much *shorter*, i.e. the number of tokens has been reduced. This is due to the fact that many words "accidentally" end in *–chen* (such as *Groschen* "ten-pfennig piece", *Drachen* "dragon", or *Zeichen* "sign"). Because of the resulting smaller sample size the manually corrected curve cannot be compared directly to the other curves (compare Section 2). However, the shape of the former suggests that *–chen* is much less productive than the automatically corrected curves would predict.

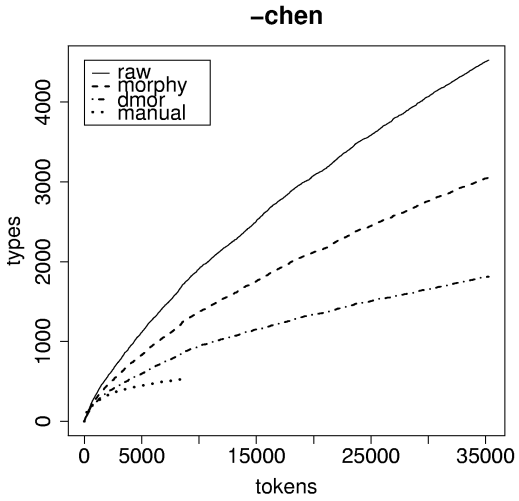

Figure 4: Vocabulary growth curves for *–chen*.

To summarize the results so far: The improvement achieved by automatic preprocessing differs considerably for the various processes. It never reaches the quality of manual preprocessing: in the case of *–bar* there is little improvement over the raw data, in the case of *–sam* the automatically corrected data comes close to the manual curve, and for *–chen*, automatically corrected curves are approximately in the middle between the uncorrected curve and the manual curve. For *–chen*, we also observe the largest difference between the two morphology systems. Since we cannot predict how good the results are compared to manually corrected data, automatic preprocessing is no real alternative for manual correction. However, it would seem that the morphology systems always lead to some improvement and should therefore always be applied (manual checks may then be done on the automatically cleaned-up data).

However, when we look at another diminutive pattern (suffixation with *–lein*) and a process that is semantically similar (compounding with *klein* "small"), we see a strikingly different pattern (Figure 5). The vocabulary growth curves of *klein* compounds show the familiar pattern of a gradual improvement from raw to manually corrected data. If we look at the plots for *–lein*, the uncorrected curve and the manually corrected curve are very similar to those of the *klein* compounds. However, when we automatically correct the data with DMOR we obtain a curve that is far *below* the manual curve, and hence predicts a considerably lower productivity than we actually find. This is due to the fact that DMOR contains the noun stem *Lein* "flax" and therefore analyses many derived words as compounds with the head *Lein*.

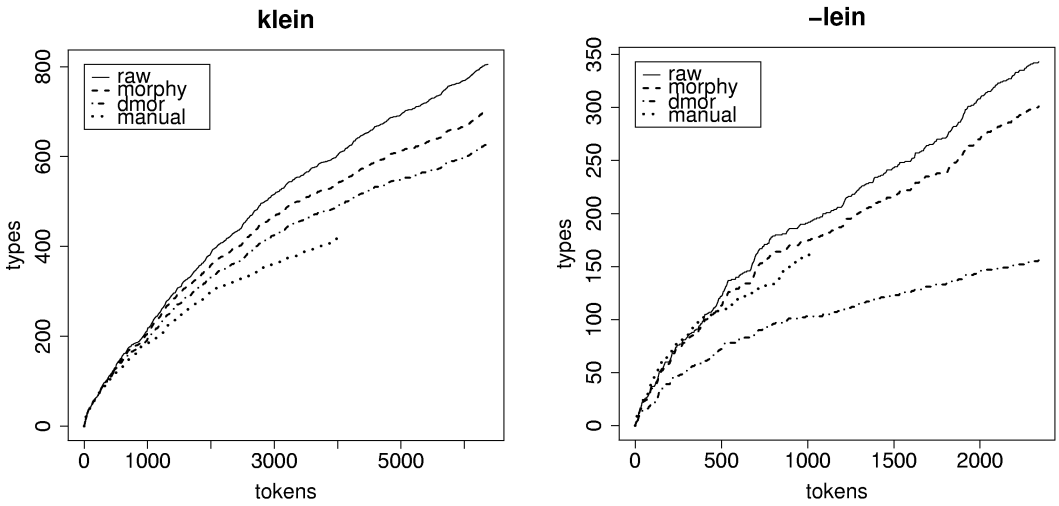

Figure 5: Vocabulary growth curves for *klein* and *–lein*.

One could argue that the *Lein/-lein* homography might be a single example, and can easily be corrected if we disallow reduction to the simplex *Lein.* The curves are given in Figure 6 which shows the familiar pattern where the automatically corrected curves are between the manually corrected curve and the uncorrected curve.

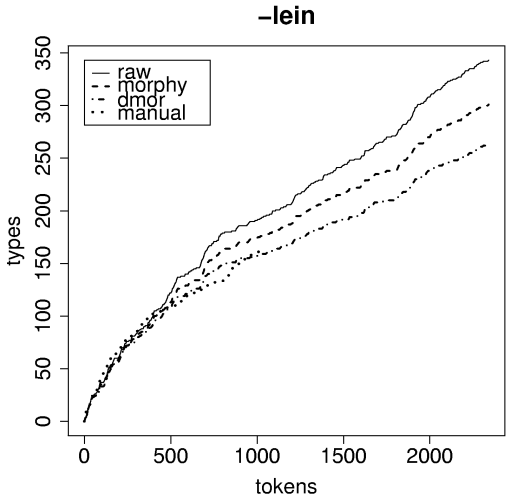

Figure 6: Vocabulary growth curves for *–lein* (DMOR curve corrected).

However, Figure 7 shows the same pattern for the noun-forming affix *–tum.* Again, the curve produced by MORPHY is above the reference curve, whereas the curve produced by DMOR is below the reference curve. *–tum* is an interesting case: DMOR finds roughly the same number of types as the human annotator, but a considerably larger number of tokens. Although this would explain the seemingly lower productivity indicated by the DMOR curve, the actual types are different, due to two factors: as in the case of *–chen*, there are many words ending with the letter sequence *tum* which are not derived by the suffix *–tum* (examples are *Datum* "date", *Faktum* "fact", or *Quantum* "quota") – DMOR mistakenly counts these types. The other factor has to do with hierarchical structure: There are many compounds of the type NOUN+*reich*+*tum* which DMOR reduces to *Reichtum* "wealth". The correct structure, however, is ((NOUN+*reich*)+*tum*) which means that these cases have to be counted as different types (an example is *Kinderreichtum* "having many children" which is derived from *kinderreich* "prolific", and not a compound of *Kinder* "children" and *Reichtum*). Here we need careful linguistic analysis – a simple automatic solution is not possible.

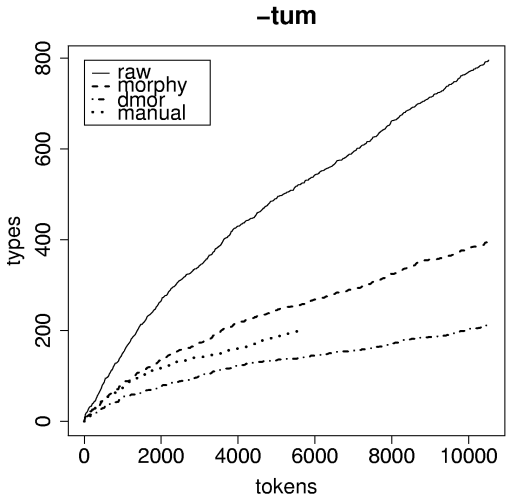

Figure 7: Vocabulary growth curves for *–tum*.

5. Conclusion

Corpus data are necessary for the application of statistical models to the quantitative analysis of morphological productivity. We summarised Lüdeling, Evert, and Heid (2000), who showed that corpus data have to be thoroughly preprocessed before they can be used in these models. Since manual preprocessing is not feasible for large word formation patterns, we wanted to find out whether automatic preprocessing is a usable alternative. We found that although automatic correction, based on the currently available morphology systems, yields an improvement over the uncorrected data in many cases, it is no replacement for manual corrections.

In some cases, automatic preprocessing is possible – if only as a basis for further manual correction. However, the over-compensation we observed for *–lein* and *–tum* shows that fully automatic preprocessing may produce misleading results. Without a manually corrected reference curve we do not know where the automatic curves for a given word formation process lie. This means that we cannot even use automatic preprocessing as a reliable basis for manual correction. Only a morphology system which, in addition to derivation and compounding components, includes a model of the order in which processes operate on a simplex form (producing a hierarchical analysis of complex words) can be expected to lead to sufficiently reliable automatic correction results for quantitative studies of productivity.[6]

6. References

Aronoff M 1976 Word Formation in Generative Grammar. Cambridge, MIT Press.

Baayen R H 1992 Quantitative aspects of morphological productivity. *Yearbook of Morphology1991*: 109 – 150.

Baayen R H 1994 Derivational productivity and text typology. *Journal of Quantitative Linguistics* 1: 16 – 34.

Baayen R H 2001 *Word Frequency Distributions*. Dordrecht, Kluwer Academic Publishers.

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

Baayen R H, Neijt A 1997 Productivity in context: a case study of a Dutch suffix. *Linguistics* 35: 565 – 587.

Lezius W, Rapp R, Wettler M 1998 A freely available morphological analyser, disambiguator, and context sensitive lemmatizer for German. In *Proceedings of the COLING-ACL*, pp 743 – 747.

[6] Work leading to such a morphology system is currently under way in the DeKo project (Schmid et al 2001).

Lüdeling A, Evert S, Heid U 2000 On measuring morphological productivity. In *Proceedings of the KONVENS 2000*, Ilmenau, pp. 57 – 61.

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

Schiller A 1996 Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO. In Hausser R (ed), *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. Tübingen, Niemeyer

Schmid T, Lüdeling A, Säuberlich B, Heid U, Möbius B 2001 DeKo: Ein System zur Analyse komplexer Wörter. In: *Proceedings der GlDV*, Giessen

Schultink H 1961 Produktiviteit als morphologisch fenomeen. *Forum der Letteren*: 110 – 125.

van Marle J 1985 *On the Paradigmatic Dimension of Morphological Productivity*. Dordrecht, Foris.