Zipf's law and the grammar of languages

Synthetic and analytic encoding strategies across languages of the world

Christian Bentz, University of Cambridge *

Arts & Humanities Research Council

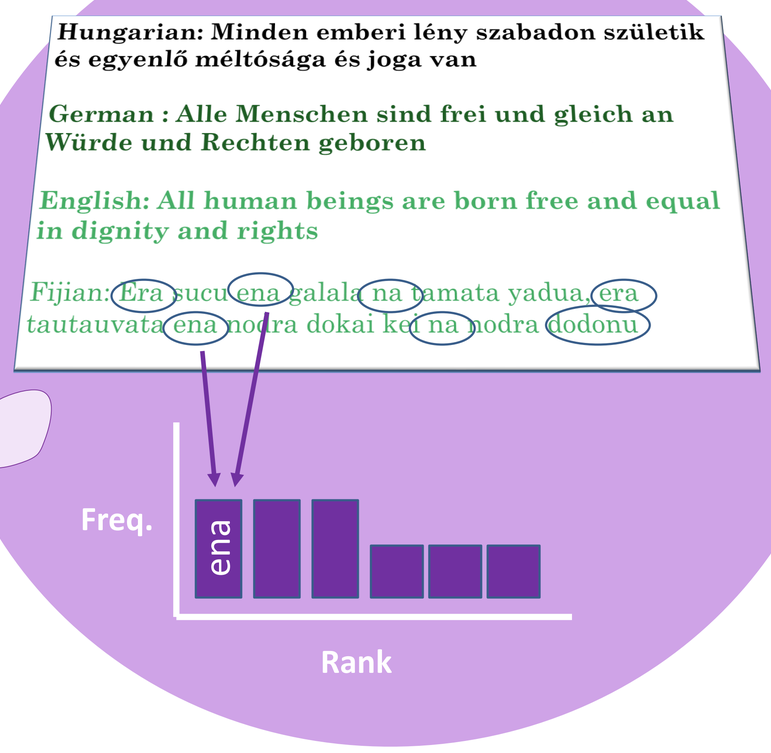
CAMBRIDGE ASSESSMENT

Introduction

Zipf's law (Zipf, 1949) denotes one of the most well-known quantitative relationships in language, but it has been argued to be linguistically 'shallow' (Miller, 1957). However, a series of studies showed that Zipf's law systematically differs across texts and languages (Popescu et al., 2009), that it reflects language complexity (Baixeries, Elvevåg, & Ferrer-i-Cancho, 2013) and that it changes systematically according to whether languages use analytic strategies of encoding (i.e. repeating highly frequent words) or synthetic strategies (i.e. complex morphology, compounding etc.) (Bentz, Kiela, Hill & Buttery, forthcoming).
Based on these findings, this poster presents a „recipe" for a syntheticity scale on which over 350 languages can be rated. Moreover, it will discuss how such ratings can help to address current issues in language typology, historical language change and language evolution.

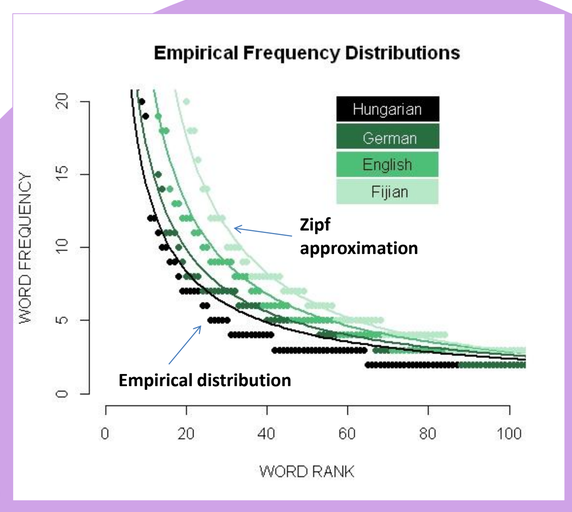
Recipe: Syntheticity Scale

1.



Take parallel translations (e.g. *Universal Declaration of Human Rights* for 363 languages) and create Zipf distributions for the texts of every language by ordering the occurring words according to their frequencies.

2.



Rather synthetic languages will have Zipf curves with longer tails and lower frequencies towards the higher ranks. This is because complex morphology and compounding create low frequency items (Bentz, Kiela, Hill & Buttery [forthcoming]).

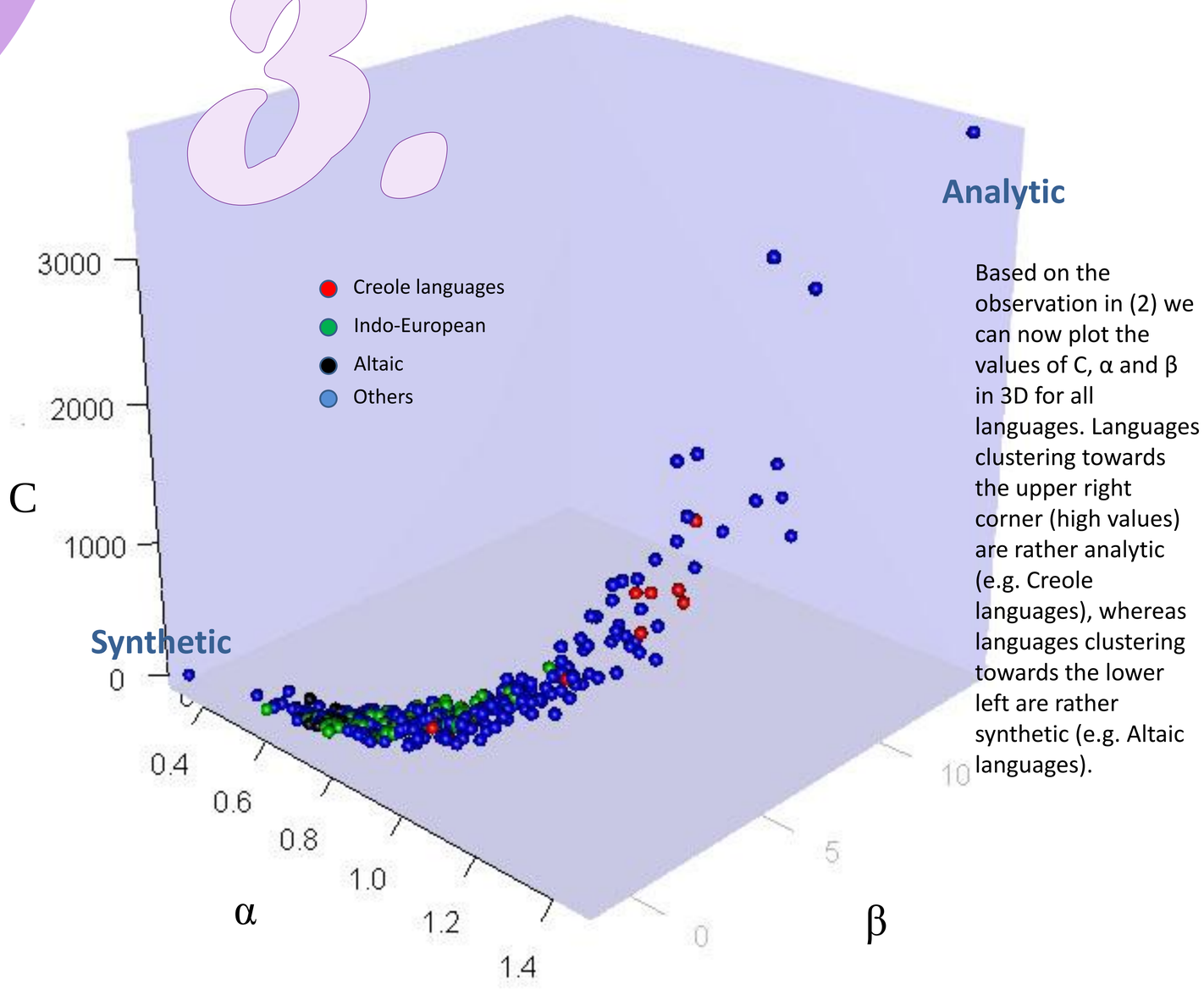
The empirical distributions of words can now be approximated by the **Zipf-Mandelbrot law**

$$f(r) = \frac{C}{(\beta+r)^{\alpha}}$$

The differences in Zipf curves are reflected in the parameters of the ZM law: The constant C, the parameter β and the parameter α are all systematically lower for synthetic languages.

C < C < C < C
β < β < β < β
α < α < α < α

3.



Based on the observation in (2) we can now plot the values of C, α and β in 3D for all languages. Languages clustering towards the upper right corner (high values) are rather analytic (e.g. Creole languages), whereas languages clustering towards the lower left are rather synthetic (e.g. Altaic languages).

Implications

a) **Zipf's law is not just a statistical artifact** and linguistically 'shallow'. It reflects details about how languages encode information. This is reflected in varying parameters of the Zipf-Mandelbrot modification (Mandelbrot 1953).

b) **The syntheticity scale** can help to quantitatively measure differences in how languages encode information. For example, it has been argued that creole grammars are amongst the simplest grammars of the world due to their lack of overt morphology. This hypothesis is supported by evidence from the syntheticity scale.

c) It is a long held assumption that languages are all equally complex. However, it still remains notoriously difficult to define and measure **language complexity** in an unifying way. The syntheticity scale could help to do that.

d) Zipf's law could be applied diachronically to measure changing morphological marking strategies on **historical and evolutionary time scales** (see also Bentz, Kiela, Hill & Buttery, forthcoming).

* **Contact:**
chris@christianbentz.de (e-mail)
christianbentz.de (website)

References

Bentz, C. ,Kiela, D., Hill, F. & Buttery, P. (forthcoming). Zipf's law and the grammar of languages. A quantitative study of Old and Modern English parallel texts.
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
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.
Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, *70*(2), 311–314.
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.