Testing the extrapolation quality of word frequency models

Stefan Evert and Marco Baroni
Cognitive Science Department / SSLMIT
University of Osnabrück / University of Bologna
stefan.evert@uos.de / baroni@sslmit.unibo.it

1 Introduction

Many studies in corpus linguistics and related disciplines aim to determine the characteristic aspects of a language, a particular genre, a group of speakers, an individual speaker or a linguistic process. In order to do so, they compute certain numerical quantities from an available text sample (i.e., a corpus) and extrapolate them to the full language (genre, speaker, process, etc.), or at least to much larger samples. For example, a corpus linguist might use the Brown and LOB corpora in this way to draw inferences about the differences between American and British English; a stylometrist might count the different words in the Shakespeare canon in order to estimate the richness of his vocabulary; and a morphologist might try to determine whether a certain word formation process is more productive than another by comparing the number of nonce words formed by each of the processes.

Some of the numerical quantities used by such studies tend to be very stable across all but the smallest sample sizes, so that they can be reliably estimated from any given corpus (typical examples are average word and sentence length). Other quantities, however, change systematically with the sample size even for large samples such as the 100 million words of the British National Corpus. In order to compare samples of different sizes or to make generalizations about the full language, it is thus necessary to extrapolate their observed values to much larger samples. In this study, we focus on two quantities that play a central role in the examples listed above: the number of distinct word *types* (the vocabulary size V) and the number of *hapax legomena* (types occurring just once, V_1) in a sample of N word *tokens*. The vocabulary size V and the type-token ratio V/N are often used to measure vocabulary richness and lexical variety in stylometry, authorship attribution, language acquisition and similar fields (see, e.g., Youmans 1990; Chipere *et al.* 2004). In morphology, an established measure for the productivity of a word formation process is Baayen's productivity index $p = V_1/N$ (Baayen 1991), which is based on the number of hapax legomena.

The most sophisticated techniques for predicting vocabulary growth (i.e., the development of V and V_1 for increasing sample size N) rely on statistical models of the distribution of word frequencies (Baayen 2001). Once the parameters of these models have been estimated from the observed text sample, they can be used to extrapolate V and V_1 to arbitrary sample sizes. Following Baayen's terminology, we refer to these models as LNRE models, which stands for Large Number of Rare Events. Recently, two open-source toolkits have become available that implement a number of LNRE models: LEXSTATS (http://www.mpi.nl/world/persons/private/baayen/software.html) and the UCS toolkit (http://collocations.sf.net/). It is natural to ask whether these toolkits can be used by researchers as off-the-shelf implementations for extrapolating V and V_1, and more generally, how useful and accurate the underlying statistical models are.

In the original publications, LNRE models are evaluated by computing their

goodness-of-fit with respect to the observed distribution of word frequencies (Baayen 2001:118ff) or by comparing the predicted vocabulary sizes for *smaller* samples (than the one used for parameter estimation) to the observed growth curve (e.g., Baayen 2001:88). Other researchers have based their assessment on how well predictions made by the models agree with human intuitions (e.g., Lüdeling and Evert 2003). Surprisingly, to our knowledge no direct evaluation of the extrapolation quality of different LNRE models has ever been carried out, even though this can be done in a straightforward and intuitive way. The purpose of the present paper is to fill this gap at least partially, with a comparison and discussion of the LEXSTATS and UCS models applied to data sets of various kinds.

In our experiments, we estimate the model parameters on a subset of the respective corpus or data set comprising only the first N_0 tokens (e.g., the first 50%, 25% or 10%). This allows us to compare the extrapolated vocabulary growth for each model directly to the true vocabulary growth up to the full corpus size of N_F. For instance, using the first 25% of the data for parameter estimation (i.e., $N_0=N_F/4$) we can evaluate extrapolation quality up to 4 times the estimation size N_0. We apply the LNRE models in a number of plausible real-life settings, ranging from small manually cleaned data sets to large text collections with fully automatic processing. The present paper presents a qualitative evaluation based on visual inspection of the predicted and observed growth curves for V and V_1. The empirical inadequacy of current LNRE models becomes clear enough from these plots, so that a more detailed quantitative evaluation is not called for at this time.

The rest of the paper is structured as follows: In Section 2, we briefly describe the theoretical foundations of LNRE models and introduce the specific models implemented by the LEXSTATS and UCS toolkits. In Section 3, we explain the experiments that we have performed and describe the data sets we have used. Sections 4, 5 and 6 present the results of these experiments. Finally, Section 7 summarizes our conclusions and makes some suggestions for further research.

2 Statistical models of word frequency distributions

The standard approach to modelling vocabulary growth is to assume that the observed corpus or data set is a random sample from a population of word types with associated occurrence probabilities. If the distribution of these probabilities is known, the expected vocabulary size V and number of hapax legomena V_1 can be calculated for any sample size N (Baayen 2001:41–51). A (parametric) LNRE model is then simply a formula that describes the distribution of occurrence probabilities in the population. Most LNRE models have between one and three parameters that determine the precise shape of this distribution. The process of adjusting these parameters to match the distribution of word frequencies in the observed corpus is referred to as parameter estimation. It is usually followed by an evaluation of the model's goodness-of-fit, which measures how well the shape of the probability distribution could be brought into agreement with the observed data (see Baayen 2001:118–122 for details). While goodness-of-fit is important as an indicator of the appropriateness of a given LNRE model, it is not necessarily a measure of extrapolation quality (although it is tempting to assume a close relationship between the two).

The most well-known LNRE model is Zipf's law (Zipf 1949), which stipulates that the probability p_n of the a word type w is inversely proportional to its Zipf rank n, i.e., the rank of w in a list of all word types ordered by decreasing frequency. An extension

of this model leads to a general power-law relationship expressed by the equation $p_n = C \cdot n^{-\alpha}$, with α ranging from 1 to 2 and a C normalizing constant that ensures that probabilities add up to one. A further extension, the Zipf-Mandelbrot law, is implemented by the ZM and fZM models that are part of the UCS toolkit (Evert 2004). While the ZM model assumes a population with an infinite number of types, the fZM introduces population diversity (i.e., the finite number of types) as an additional parameter of the model.

An intuitively unrealistic aspect of the fZM model (which it shares with the original form of Zipf's law) is that it simply cuts off the power law after a certain number of types. This implies a probability threshold for words, with many types having occurrence probabilities slightly above the threshold but none at all below. The GIGP model (see Baayen 2001:89–93) is a more sophisticated implementation of the Zipf-Mandelbrot law, which adjusts the probabilities of the lowest-frequency types in order to obtain a smoother distribution. In addition to GIGP, the LEXSTATS software also implements a model based on a lognormal distribution of the type probabilities (see Baayen 2001:82–88), as well as several other models that are not LNRE models in a strict sense (Baayen 2001:94–118). These models, most notably the Yule-Simon law, do not specify a probability distribution for the population, but user other techniques to extrapolate V and V_1 from the observed data.

In a comparative evaluation of the latter models based solely on their goodness-of-fit, Baayen (2001:124–131) notes that the lognormal model gives good results for small data sets, while GIGP is superior for larger sample sizes (up to 6 million tokens). Overall, the Yule-Simon model seems to achieve the best goodness-of-fit independent of corpus size. Evert (2004) evaluates the ZM and fZM models in a similar manner on larger data sets (up to 48 million tokens). His findings indicate that the fZM model is far superior to ZM, and similar in quality to the best models implemented in the LEXSTATS package. It has to be emphasized once again that these results *per se* tell us nothing about the extrapolation quality of the models.

3 Experiments

3.1 Data sets used for the comparison

In our experiments, we used the following corpora and data sets, which represent a number of typical situations that a researcher is likely to encounter in real-life studies:

- **The British National Corpus (BNC)**: a balanced corpus of approximately 100 million words of written and spoken British English from the years 1975–1994 (Aston and Burnard 1998).
- **The Lancaster-Oslo/Bergen corpus (LOB)**: a balanced corpus of approximately 1 million words of written British English from 1960, designed as an analogue to the Brown corpus (Johansson et al. 1978).
- **Süddeutsche Zeitung (SZ)**: a collection of German newspaper articles from several volumes of the *Süddeutsche Zeitung*, with a total of more than 250 million words.
- **The JP Web corpus**: approximately 3.5 million words of text obtained by crawling Japanese Web pages (Ueyama and Baroni 2005).
- **German suffix data**: manually corrected occurrences of German adjectives formed by the suffixes *–bar* and *–lich* in a collection of German newspaper articles from two volumes of the *Stuttgarter Zeitung* (Evert and Lüdeling 2001).

Except for the German suffix data, we computed frequency distributions and vocabulary growth for all word forms in the corpora. The tokens in the BNC, LOB and SZ corpora were normalized to lowercase, and simple regular expressions were applied to filter out numbers, punctuation marks and other non-linguistic "junk" (we also ran preliminary experiments with a lemmatized version of the BNC but did not further pursue this avenue because the results did not differ considerably from those obtained for word forms). The Japanese corpus was tokenized with ChaSen (Matsumoto et al. 2000) and all words containing non-Japanese characters were discarded.

The four corpora and the suffix data sets were chosen in order to illustrate a number of typical scenarios arising in corpus-based studies: small and restrictive lists of manually cleaned data (the suffix data sets); a relatively small and clean sample of language data (the LOB); a larger and more varied balanced corpus (BNC); and a very large and noisy collection of texts from a single source (SZ). We also conducted some experiments with a 380 million word collection of Italian newspaper articles, which gave results that are similar to those reported for the BNC but caused considerable computational problems for the LEXSTATS tools. Finally, the JP Web corpus illustrates both the increasingly common usage of the Web as a source of corpus data and, more importantly, how the models handle a language that is typologically very different from the Indo-European family. Table 1 lists the sample size N_F and vocabulary size V_F for each of the data sets (notice that these are sizes computed *after* data-cleaning, and thus in some cases they are considerably smaller than those reported above).

Data set	Sample size N	Vocabulary size V
Suffix *–bar*	36,164	544
Suffix *–lich*	278,364	3,120
LOB	994,469	44,485
BNC	96,903,342	487,221
SZ	226,147,264	2,584,543
JP Web	2,175,736	137,060

Table 1. Descriptive statistics (sample size and vocabulary size) for the data sets used in the experiments.

3.2 Experimental procedure

We applied all LNRE models described in Section 2 to the data sets listed above, using the implementations provided by the LEXSTATS and UCS packages. In general, we relied on the default settings for parameter estimation. Although goodness-of-fit can sometimes be improved considerably by hand-tuning this procedure, most researchers will lack the necessary experience and are more likely to use the packages as off-the-shelf solutions. The only changes we made are to use the cost function C_2 rather than the default function C_1 for parameter estimation of the LEXSTATS models (Baayen 2001:123), wherever this was appropriate and computationally feasible. Our rationale for this is that C_2 uses more information from the training data, so we expect it to result in better fits than the default C_1 estimation scheme, which relies solely on the three quantities N, V and V_1. For this same reason,

however, C_2 is more likely to break down than C_1, especially on the larger data sets. In addition, we initialized the parameter estimation procedure for the lognormal model with hand-tuned values since the default settings led to very poor goodness-of-fit.

For each model and each data set, we ran experiments with five different estimation sizes N_0, using the first 75%, 50%, 25%, 10% and 1% of the available data, respectively. The results of these experiments are reported in Section 4. In most cases, we focus on the runs with $N_0=N/2$ (50%) and $N_0=N/4$ (25%), which are the most interesting situations. Extrapolation quality rapidly degrades when less than 25% of the data are used for estimation. In Section 5, we repeat all experiments on randomized versions of the data sets in order to study the influence that violations of the randomness assumption underlying all LNRE models have on the predicted vocabulary growth. Finally, we use the models to extrapolate the number of hapaxes V_1 in Section 6, which is particularly relevant for studies of (morphological) productivity.

Although we performed the experiments with all available LNRE models, we only report the results of GIGP, lognormal, ZM and fZM for reasons of clarity and space. The LEXSTATS package implements two versions of the GIGP model, of which we used the more general one. The Yule-Simon model performed considerably worse than lognormal and GIGP (with a few exceptions). Another model from the same family (referred to as the Zipf model in LEXSTATS) showed equally bad performance and its implementation broke down for the larger sample sizes. The reader interested in these models is welcome to ask us for the relevant data sets and results.

4 Results for the extrapolation of V

4.1 All word forms (BNC, LOB, SZ, JP Web)

The results of our experiments are presented in the form of intuitively understandable vocabulary growth curves, i.e., plots of vocabulary size V against sample size N as illustrated in Figure 1 for the BNC data. For the panel in the top row, the first 25% of the 96 million BNC tokens were used for parameter estimation (i.e., N_0 = 24 million), indicated by a vertical line in the plots. Note that the numbers on the x-axis refer to millions of tokens (M) – or thousands of tokens (k) in some of the later plots – and those on the y-axis to thousands of types (k). This experiment simulates a situation where vocabulary growth has to be extrapolated from an observed corpus of 24 million tokens. The part of the growth curves to the *right* of the vertical line represent the predictions of the various LNRE models up to 4 times the estimation size N_0 (also referred to as the *expected* growth curves) and compare them with the actual vocabulary growth through the remaining 72 million tokens of the BNC (the *observed* growth curve, labelled “Corpus” in the plots). The part of the curves to the *left* of the vertical line represents the information that would be available to the researcher for assessing the quality of an LNRE model, either by calculating goodness-of-fit or by visual comparison of the interpolated vocabulary growth according to the model with the observed growth curve up to N_0 (cf. Section 4.3).

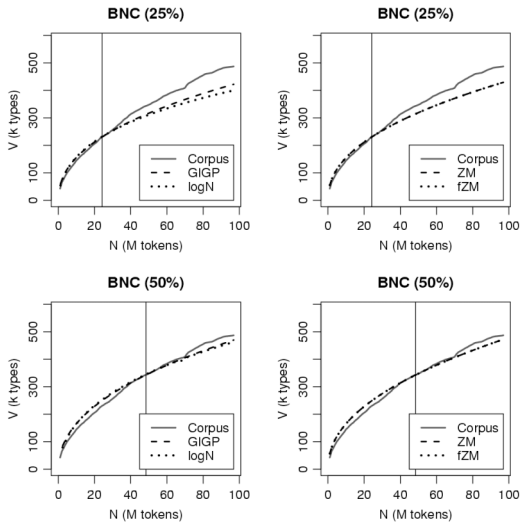

Figure 1. Extrapolated vocabulary growth curves for the BNC data.

Taking a closer look at Figure 1, we see that for extrapolation from 25% of the data, all models underestimate the true vocabulary growth substantially. While GIGP, ZM and fZM are virtually indistinguishable, lognormal gives slightly worse results than the other three models. Naturally, the accuracy of the prediction degrades with greater extrapolation distance. When the models are estimated on 50% of the data (bottom row), the performance of the models improves considerably, although they still do not coincide fully with the true vocabulary growth curve in the last 25% of the corpus. Here, the observed curve shows a distinct irregularity in the form of a hump, i.e. faster vocabulary growth than in previous parts of the corpus. This phenomenon is explained by the fact that the spoken parts of the BNC are concentrated at the end of the corpus, where suddenly new words from spoken English are injected into the vocabulary. It is hardly surprising that the LNRE models, which were trained only on written English, are unable to predict the vocabulary growth of spoken English accurately. A similar hump in the observed growth curve appears after the first 25% of the corpus, which explains why the models in the top row perform considerably worse than those in the bottom row, even when we consider only extrapolation up to $2N_0$. In Section 5, we will take a closer look at the extent to which the non-random arrangement of texts in a corpus (and of words within the texts) affects the LNRE models.

Of course, the extrapolated growth curves show the *expected* value of V according to each model, which can be interpreted as average values across many different random samples. Part of the discrepancy between them and the observed curve may thus be explained by random variation in the one particular language sample (namely the BNC) that we used as a basis for our experiments. While it should be immediately obvious from the top row of Figure 1 that we are dealing with systematic underestimation rather than random effects, we have also calculated confidence intervals for V based on the standard deviation predicted by the LNRE models. Due to the large sample sizes of millions of tokens that we are dealing with, these confidence intervals are so small as to be indistinguishable from the expected growth curves, even at a confidence level of 99.9%. Consequently, we have omitted them from the plots and note that extrapolation errors can by no means be attributed to random effects.

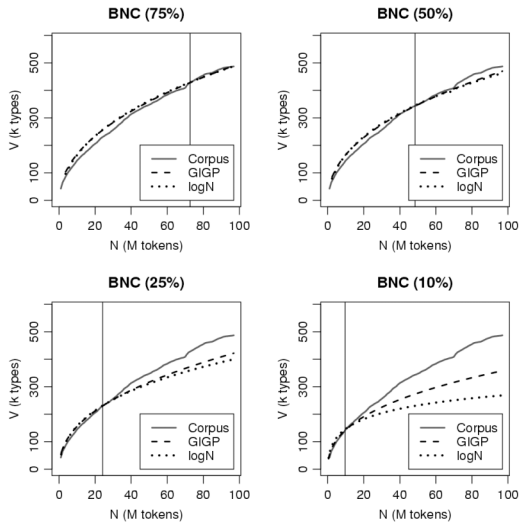

Figure 2. Extrapolated vocabulary growth curves for the BNC data, with estimation sizes ranging from 10% to 75%.

In Figure 1 and most of the rest of the paper, we only present results for LNRE models estimated from 25% and 50% of the available data, respectively. We chose these two settings since they give a good impression of the general trends, and quite often the step from 25% to 50% constitutes the "boundary" between extrapolation results that are clearly off-the-mark and results that are at least qualitatively plausible. It is hardly surprising that extrapolation from a 10% sample, shown in the bottom right panel of Figure 2, confirms the negative trend and magnifies the differences between the curves (note that the curves for ZM and fZM, which are similar to those of the GIGP model, were omitted from the plots to save space). At 10 times the estimation size, the extrapolated vocabulary growth curves have little in common with the true growth curve any more, especially for the lognormal model. In the opposite direction, extrapolation from 75% of the data shows very good agreement with the observed curve, but is of little practical relevance (because extrapolation goes only from 72.6 million tokens to 96.9 million tokens, i.e. 33% beyond the estimation size).

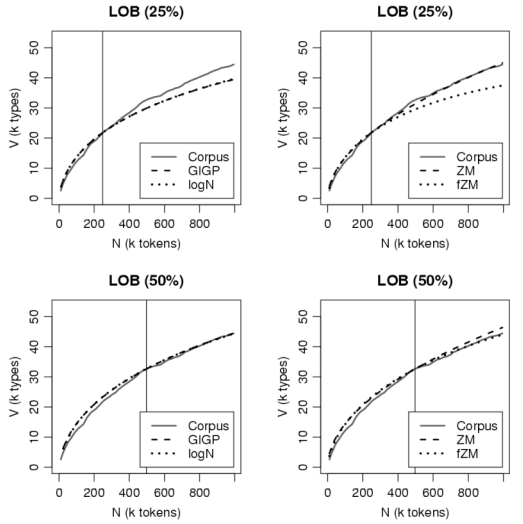

Figure 3. Extrapolated vocabulary growth curves for the LOB data.

Figure 3 presents extrapolation results for the LOB corpus, which is much smaller than the BNC and more homogeneous (because it does not contain spoken English).

Despite these differences, the overall trends are similar to those encountered for the BNC, indicating that the accuracy of the model predictions depends more on how far the growth curves are extrapolated beyond the estimation size than on the absolute amount of training data used. There are two interesting aspects of the LOB data: the GIGP and lognormal models are practically indistinguishable, and the ZM-based extrapolation is spot on for an estimation size of 25%, where it is also much better than the fZM model. This observation is quite surprising, considering that Evert (2004) reports far superior goodness-of-fit for the fZM model (which is an extension of ZM and should thus always perform equal to or better than the latter).

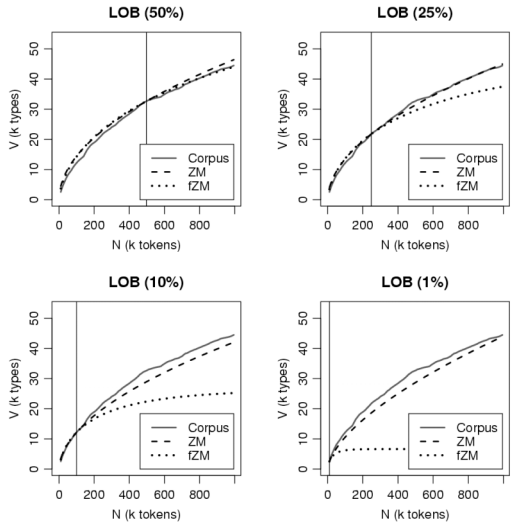

Figure 4. Extrapolated vocabulary growth curves for the LOB data, with estimation sizes ranging from 1% to 50%.

In order to find out whether this trend continues for even smaller estimation sizes, the plots in Figure 4 compare the ZM and fZM models estimated on as little as 1% of the full corpus data (i.e., only 10,000 tokens of text). Astonishingly, the ZM model still achieves an excellent prediction of the true vocabulary growth, while the fZM quality degrades quickly. In the bottom right panel, the fZM model even predicts that the total vocabulary of written British English contains less than 10,000 word form types. Since the surprising accuracy of the ZM model cannot be replicated in the experiments with other corpora, it has to be regarded as a “freak accident” for this particular data set. In Section 5, we will have a look at the role that the non-random order of texts in the LOB plays in this context (e.g., we gather from the manual that all the humour appears at the end of the corpus).

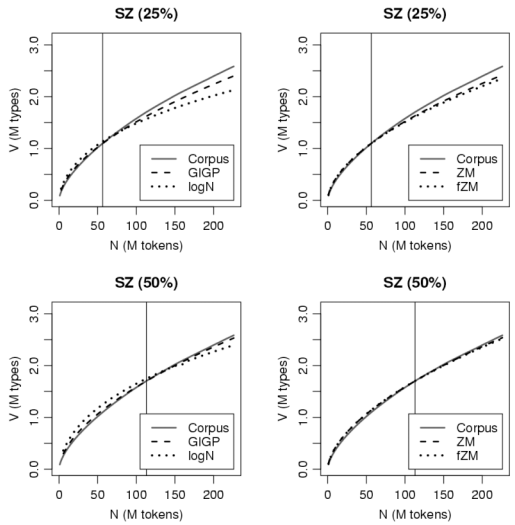

Figure 5. Extrapolated vocabulary growth curves for the SZ data.

The SZ corpus can be seen as the opposite of the LOB in terms of size (more than 200 million tokens), cleanness (fully automatic processing with off-the-shelf tools), and diversity (articles from a single newspaper only). From the plots in Figure 5 we see that this kind of "big and noisy" data set is much more amenable to LNRE modelling than a balanced language sample such as the LOB or BNC. One possible reason for this result is that in the newspaper data, different text types and genres are mixed (the articles being ordered by publication date) rather than lumped together at the beginning or end of the corpus (as it was the case with BNC and LOB). This explanation is supported by the much smoother shape of the observed vocabulary growth curves for the SZ data. Apart from the better overall results, the general trends that we have encountered in the BNC are confirmed: GIGP, ZM and fZM achieve very similar extrapolation quality (recall that all three are based on the Zipf-Mandelbrot law), and are considerably better than the lognormal model. All models have a tendency to underestimate the true vocabulary growth, which becomes more pronounced for smaller estimation sizes (as do the differences between individual models).

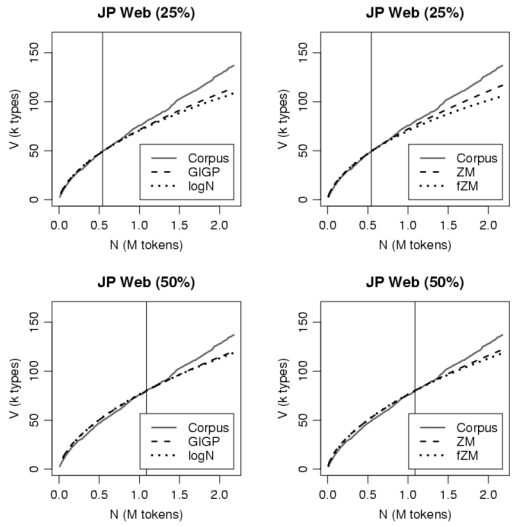

Figure 6. Extrapolated vocabulary growth curves for the JP Web data.

The good results obtained on the relatively noisy SZ corpus are encouraging for an

application of the models to language data collected from the Web, such as the JP Web corpus. However, the observed vocabulary growth curve of this corpus has a very unusual shape that is almost a straight line (see Figure 6). After a short initial phase (comprising roughly the first 500,000 tokens), the vocabulary continues to grow at a constant rate (while all the growth curve plots in Baayen (2001) and similar publications show a concave shape with decreasing slope). All four LNRE models underestimate this untypical growth curve considerably, with fZM delivering the worst performance. It is not clear yet whether this phenomenon is a consequence of the typological differences between Japanese and the European languages, or whether it has to do with the fact that word boundaries are not marked orthographically in Japanese and are often difficult to establish by other means (cf. Ha *et al.* 2002). For the JP Web corpus, word units were determined in a fully automatic way by the ChaSen tokenizer.

4.2 Suffix data sets (*–bar*, *–lich*)

We turn now to the two German suffix data sets of Evert and Lüdeling (2001). Unlike the previous examples where the goal was to model the frequency distribution of all word forms in a corpus, these data sets were created for a much more narrowly defined purpose, namely measuring and comparing the productivity of the German adjective-forming suffixes *–bar* and *–lich*. Consequently, they are much smaller than the data sets in Section 4.1 and have undergone extensive manual corrections, so that the remaining data represent a specific linguistic phenomenon (a word-formation process) rather than a broad mixture of factors that contribute to the overall vocabulary growth in a corpus.

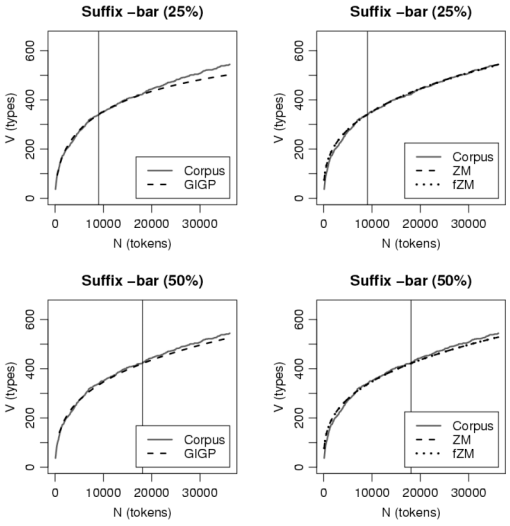

Figure 7. Extrapolated vocabulary growth curves for the *–bar* suffix data.

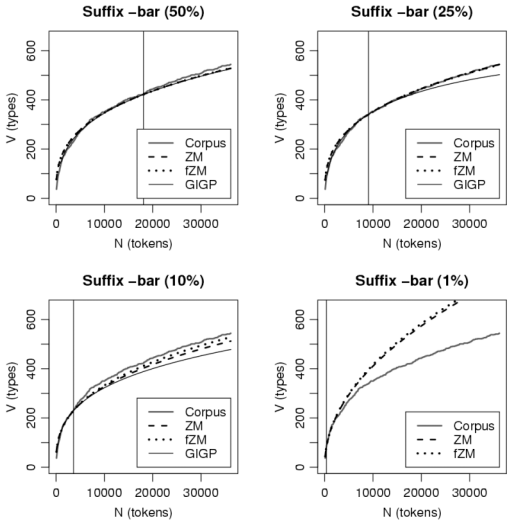

Figure 8. Extrapolated vocabulary growth curves for the *–bar* suffix data, with estimation sizes ranging from 1% to 50%.

Figure 7 presents the extrapolation results for the *–bar* data set. Note the absence of the lognormal model from the left column, which is due to the failure of the estimation procedure implemented for this model in the LEXSTATS package. The other three models achieve very good extrapolation accuracy, with ZM and fZM being spot on for an estimation size of 25%. These results are very encouraging, since one of the main goals of any quantitative study of morphological productivity is to predict the rate at which new words are created by a word-formation process.

Figure 8 shows that even when extrapolating to 10 times the estimation size, the accuracy of the predictions degrades gracefully (bottom left panel), allowing researchers to compare the productivity of morphological processes even when the corresponding sample sizes differ substantially. The drastic overestimation seen in the bottom right panel is hardly surprising, since the parameter estimates are based on a sample containing only 360 tokens in this case.

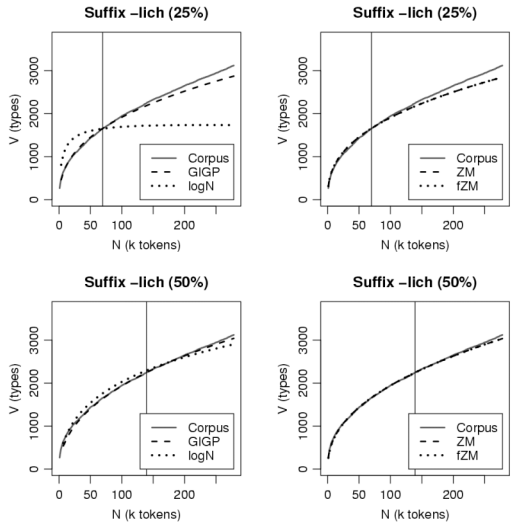

Figure 9. Extrapolated vocabulary growth curves for the *–lich* suffix data.

Unfortunately, Figure 9 reveals that the high extrapolation quality for *–bar* is not shared by all morphological processes. When the models are applied to the *–lich* suffix data, we see the familiar underestimation pattern. Using any of the models to extrapolate beyond $4N_0$ would be highly dangerous. While the lognormal estimation does not fail in this case, it is completely off the mark for an estimation size of 25%. A blind use of this model, perhaps inspired by the good results that Baayen (2001:126) has obtained for data sets of similar size, would lead to the erroneous and counter-intuitive conclusion that the suffix *–lich* is completely unproductive.

4.3 Predicting extrapolation quality

Although our experiments show a common pattern of underestimation of the vocabulary growth and decreasing accuracy, depending on how far beyond the estimation size extrapolation is carried out, there are considerable differences between the four LNRE models and between the data sets. For instance, while the ZM model can be used to extrapolate from a small subset of the LOB to the full corpus (and presumably beyond), the other models perform considerably worse and would make entirely misleading predictions when estimated on 10% of the data or less. None of the models is able to predict the vocabulary growth of the BNC accurately. While these differences are obvious in our experiments, any real-life application will use all the available corpus data for parameter estimation (N_0=N_F) and extrapolate to even larger values of N. Since there is no empirical growth curve to which the model predictions could be compared, we have no way of knowing whether we can trust our LNRE model or whether we have chosen an infelicitous combination of model and data set. If we want to draw meaningful conclusions, we need a means for predicting extrapolation quality based on the data available for parameter estimation.

The most straightforward solution is to compute the goodness-of-fit of a LNRE model after parameter estimation, i.e. how well the model matches the observed frequency distribution. Using a multivariate chi-squared test (Baayen 2001:118–122), we can compute a X^2 value as a measure of goodness-of-fit, with smaller values indicating better agreement between expected and observed frequency distribution. It is natural to assume that better goodness-of-fit should lead to better extrapolation quality, so that we should use the model with the smallest X^2 value for the extrapolation. However, for the LOB experiment with an estimation size of 10%, we obtain $X^2=6896.6$ (df=14) for the ZM model compared to $X^2=472.8$ (df=13) for the fZM model: even though the fZM has much better goodness-of-fit than the ZM model, its vocabulary growth extrapolation is fundamentally misleading while that of the ZM model is very accurate (cf. Figure 4). These findings are corroborated by the even better goodness-of-fit of the lognormal model ($X^2=98.4$, df=14), whose extrapolation quality is similar to that of fZM (not shown in the plots). It is thus obvious that we cannot use goodness-of-fit as a predictor of extrapolation accuracy.

An intuitively appealing alternative is to look at the interpolated vocabulary growth curves (i.e., the part of the curves to the left of the vertical line in our experiments) and compare them with the observed vocabulary growth up to the estimation size N_0. Since the parameter estimation ensures that the models predict the correct vocabulary size V at N_0, underestimation of V in the extrapolation should go hand in hand with overestimation in the interpolated part of the curve. This pattern is clearly visible for the BNC data in the top row of Figure 1. However, Figure 4 paints a different picture. While the interpolated curves of the ZM and fZM models are practically indistinguishable, their predictions for larger values of N are totally different. Moreover, at an estimation size of 25%, the ZM model overestimates vocabulary growth in the interpolated part, but its extrapolation is highly accurate nonetheless. Despite such counterexamples, this approach is much more promising as a predictor of extrapolation accuracy than goodness-of-fit.

5 Results for randomized data sets

As has been pointed out in Section 2, all the LNRE models in our evaluation assume that the observed corpus is a random sample from a population of word types. Of course, this assumption is unrealistic for natural language and the causes and consequences of non-randomness are discussed at length in the technical literature on LNRE models (see Baayen 2001:161–173). In order to assess the influence of non-randomness on the extrapolation quality of our models, we have repeated the experiments reported in Section 4 on randomized versions of the data sets (i.e., the tokens have been re-arranged in random order for each data set).

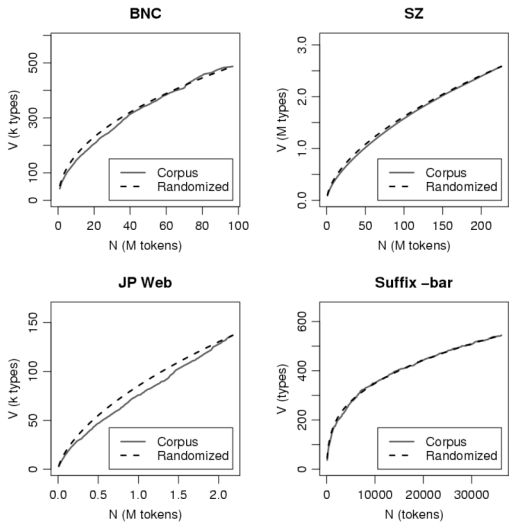

Figure 10. Comparison of vocabulary growth in randomized and non-randomized versions of the BNC, SZ, JP Web and *–bar* data sets.

Figure 10 compares the original and randomized versions of some of the data sets, giving a first indication to what extent non-randomness affects the growth curves. The BNC and JP Web data sets, both of which were problematic for the LNRE models, show strong non-randomness effects. The much smoother growth curves of the SZ corpus differs only slightly from the randomized version, and for the *–bar* suffix data there is no visible difference (that cannot be explained by random variation). A possible explanation for the high degree of randomness in the *–bar* data can be found in the relatively small number of tokens and their "sparse" distribution across the corpus, where they are unlikely to lump together in a single article. Figure 10 indicates that violations of the randomness assumption may well be the cause for underestimation on the BNC and JP Web corpora, whereas the other two data sets should not be affected. To test this hypothesis, the figures below compare extrapolation performance on the randomized and original versions of the four data sets, at an estimation size of 25%.

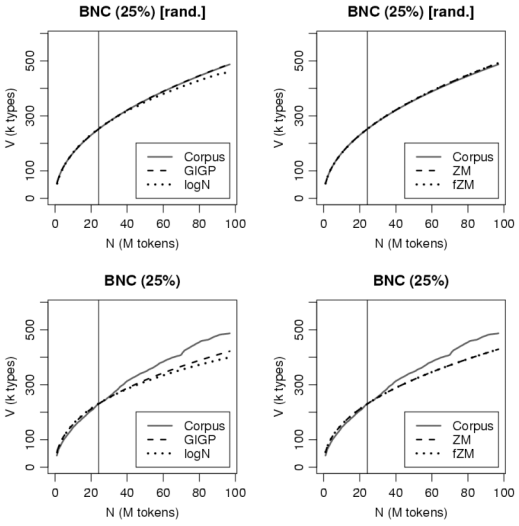

Figure 11. Comparison of extrapolation accuracy for randomized and non-randomized versions of the BNC corpus.

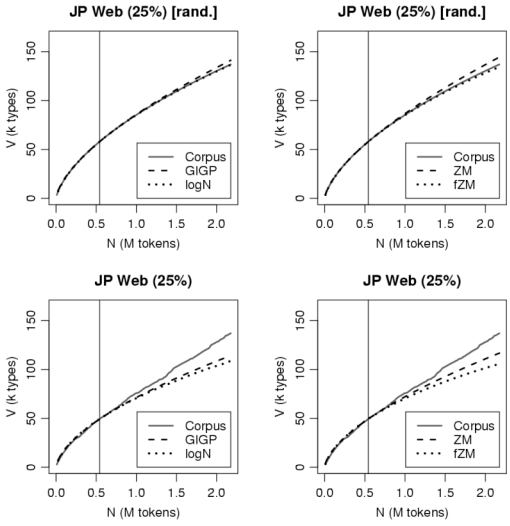

Figure 12. Comparison of extrapolation accuracy for randomized and non-randomized versions of the JP Web corpus.

Figure 11 and Figure 12 confirm our expectations. On the randomized version of the BNC, all models except lognormal show excellent extrapolation quality. For the JP Web data, the results are also very good. Interestingly, the fZM model now performs much better than the ZM model, in line with its better goodness-of-fit ($X^2=167.9$, df=13 vs. $X^2=5374.0$, df=14).

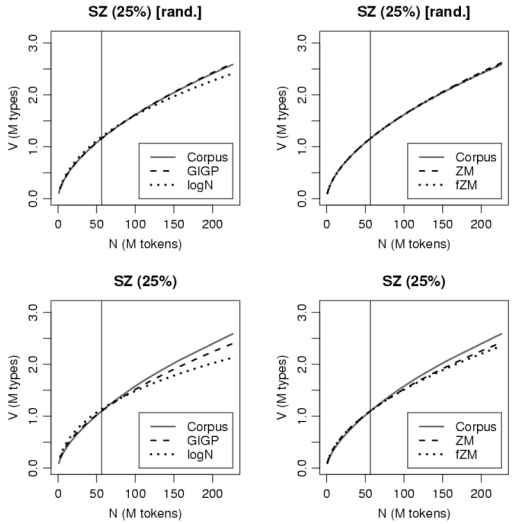

Figure 13. Comparison of extrapolation accuracy for randomized and non-randomized versions of the SZ corpus.

Although Figure 10 showed hardly any visible differences between the randomized and the original vocabulary growth curve of the SZ corpus, suggesting that there are no serious violations of the randomness assumption, all models achieve much better extrapolation performance on the randomized version of the corpus (Figure 13). Again all models except lognormal give an excellent prediction of the true vocabulary growth.

Our preliminary conclusion at this point is that a substantial (if not the largest) part of the extrapolation problems reported in Section 4 can be attributed to non-randomness of the corpus data. On randomized data sets, most LNRE models give an excellent approximation to the true vocabulary growth up to $4N_0$. Moreover, the better extrapolation quality of the ZM model compared to fZM seems to stem from a greater robustness against non-randomness. On the randomized data, fZM is as good as or even better than ZM. Further evidence comes from the LOB corpus (see Figure 4, randomized version not shown here). The astonishing accuracy of the ZM model on the original corpus gives way to serious overestimation when the data are randomized, confirming our assumption that it was a "freak accident". The fZM model performs distinctly better than ZM now, and the GIGP model achieves excellent results.

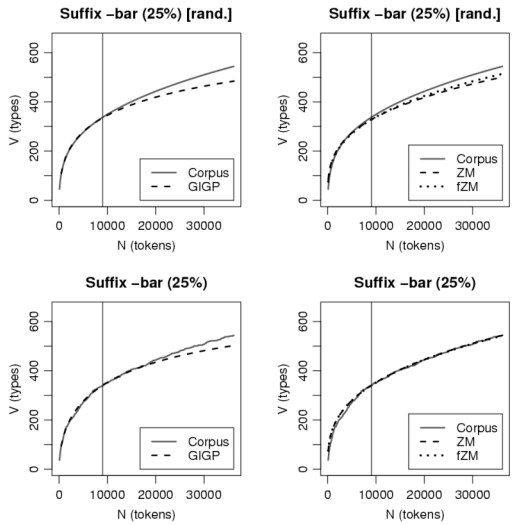

Figure 14. Comparison of extrapolation accuracy for randomized and non-randomized versions of the *–bar* suffix data.

Finally, Figure 14 presents the results for the randomized *–bar* suffix data set. Since there was no visible difference between the original and randomized vocabulary growth curves and since the ZM and fZM models achieved excellent extrapolation quality, we would not expect randomization to have substantial effects. Surprisingly, though, the extrapolation accuracy deteriorates for all three LNRE models (recall that the estimation procedure for the lognormal model failed on this data set). This as yet unexplained result demonstrates clearly that not all extrapolation problems can be attributed to non-randomness in the data, which sometimes seems to counterbalance other inadequacies of the LNRE models.

6 Results for the extrapolation of V_1

So far, our experiments have concentrated on the development of the vocabulary size V for increasing sample size N. In addition to V, the number V_1 of hapaxes also plays an important role for studies of morphological productivity and vocabulary richness (e.g., Baayen's (1991) productivity index $p = V_1/N$). In this section, we present results on the accuracy of the extrapolation of V_1 for selected data sets.

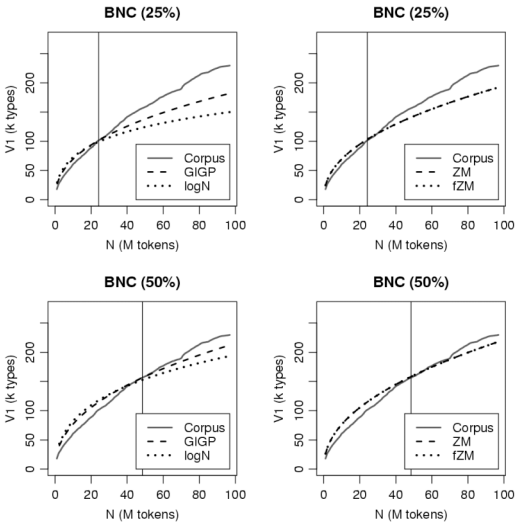

Figure 15. Extrapolation of the number of hapaxes for the BNC data.

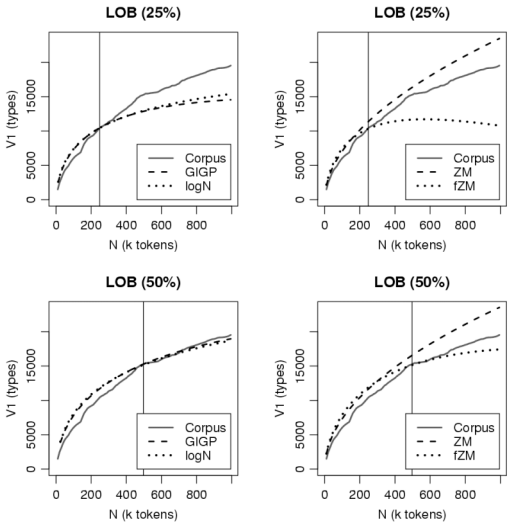

Figure 16. Extrapolation of the number of hapaxes for the LOB data.

Figure 15 and Figure 16 show that extrapolation quality is considerably lower for the number of hapaxes V_1 than for the vocabulary size V. For the LOB corpus with an estimation size of 25%, the extrapolated curve according to the fZM model has a maximum at around 500,000 tokens and begins to fall afterwards, (wrongly) indicating that the corpus leaves the LNRE zone (Baayen 2001:55) where the distribution of word frequencies follows Zipf's law.

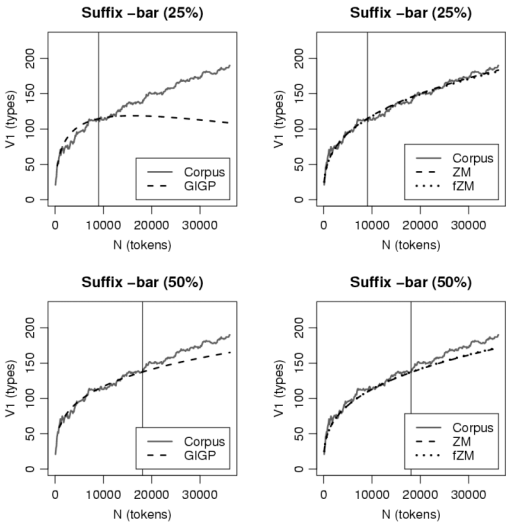

Figure 17. Extrapolation of the number of hapaxes for the *–bar* suffix data.

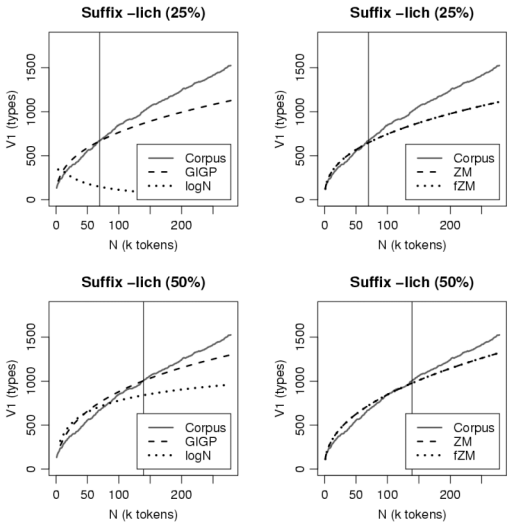

Figure 18. Extrapolation of the number of hapaxes for the *–lich* suffix data.

Since the extrapolation of hapax counts is of particular importance for morphological productivity, we show the results for both German suffix data sets in Figure 17 and Figure 18. While the predictions made by the ZM and fZM models are fairly accurate for the *–bar* data, they are at best qualitatively acceptable as an indication of the overall trend (suggesting a productive process) for the *–lich* data. However, the most worrying aspect of these experiments is that extrapolation quality seems to break down suddenly when the estimation size becomes too small (which is often the case for highly restricted data such as morphological affixes). This happens to the GIGP model on the *–bar* data (at $N_0=9{,}050$; top left panel in Figure 17) and to the

lognormal model on the *–lich* data (at N_0=*69,600*; top left panel in Figure 18), which suddenly indicate an unproductive process or an unrealistically low degree of productivity. Without evidence from further experiments, we have to assume that this is not a particular weakness of the GIGP and lognormal models, but that we would find the same behaviour with ZM and fZM on other morphological data sets (especially considering the fact that fZM is mathematically very similar to the GIGP model). Consequently, at the current time LNRE models seem to be entirely unusable for the extrapolation of quantitative measures of morphological productivity that are based on the number of hapaxes.

7 Conclusions and directions for further work

In this paper, we have tested the extrapolation accuracy of four widely-used LNRE models empirically, by estimating their parameters on a subset of a given corpus and comparing the true vocabulary growth for the rest of the corpus with the growth curves predicted by the models. Using six different data sets that cover a range of typical real-life situations, our overall conclusion is that most of the models (with the exception of lognormal) provide a plausible extrapolation of the vocabulary size V up to 2 times the size N_0 of the estimation corpus, and in many cases they at least capture the right trend up to 4 times N_0. The best results were often achieved by the GIGP and ZM models. Extrapolation accuracy is much lower for the number of hapaxes V_1, rendering the current models unusable for studies of morphological productivity.

While LNRE models are a powerful and important tool for quantitative studies related to the distribution of word frequencies, vocabulary diversity and morphological productivity, it is obvious that improvements over the current state of the art are urgently needed. Our experiments with randomized data sets in Section 5 suggest that the inaccuracy of the extrapolated growth curves is to a large part caused by non-random ordering of tokens in the corpus (but not all errors of the models can be blamed on non-randomness). Of course, randomization is not an option in real studies, where all the available corpus data are used for parameter estimation. Therefore, the logical next step in LNRE research is to develop models that can detect, and correct for non-randomness in the training data. Baayen (2001) suggests a range of parameter-adjusted models and mixture models for this purpose, which are also implemented in the LEXSTATS package. However, these extensions add to the general slowness (in some of our experiments, parameter estimation and extrapolation took several hours to complete) and frailty of the LEXSTATS models and depend crucially on hand-tuned parameter values. Similar extensions to the faster and more robust UCS models will hopefully turn out to be a viable alternative.

In order to support theoretical and practical improvements of the LNRE models, further evaluation experiments will play a key role. We plan to extend our randomization experiments to the extrapolation of V_1, as well as to consider a broader range of data sets and LNRE models in the evaluation (especially versions of the LEXSTATS models that adjust for non-randomness). Balanced corpora such as the BNC and LOB allow us to determine whether the relevant non-randomness effects depend on the arrangement of documents (which are grouped by genre etc.) or rather on term clustering within individual documents as has often been suggested (e.g. Katz 1996). A final important question is whether the extrapolation quality of a given model on a given data set can be predicted in a reliable way, using only the part of the data available for parameter estimation.

References

Aston, G. and Burnard, L. (1998) *The BNC handbook: Exploring the British National Corpus with SARA* (Edinburgh: Edinburgh University Press).

Baayen, H. (1991) Quantitative aspects of morphological productivity, in G. Boij and J. van Marle (eds.) *Yearbook of Morphology 1991* (Dordrecht: Springer), 109–149.

Baayen, H. (2001) *Word frequency distributions* (Dordrecht: Kluwer).

Chipere, N., Malvern, D. and Richards, B. (2004) Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios, in G. Aston, S. Bernardini and D. Stewart (eds.) *Corpora and language learners* (Amsterdam: Benjamins), 139–147.

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

Evert, S. and Lüdeling, A. (2001) Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of the Corpus Linguistics 2001 Conference* (Lancaster, UK), 167–175.

Ha, L. Q.; Sicilia-Garcia, E. I.; Ming, J. and Smith, F. J. (2002) Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, 315–320.

Johansson, S., Leech, G. and Goodluck, H. (1978) *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers* (Department of English, University of Oslo).

Katz, S. M. (1996) Distribution of content words and phrases in text and language modelling. *Natural Language Engineering* **2**(2), 15–59.

Lüdeling, A. and Evert, S. (2003) Linguistic experience and productivity: corpus evidence for fine-grained distinctions. *Proceedings of the Corpus Linguistics 2003 Conference* (Lancaster, UK), 475–483.

Matsumoto, Y.; Kitauchi, A.; Yamashita, T.; Hirano, Y.; Matsuda, H.; Takaoka, K. and Asahara, M. (2000) *Morphological analysis system ChaSen version 2.2.1 manual* (Computational Linguistics Laboratory, Nara Institute of Science and Technology).

Ueyama, M. and Baroni M. (2005) Automated construction and evaluation of a Japanese web-based reference corpus. *Proceedings of Corpus Linguistics 2005* (Birmingham, UK).

Youmans, G. (1990) Measuring lexical style and competence: The type-token vocabulary curve. *Style* **24***,* 584–599.

Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort* (Cambridge, MA: Addison-Wesley).