# Zipf's law of abbreviation as a language universal

Christian Bentz
Department of General Linguistics
University of Tübingen
Nauklerstraße 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. We argue that this finding has important implications for the discussions surrounding language universals, and the search for the biological underpinnings of human language.

## 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/

[2] Version 2.7, http://glottolog.org, accessed on 2016-02-25

TABLE I
The concordance with Zipf's law of abbreviation across 986 languages. For each dataset, $N$ is the number of texts or languages, $N_\alpha^-$ is the number of negative correlations between word frequency and word length with p-values not exceeding $\alpha$; $N_\alpha^+$ is the converse of $N_\alpha^-$ for positive correlations.

| | Texts | | Languages | |
|---|---|---|---|---|
| | PBC | UDHR | PBC | UDHR |
| $N$ | 907 | 355 | 801 | 332 |
| $N_1^-$ | 907 | 355 | 801 | 332 |
| $N_1^+$ | 0 | 0 | 0 | 0 |
| $N_{0.05}^-$ | 907 | 328 | 801 | 307 |
| $N_{0.01}^-$ | 907 | 316 | 801 | 296 |
| $N_{0.001}^-$ | 907 | 283 | 801 | 265 |
| $N_{0.0001}^-$ | 907 | 245 | 801 | 230 |

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

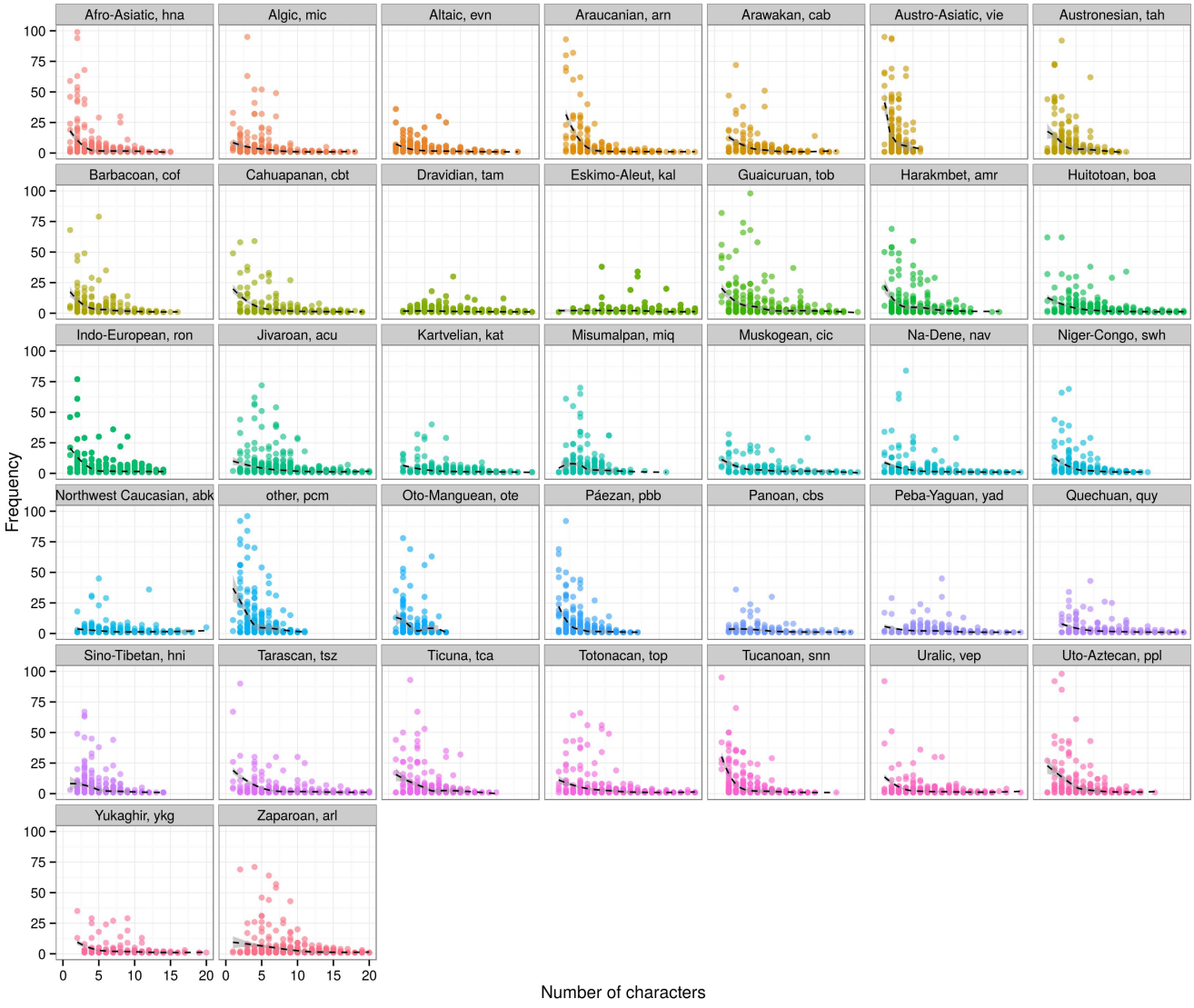


Fig. 1. The relationship between frequency of occurrence (y-axis) and number of characters (x-axis) for different language families in the UDHR. The languages selected represent the *mode* of kendall's $\tau$ values across all the languages of the family. For illustration purposes, a LOESS smoother with 95% confidence intervals is added (dashed black). The grey boxes above plots give names of language families and the ISO 639-3 codes for languages chosen to represent them.

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 EVOLAEMP 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

[1] E. Bates, S. D'Amico, T. Jacobsen, A. Székely, E. Andonova, A. Devescovi, D. Herron, C.C. Lu, T. Pechmann, C. Pléh, N. Wicha, K. Federmeier, I. Gerdjikova, G. Gutierrez, D. Hung, J. Hsu, G. Iyer, K. Kohnert, T. Mehotcheva, A. Orozco-Figueroa, A. Tzeng, and O. Tzeng. Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, 10:344–380, 2003.

[2] B. Bickel. Typology in the 21st century: major current developments. *Linguistic Typology*, 11(1):239–251, 2007.

[3] B. Bickel. Absolute and statistical universals. *The Cambridge Encyclopedia of the Language Sciences*, pages 77–79, 2010.

[4] B. Bickel. Linguistic diversity and universals. *Cambridge Handbook of Linguistic Anthropology. Cambridge: Cambridge University Press, to appear*, 2013.

[5] J. Blevins. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press, 2004.

[6] J. L. Bybee. The diachronic dimension in explanation. In *Explaining language universals*. 1988.

[7] M. H. Christiansen and N. Chater. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509, 2008.

[8] W. J. Conover. *Practical nonparametric statistics*. Wiley, New York, 1999. 3rd edition.

[9] B. Conrad and M. Mitzenmacher. Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Transactions on Information Theory*, 50(7):1403–1414, 2004.

[10] R. Ferrer-i-Cancho, C. Bentz, and C. Seguin. Compression and the origins of Zipf's law of abbreviation. *http://arxiv.org/abs/1504.04884*, 2015.

[11] R. Ferrer-i-Cancho and B. Elvevåg. Random texts do not exhibit the real Zipf's-law-like rank distribution. *PLoS ONE*, 5(4):e9411, 2009.

[12] R. Ferrer-i-Cancho, N. Forns, A. Hernández-Fernández, G. Bel-Enguix, and J. Baixeries. The challenges of statistical patterns of language: the case of Menzerath's law in genomes. *Complexity*, 18(3):11–17, 2013.

[13] R. Ferrer-i-Cancho and A. Hernández-Fernández. The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4(1), 2013.

[14] R. Ferrer-i-Cancho, A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M. J. Hsu, and S. Semple. Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578, 2013.

[15] R. Ferrer-i-Cancho and D. Lusseau. Efficient coding in dolphin surface behavioral patterns. *Complexity*, 14(5):23–25, 2009.

[16] J. P. Hailman, M. S. Ficken, and R. W. Ficken. The 'chick-a-dee' calls of *Parus atricapillus*: a recombinant system of animal communication compared with written English. *Semiotica*, 56:121–224, 1985.

[17] C. J. Hall. Integrating diachronic and processing principles in explaining the suffixing preference. In J. A. Hawkins, editor, *Explaining language universals*, pages 321–349. Oxford: Basil Blackwell, 1988.

[18] S. Kirby. *Function, Selection, and Innateness: The Emergence of Language Universals: The Emergence of Language Universals*. Oxford University Press, 1999.

[19] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions of Information Theory*, 38(6):1842–1845, 1992.

[20] B. Luo, T. Jiang, Y. Liu, J. Wang, A. Lin, X. Wei, and J. Feng. Brevity is prevalent in bat short-range communication. *Journal of Comparative Physiology A*, 199:325–333, 2013.

[21] K. Mahowald, E. Fedorenko, S. T. Piantadosi, and E. Gibson. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318, 2013.

[22] T. Mayer and M. Cysouw. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik,*

*Iceland, May 26-31, 2014.*, pages 3158–3163. European Language Resources Association (ELRA), 2014.

[23] G. A. Miller. Some effects of intermittent silence. *The American Journal of Psychology*, 70:311–314, 1957.

[24] S. T. Piantadosi and E. Gibson. Quantitative standards for absolute linguistic universals. *Cognitive Science*, 38(4):736–756, 2013.

[25] S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[27] S. Semple, M. J. Hsu, and G. Agoramoorthy. Efficiency of coding in macaque vocal communication. *Biology Letters*, 6:469–471, 2010.

[28] U. Strauss, P. Grzybek, and G. Altmann. Word length and word frequency. In P. Grzybek, editor, *Contributions to the science of text and language*, pages 277–294. Springer, Dordrecht, 2007.

[29] G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann. Towards a theory of word length distribution*. *Journal of Quantitative Linguistics*, 1(1):98–106, 1994.

[30] G. K. Zipf. *Selected studies of the principle of relative frequency in language*. Harvard University Press, Cambridge (Massachusetts), 1932.

[31] G. K. Zipf. *The psycho-biology of language*. The M.I.T Press, Cambridge (Massachusetts), 1935.

[32] G. K. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge (MA), USA, 1949.