ZIPF'S LAW ACROSS LANGUAGES OF THE WORLD: TOWARDS A QUANTITATIVE MEASURE OF LEXICAL DIVERSITY

CHRISTIAN BENTZ
Department of Theoretical and Applied Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge, CB3 9DA
United Kingdom

DOUWE KIELA
Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge, CB3 0FD, United Kingdom

Zipf's law (Zipf, 1949) has been argued to differ systematically between texts and languages (Popescu et al., 2009), to change throughout time (Bentz, Kiela, Hill, & Buttery, forthcoming) and to reflect language complexity (Baixeries, Elvevåg, & Ferrer-i-Cancho, 2013). Furthermore, we argue that Zipf's law can be used as cross-linguistic, quantitative measure of the lexical diversity of languages (in parallel to biodiversity indices). In this context, lexical diversity is defined as the breadth of word forms used to encode a constant information content. A quantitative measure of lexical diversity can help to a) model factors driving the evolution of lexical encoding strategies on historical and evolutionary timescales, b) to determine the range of lexical diversities in natural languages and distinguish them from other symbolic encoding systems and animal communication.

To show this, we estimated parameters of the Zipf-Mandelbrot law (Mandelbrot, 1953) for 363 parallel translations of the *Universal Declaration of Human Rights* (UDHR) using a maximum likelihood method. The theoretical ZM distribution is assumed to be

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

where $f(r_i)$ is the frequency of a word of i^{th} rank (r_i) in a rank-frequency profile, n is the number of ranks, C is a normalizing factor and β and α are parameters.

The ML estimation shows that parameters differ systematically between languages. Namely, languages with low lexical diversity (e.g. Pidgin Nigerian, Fijian) have higher parameters, whereas lexically rich languages display lower parameters (e.g. Greenlandic, Hungarian). Moreover, we present evidence that a) all 363 languages in our sample fall within a relatively narrow range of lexical diversity, b) languages of the same family cluster according to lexical diversity, but also c) languages with more non-native speakers (more language contact) have systematically reduced lexical diversities.

Based on these quantitative findings, we argue that lexical diversity can be modeled by taking into account genealogical and sociolinguistic factors. This will help us understand how and why lexical encoding systems in natural languages differ from other encoding systems and how they evolved over time.

References

Baixeries, J., Elvevåg, B., & Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipf 's law in language ontogeny. *PloS one*, *8*(3), e53227. doi:10.1371/Citation

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (n.d.). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. Jackson (Ed.), *Communication Theory* (pp. 468–502). London: Butterworths Scientific Publications.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., et al. (2009). *Word frequency studies*. Berlin & New York: Mouton de Gruyter.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge (Massachusetts): Addison-Wesley.