Zipf’s law of abbreviation as a language universal

Christian Bentz
Department of General Linguistics
University of Tübingen
Naukerstraße 35, 72074 Tübingen
Email: chris@christianbentz.de

Ramon Ferrer-i-Cancho
Departament de Ciencies de la Computacio
Universitat Politecnica de Catalunya (UPC)
Campus Nord, Edifici Omega, Despatx S124
08034 Barcelona, Catalonia, Spain

Abstract—Words that are used more frequently tend to be shorter. This statement is known as Zipf’s law of abbreviation. Here we perform the widest investigation of the presence of the law to date. In a sample of 1262 texts and 986 different languages - about 13% of the world’s language diversity - a negative correlation between word frequency and word length is found in all cases. In line with Zipf’s original proposal, we argue that this universal trend is likely to derive from fundamental principles of information processing and transfer.

I. INTRODUCTION

Zipf’s law of abbreviation [32], [30], [31] states that frequently used words tend to be shorter. In English, for example, some of the most frequent words are short articles and prepositions such as the, and, of and a, as opposed to low-frequency content words such as harpsichord, deforestation and marginalizing. The negative association between lengths and frequencies of words holds in every language for which it has been tested [32], [1], [28], [25], [13], [29]. However, the focus of earlier studies was not on testing the law as a linguistic universal, and the language samples are small (e.g. 7 languages in [1] and [13], 11 languages in [25]).

Recently, massively parallel corpora have become available. These allow us to test quantitative laws more systematically across many languages. The study reported here uses a sample of 1263 texts written in 986 different languages of 80 different families. It is shown that a negative correlation between word frequency and word length is found in all texts and languages. Moreover, all texts in the PBC have a negative correlation with a p-value smaller than 0.05.

II. MATERIALS AND METHODS

The texts in our sample are parallel translations of the Universal Declaration of Human Rights (UDHR)\(^1\) and the Parallel Bible Corpus (PBC) [22]. The UDHR comprises 376 parallel translations converted into unicode. The PBC currently comprises 918 parallel translations that have been assigned 810 unique ISO 639-3 codes. Overall, the samples amount to 1263 texts with 986 different ISO 639-3 codes, i.e. unique languages. The Glottolog\(^2\) lists 7748 unique languages currently spoken in the world. The text sample used here covers 12.7% of these.

\(^1\)http://unicode.org/udhr/


Here word frequencies are counted as the number of occurrences of word types. Word types, in turn, are defined as strings of unicode characters delimited by non-alphanumeric characters (e.g. punctuation, white spaces, etc.). The number of unicode characters per word type in the PBC and UDHR are calculated using the function `nchar()` in R [26].

To investigate the association between the frequency of a type and its length the *Kendall rank correlation* [8], [15], [27], [13] is used. The statistic is chosen for its capacity to capture non-linear dependencies and for its intimate relationship with the minimization of the energetic cost of a vocabulary [10].

III. RESULTS

Table I (rows 2 and 3) shows that the correlation between word frequency and word length is always negative for all texts and languages. Moreover, all texts in the PBC have a negative correlation with a p-value smaller than $\alpha = 0.05$. This is also the case for 92.3% of texts in the UDHR. However, 27 languages of the UDHR have a p-value bigger than 0.05.

Fig. 1 is a visual illustration of the frequencies of words and their lengths for texts across 37 different language families of the UDHR. For each family the language with the mode Kendall’s $\tau$ value was selected to represent that family.

IV. DISCUSSION

The frequency/length relationship for words as noticed by Zipf emerges as an empirical universal across 1263 texts.
written in 986 languages of 80 language families. This result is very unlikely to occur by chance. Its cross-linguistic strength qualifies it as a candidate for a linguistic universal. However, there are several further caveats that need to be addressed in future research:

A. Absolute, statistical, and evolutionary universals

Linguistic research traditionally distinguishes between absolute and statistical universals [4], [3]. Absolute universals are supposed to hold across all human languages, be it extant or extinct. Absolute universality in this sense might well turn out to be impossible to prove empirically [24]. Statistical universals, on the other hand, are merely strong tendencies found in large-scale comparative data.

Furthermore, a growing body of research points to biological and communicative constraints that universally shape the evolution of languages. From this perspective, universals are not only properties of currently attested languages (synchronic universals), but rather universal processing constraints that play out on the evolutionary time scale [6], [17], [18], [5], [2], [7]. These could be called evolutionary universals.

Based on the results reported in this study it is reasonable to assume that Zipf’s law of abbreviation surfaces in all, or at least a very high percentage of attested languages. It is thus a candidate for an absolute synchronic universal. The diachronic, evolutionary pressures towards the shortening of word forms still need to be uncovered.
B. Is this result trivial?

The inevitability of quantitative laws of language has been a matter of ongoing discussion over decades [23], [19], [11], [12]. For example, random typing models [23], [19], [9] have been invoked to show that quantitative laws in language might be statistical artefacts and hence “linguistically shallow”. Preliminary analyses suggest that there are systematic differences between Zipf’s law of abbreviation in natural languages and random typing models. Further analyses are necessary to clearly delimit the occurrence of the law in natural languages from its occurrence in random typing.

C. The problem of explanation

If the law of abbreviation is not trivial, then how can it be explained? We put forward the hypothesis that the law is a mediation between two major constraints: the pressure to reduce the cost of production, i.e. the pressure for brevity, on the one hand, and the pressure to maximize transmission success [14], on the other hand. This idea goes back to Zipf and his principle of least effort [32]. This principle of reducing the cost of production might also be related to the informativeness of words [21], [25].

D. Communication or language?

There is evidence suggesting that the principle of least effort also acts upon communicative and (potentially) non-communicative behavior of other species [15], [16], [27], [20]. Ultimately, Zipf’s law of abbreviation might emerge as a universal of communication systems in which the principle of compression outweighs the principle of transmission success. This distinction, however, does not necessarily intersect with the human/non-human distinction.

E. The problem of text size

The law of abbreviation emerges from the accumulation of lengths and frequencies of hundreds and thousands of word tokens. How does the correlation depend on text size, and what is the minimum number of tokens required to get a robust correlation? Preliminary analyses suggest that text sizes of around 250 tokens are sufficient (for most languages).

V. CONCLUSIONS

Zipf’s law of abbreviation holds across all 1263 texts and 986 languages tested here. The robustness of the law calls for theoretical explanation. This is particularly important since it can shed new light on the discussion about linguistic universals. Universal properties of language might, after all, exist. However, it is possible that they derive from fundamental principles of information transfer, rather than language and human specific biases.

ACKNOWLEDGMENT

CB was funded by an Arts and Humanities Research Council (UK) doctoral grant and Cambridge Assessment (reference number: RG 69405), as well as a grant from the Cambridge Home and European Scholarship Scheme. At a later stage this project was also supported by the ERC Advanced Grant 324246 EVOLAE and the DFG-KFG 2237 Words, Bones, Genes, Tools.

RFC was supported by the grant APCOM project (TIN2014-57226-P) from the Spanish Ministry of Science and Innovation. RFC was partially supported by the grant 2014SGR 890 (MACDA) from AGAUR, Generalitat de Catalunya.

REFERENCES


