Zipf's law and the grammar of languages
Synthetic and analytic encoding strategies across languages of the world
Christian Bentz, University of Cambridge *

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 (Baxieres, Evelvæg, & Ferrer-i-Cancho, 2013) and that it changes systematically according to whether languages use analytic strategies of encoding (i.e. repeating frequently used 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: Synthetic Scale

1. Hungarian: Minden emberi lény szabadon születik és egyenlő méltóságban és joga van
German: Alle Menschen sind frei und gleich an Würde und Rechten geboren
English: All human beings are born free and equal in dignity and rights
Finnish: Kaikkien ihmisten on oikeus olla vapaa ja arvostettava

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

\[ f(r) = \frac{c}{(r + \gamma)^\alpha} \]

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.

3. The differences in Zipf curves are approximated by the Zipf-Mandelbrot law.

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).

References


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