Towards a computational model of grammaticalization and lexical diversity

Christian Bentz
University of Cambridge, DTAL
9 West Road, CB3 9DA
cb696@cam.ac.uk

Paula Buttery
University of Cambridge, DTAL
9 West Road, CB3 9DA
pjb48@cam.ac.uk

Abstract

Languages use different lexical inventories to encode information, ranging from small sets of simplex words to large sets of morphologically complex words. Grammaticalization theories argue that this variation arises as the outcome of diachronic processes whereby co-occurring words merge to one word and build up complex morphology. To model these processes we present a) a quantitative measure of lexical diversity and b) a preliminary computational model of changes in lexical diversity over several generations of merging higly frequent collocates.

1 Introduction

All languages share the property of being carriers of information. However, they vastly differ in terms of the exact encoding strategies they adopt. For example, German encodes information about number, gender, case, tense, aspect, etc. in a multitude of different articles, pronouns, nouns, adjectives and verbs. This abundant set of word forms contrasts with a smaller set of uninflected words in English.

Crucially, grammaticalization theories (Heine and Kuteva, 2007, 2002; Bybee 2006, 2003; Hopper and Traugott, 2003; Lehmann, 1985) demonstrate that complex morphological marking can derive diachronically by merging originally independent word forms that frequently co-occur. Over several generations of language learning and usage such grammaticalization and entrenchment processes can gradually increase the complexity of word forms and hence the lexical diversity of languages.

To model these processes Section 2 will present a quantitative measure of lexical diversity based on Zipf-Mandelbrots law, which is also used as a biodiversity index (Jost, 2006). Based on this measure we present a preliminary computational model to reconstruct the gradual change from lexically constrained to lexically rich languages in Section 3. We therefore use a simple grammaticalization algorithm and show how historical developments towards higher lexical diversity match the variation in lexical diversity of natural languages today. This suggests that *synchronic* variation in lexical diversity can be explained as the outcome of *diachronic* language change.

The computational model we present will therefore help to a) understand the diversity of lexical encoding strategies across languages better, and b) to further uncover the diachronic processes leading up to these synchronic differences.

2 Zipf's law as a measure of lexical diversity

Zipf-Mandelbrot's law (Mandelbrot, 1953; Zipf, 1949) states that ordering of words according to their frequencies in texts will render frequency distributions of a specific shape: in general, few words have high frequencies, followed by a middle ground of medium frequencies and a long tail of low frequency items.

However, a series of studies pointed out that there are subtle differences in frequency distributions for different texts and languages (Bentz et al., forthcoming; Ha et al., 2006; Popescu and Altmann, 2008). Namely, languages with complex morphology tend to have longer tails of low frequency words than languages with simplex morphology. The parameters of Zipf-Mandelbrot's law reflect these differences, and can be used as a quantitative

Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL) @ EACL 2014, pages 38–42,
Gothenburg, Sweden, April 26 2014. ©2014 Association for Computational Linguistics

measure of lexical diversity.

2.1 Method

We use the definition of ZM's law as captured by equation (1):

$$f(r_i) = \frac{C}{\beta + r_i^{\alpha}}\,, \\ C > 0, \alpha > 0,\ \ \beta > -1,\ \ i = 1, 2, \ldots, n \qquad (1)$$

where $f(r_i)$ is the frequency of the word of the i^{th} rank (r_i), n is the number of ranks, C is a normalizing factor and α and β are parameters. To illustrate this, we use parallel texts of the *Universal Declaration of Human Rights* (UDHR) for Fijian, English, German and Hungarian. For frequency distributions of these texts (with tokens delimited by white spaces) we can approximate the best fitting parameters of the ZM law by means of maximum likelihood estimation (Izsák, 2006; Murphy, 2013). In double logarithmic space (see Figure 1) the normalizing factor C would shift the line of best fit upwards or downwards, α is the slope of this line and β is Mandelbrot's (1953) corrective for the fact that the line of best fit will deviate from a straight line for higher frequencies (upper left corner in Figure 1).

As can be seen in Figure 1 Fijian has higher frequencies towards the lowest ranks (upper left corner) but the shortest tail of words with frequency one (horizontal bars in the lower right corner). For Hungarian the pattern runs the other way round: it has the lowest frequencies towards the low ranks and a long tail of words with frequency one. German and English lie between these. These patterns are reflected in ZM parameter values. Namely, Fijian has the highest parameters, followed by English, German and Hungarian. By trend there is a negative relationship between ZM parameters and lexical diversity: low lexical diversity is associated with high parameters, high diversity is associated with low parameters. Cross-linguistically this effect can be used to measure lexical diversity by means of approximating the parameters of ZM's law for parallel texts.

In the following, we will present a computational model to elicit the diachronic pathways of grammaticalization through which a low lexical diversity language like Fijian might develop towards a high diversity language like Hungarian.

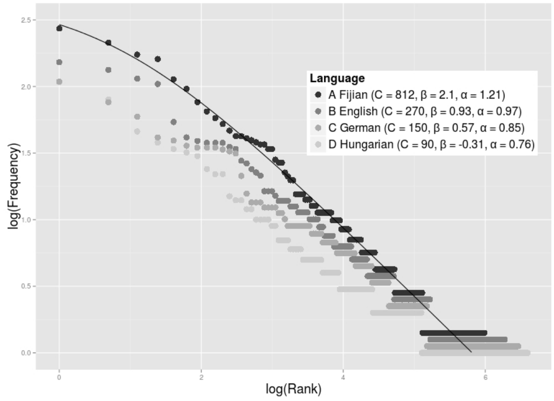

Figure 1: Zipf frequency distributions for four natural languages (Fijian, English, German, Hungarian). Plots are in log-log space, values 0.15, 0.1 and 0.05 were added to Fijian, English and German log-frequencies to avoid overplotting. Values for the Zipf-Mandelbrot parameters are given in the legend. The straight black line is the line of best fit for Fijian.

3 Modelling changes in lexical diversity

Grammaticalization theorists have long claimed that synchronic variation in word complexity and lexical diversity might be the outcome of diachronic processes. Namely, the grammaticalization cline from *content item >grammatical word >clitic >inflectional affix* is seen as a ubiquitous process in language change (Hopper and Traugott, 2003: 7). In the final stage frequently co-occurring words merge by means of phonological fusion (Bybee, 2003: 617) and hence 'morphologize' to built inflections and derivations.

Typical examples of a full cline of grammaticalization are the Old English noun *līc* 'body' becoming the derivational suffix *-ly*, the inflectional future in Romance languages such as Italian *canterò* 'I will sing' derived from Latin *cantare habeo* 'I have to sing', or Hungarian inflectional elative and inessive case markers derived from a noun originally meaning 'interior' (Heine and Kuteva, 2007: 66). These processes can cause languages to distinguish between a panoply of different word forms. For example, Hungarian displays up to 20 different noun forms where English would use a single form (e.g. *ship* corresponding to Hungarian *hajó* 'ship', *hajóban* 'in the ship', *hajóba* 'into the ship', etc.).

As a consequence, once the full grammaticalization cline is completed this will increase the lexical diversity of a language. Note, however, that borrowings (loanwords) and neologisms can also increase lexical diversity. Hence, a model of changes in lexical diversity will have to take both grammaticalization and new vocabulary into account.

3.1 The model

Text: We use the Fijian UDHR as our starting point for two reasons: a) Fijian is a language that is well known to be largely lacking complex morphology, b) the UDHR is a parallel text and hence allows us to compare different languages by controlling for constant information content. Fijian has relatively low lexical diversity and high ZM parameter values (see Figure 1). The question is whether we can simulate a simple merging process over several generations that will transform the frequency distribution of the original Fijian text to fit the frequency distribution of the morphologically and lexically rich Hungarian text. To answer this question, we simulate the outcome of grammaticalization on the frequency distributions in the following steps:

Simulation: Our program takes a given text of generation i, calculates a frequency distribution for this generation, changes the text along various operations given below, and gives the frequency distribution of the text for a new generation $i + 1$ as output.

We take the original UDHR in Fijian as our starting point in generation 0 and run the program for consecutive generations. We simulate the change of this text over several generations of language learning and usage by varying the following variables:

- p_m: Rank bigrams according to their frequency and merge the highest p_m percent of them to one word. This simulates a simple grammaticalization process whereby two separate words that are frequent collocates are merged to one word.
- p_v: Percentage of words replaced by new words. Choose p_v of words randomly and replace all instances of these words by inverting the letters. This simulates neologisms and loanwords replacing deprecated words.
- r_R: Range of ranks to be included in p_v replacements. If set to 0, vocabulary from anywhere in the distribution will be randomly replaced.
- n_G: Number of generations to simulate.

This simulation essentially allows us to vary the degree of grammaticalization by means of varying p_m, and also to control for the fact that frequency distributions might change due to loanword borrowing and introduction of new vocabulary (p_v). Additionally, r_R allows us to vary the range of ranks where new words might replace deprecated ones. For frequency distributions calculated by generations we approximate ZM parameters by maximum likelihood estimations and therefore document the change of their shape.

Results: Figure 2 illustrates a simulation of how the low lexical diversity language Fijian approaches quantitative lexical properties similar to the Hungarian text just by means of merging high-frequent collocates. While the frequency distribution of Fijian in generation 0 still reflects the original ZM values, the ZM parameter values after 6 generations of grammaticalization have become much closer to the values of the Hungarian UDHR:

Fij ($n_G = 0$): $\alpha = 1.21$,$\beta = 2.1$,$C = 812$
Fij ($n_G = 6$): $\alpha = 0.70$,$\beta = -0.22$,$C = 73$
Hun ($n_G = 0$): $\alpha = 0.76$,$\beta = -0.31$,$C = 90$

Note, that in this model there is actually no replacement of vocabulary necessary to arrive at frequency distributions that correspond to high lexical diversity variants. After only six generations of merging 2.5% of bigrams to a single grammaticalized word the Fijian UDHR has ZM parameter properties very close to the Hungarian UDHR. However, in future research we want to scrutinize the effect of parameter changes on frequency distributions in more depth and in accordance

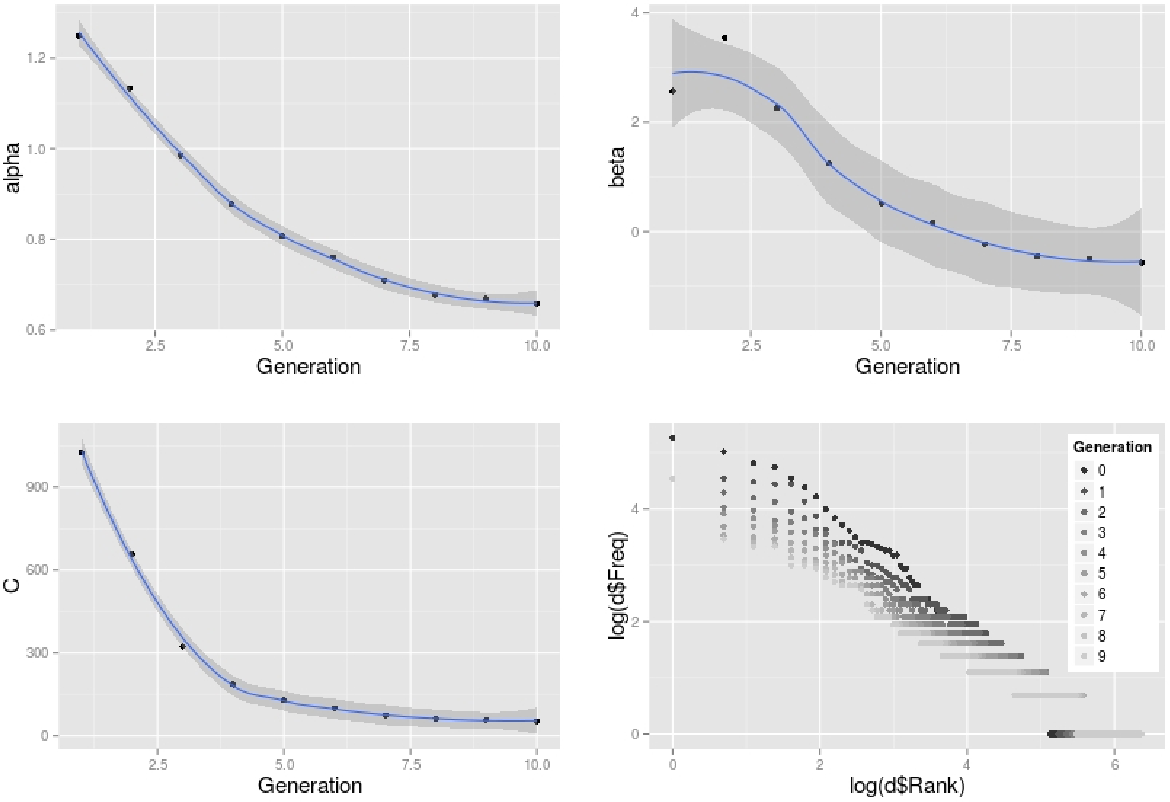

Figure 2: Simulation of grammaticalization processes and their reflections in Zipf distributions for variable values $p_m = 2.5,\ p_v = 0, r_R = 0, n_G = 10$. Changes of α are shown in the upper left panel, changes in β are shown in the upper right panel, changes in C are shown in the lower left panel, and changes in log-transformed frequency distributions are illustrated in the lower right panel.

with estimations derived from historical linguistic studies.

4 Discussion

We have pointed out in Section 2 that lexical diversity can be measured cross-linguistically by means of calculating frequency distributions for parallel texts and approximating the corresponding ZM parameters in a maximum likelihood estimation.

It is assumed that cross-linguistic variation is the outcome of diachronic processes of grammaticalization, whereby highly frequent bigrams are merged into a single word. The preliminary computational model in Section 3 showed that indeed even by a strongly simplified grammaticalization process a text with low lexical diversity (Fijian UDHR) can gain lexical richness over several generations, and finally match the quantitative properties of a lexically rich language (Hungarian UDHR).

However, there are several caveats that need to be addressed in future research:

- More models with varying parameters need to be run to scrutinize the interaction between new vocabulary (loanwords, neologisms) and grammaticalization.

- The grammaticalization algorithm used is overly simplified. A more realistic picture is possible by using POS tagged and parsed texts to ensure that only certain parts of speech in certain syntactic contexts grammaticalize (e.g. pre- and postpositions in combination with nouns).

- The model could be elaborated by considering not only bigram frequencies but also frequencies of the individual words and more complex frequency measures (see Schmid, 2010).

5 Conclusion

Languages display an astonishing diversity when it comes to lexical encoding of information. This *synchronic* variation in encoding strategies is most likely the outcome of *diachronic* processes of language change. We have argued that lexical diversity can be measured quantitatively with reference to the parameters of Zipf-Mandelbrot's law, and that pathways of change in lexical diversity can be modelled computationally. Elaboration and refinement of these models will help to better understand linguistic diversity as the outcome of processes on historical and evolutionary time scales.

References

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An international handbook*. Berlin/ New York, Mouton de Gruyter, pages 803-826.

Christian Bentz, Douwe Kiela, Felix Hill, and Paula Buttery. forthcoming. Zipf's law and the grammar of languages. In *Corpus Linguistics and Linguistic Theory.*

Joan Bybee. 2006. From usage to grammar: The mind's repsonse to repetition. In *Language*, volume 82 (4), pages 711-733.

Joan Bybee. 2003. Mechanisms of change in grammaticization: the role of frequency. In B. D. Joseph and J. Janda(eds.), *The Handbook of Historical Linguistics.* Oxford, Blackwell, pages 711-733.

Le Q. Ha, Darryl Stewart, Philip Hanna, and F. Smith. 2006. Zipf and type-token rules for the English, Spanish, Irish and Latin languages. In *Web Journal of Formal, Computational and Cognitive Linguistics*, volume 8.

Bernd Heine and Tania Kuteva. 2007. *The Genesis of Grammar: A Reconstruction.* Oxford University Press.

Bernd Heine and Tania Kuteva. 2002. *World lexicon of grammaticalization.* Cambridge University Press.

Paul J. Hopper and Elizabeth C. Traugott. 2003. *Grammaticalization.* Cambridge University Press.

János Izsák. 2006. Some practical aspects of fitting and testing the Zipf-Mandelbrot model: A short essay. In *Scientometrics*, volume 67(1), pages 107-120.

Lou Jost. 2006. Entropy and diversity. In *OIKOS*, volume 113(2), pages 363-375.

Christian Lehmann. 1985. Grammaticalization: Synchronic variation and diachronic change. In *Lingua e stile*, volume 20, pages 303-318.

Benoit Mandelbrot. 1953. An informational theory of the statistical structure of language. In William Jackson (ed.), *Communication Theory.* Butterworths Scientific Publications, London, pages 468-502.

Laura Murphy. 2013. *R package likelihood: Methods for maximum likelihood estimation.* Retrieved from cran.r-project.org/web/packages/likelihood

Ioan-Iovitz Popescu, and Gabriel Altmann. 2008. Hapax legomena and language typology. In *Journal of Quantitative Linguistics*, volume 15(4), pages 370378.

Hans-J. Schmid. 2010. Does frequency in text instantiate entrenchment in the cognitive system? In Dylan Glynn and Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches.* Berlin, Walter de Gruyter, pages 101-133.

George K. Zipf. 1949. *Human behavior and the principle of least effort.* Addison, Cambridge (Massachusetts).