FREQUENCY AND CHARACTER COMPLEXITY IN 27 WRITING SYSTEMS

Alexey Koshevoy[*1], Helena Miton[2,3], and Olivier Morin[1,2]

[*]Corresponding Author: akoshevoy@ens.fr
[1]Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, UMR 8129
[2]Minds and Traditions Group, Max Planck Institute for the Science of Human History, Jena, Germany
[3]Santa Fe Institute, Santa Fe, USA

We are testing whether Zipf's Law of abbreviation — the idea that more frequent symbols in a code are simpler than less frequent ones — is present at the level of individual characters. In its original and most well-known manifestation, the law of abbreviation is a correlation between the length and frequency of spoken or written words. Zipf's Law of Abbreviation has been shown to hold at the level of words in many languages, but it is unknown whether it holds at the level of characters. Characters' complexity is similar to word length, requiring more cognitive and motor effort from producing and processing the more complex symbols. We built a dataset of character complexity and frequency measures covering 27 different writing systems. Our results suggest that Zipf's Law of Abbreviation holds for every single system in our dataset – the more frequent characters have lower degrees of complexity and vice-versa. This result provides further evidence of optimization mechanisms shaping communication systems.

1. Introduction

In his pioneering work, George Kingsley Zipf observed that more frequent words tend to be shorter – a principle known as Zipf's Law of Abbreviation (Zipf, 1949). This law-like relation has since been confirmed on data coming from many different languages (see, for example, Piantadosi, Tily, & Gibson, 2011; Bentz & Ferrer-i-Cancho, 2016). Additionally, Zipf's Law of Abbreviation has been shown to arise in artificial language learning experiments (Kanwal, Smith, Culbertson, & Kirby, 2017) and in communication systems of different species (see Ferrer-i-Cancho & Lusseau, 2009; Semple, Hsu, & Agoramoorthy, 2010, amongst many). This effect is usually explained in terms of minimizing cumulative production cost: speakers intend to reduce their average articulation effort. This results in a reduction in the number of sounds that are pronounced overall.

Since scripts can be thought of as a communication system, which maps written characters to phonemes or syllables (Morin, Kelly, & Winters, 2020), we might expect that the same effect will hold for individual characters. Characters do not have length, unlike words. But visual complexity shares several relevant properties with word length. Complex characters take more effort to write and read,

just like long words are more effortful for speakers and hearers. Miton and Morin (2021) suggest that characters in writing systems are under similar pressures as words in spoken languages. Rovenchak and Vydrin (2010) have found a negative correlation between the complexity of characters and their frequency in the Nko writing system (West Africa). Similar results were reported for the Vai writing system in (Rovenchak, Mačutek, & Riley, 2008), and Mandarin Chinese characters (Shu, Chen, Anderson, Wu, & Xuan, 2003). The small number of studies that have tested this hypothesis shows a negative association between the complexity and frequency of characters – consistent with Zipf's Law of Abbreviation. However, no large-scale comparative testing was done in this domain. This study fills this gap by using a dataset that consists of 27 writing systems and computational, automated, and replicable measures to quantify character complexity. This is different from the idiosyncratic methods, primarily based on stroke counts used in previous studies (see Changizi & Shimojo, 2005 for an example of such methodology). Large-scale cross-linguistic corpora and datasets containing data on character complexity such as GraphCom (Chang, Chen, & Perfetti, 2018) or the dataset introduced in (Miton & Morin, 2021) offer a way to conduct cross-linguistic research on Zipf's Law of Abbreviation in writing systems.

We hypothesize that writing systems will follow Zipf's Law of Abbreviation. As most writing systems are largely based on handwritten characters shaped by centuries of reproduction, a minimization of the cumulative production cost is expected. There is evidence, based on high-quality but limited data, that writing systems can become less complex over time (Kelly, Winters, Miton, & Morin, 2021), indicating an overall trend towards simplification. Reduction of the complexity of symbols was also observed in interactive graphical communication experiments (Tamariz & Kirby, 2015; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007) We expect that frequency should negatively correlate with complexity, i.e., more frequent characters should have become simpler visually due to pressures to minimize production cost over the course of their constant reproduction.

2. Data

2.1. *Dataset description*

The dataset used in this study combines complexity measures from (Miton & Morin, 2021) and frequencies for each character. The complexity measures for every character include perimetric complexity and algorithmic complexity. Perimetric complexity was introduced in (Attneave & Arnoult, 1956), and it is defined as the inked surface divided by the perimeter of this inked surface. Miton and Morin (2021) computed this complexity measure using an implementation proposed in (Watson, 2012). Additionally, Pelli, Burns, Farell, and Moore-Page (2006) demonstrate that perimetric complexity closely correlates with human visual processing effort. Algorithmic complexity is the number of bytes needed to store

a compressed version of the character. The frequencies of individual characters were obtained from biblical texts extracted from `bible.com`. If data on the desired writing system was not available on `bible.com`, we used data from (Bentz & Ferrer-i-Cancho, 2016), which is also based on biblical texts, instead. Additionally, for Shavian, we extracted the data from `shavian.info/books/`. The texts were preprocessed to remove the punctuation, numbers, and characters that do not belong to the writing system of interest (for instance, Latin characters were removed). The character counts were computed from preprocessed texts and converted to relative frequencies by dividing each count by the sum of counts for the given writing system. Additionally, as the distribution of relative frequencies is highly skewed, these values were log-transformed. This transformation did not affect the results we present here.

The resulting dataset has 27 writing systems. The median corpus size (in characters) is 711,785, with the smallest values for Shavian (97,566 characters) and the largest for Thai (2,942,793 characters). The median number of characters is 42; the writing system with the lowest number of characters is Syriac (22 characters), and the largest writing system is Ethiopic (251 characters). Our dataset consists of four abjads, fourteen abugidas, five alphabets, one featural system, and four syllabaries (1560 characters in total). The geographic distribution of the writing systems in the dataset is shown in Fig. 1:

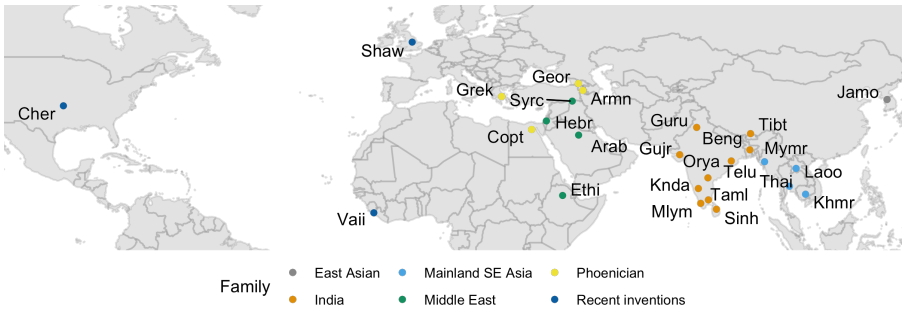

Figure 1. Geographic distribution of the writing systems in the database, annotated with the ISO 15924 codes

2.2. *Inclusion criteria*

We included writing systems in our dataset based on several criteria:

1. A writing system was included if it had available Unicode-encoded text files.

2. It is possible to identify one main language for which the writing system was

designed. The Latin and Devanagari writing systems had to be excluded because each of them is used to encode a multiplicity of languages, and each was substantially transformed to encode these languages.

3. The writing system is not combined with other writing systems. For instance, Limbu writing consists of both Devanagari and Limbu characters. Therefore, it was excluded from the sample. However, if the instances of such use are not common, these cases would be kept. For instance, Korean writing today is overwhelmingly based on the Hangul writing system, with only occasional use of Hanja (Mandarin Chinese characters). We focused on analyzing Hangul and disregarded Hanja.

4. Writing systems with less than a hundred thousand characters in the available texts were excluded.

3. Analysis

The proposed hypothesis was tested using a mixed-effect linear regression[1] predicting a character's complexity from its relative frequency (fixed effect FREQUENCY) and the writing system to which the character belongs (random effect Writing System). The model has both random slopes and random intercepts for each writing system and was run separately on our algorithmic complexity measure and on our perimetric complexity measure.

First, we measured the null model's Akaike information criterion (AIC). (The null model included only the random effect of Writing Systems.) We compared the null model's AIC with the full model's AIC. The full model included a fixed effect for FREQUENCY and the random effect of Writing Systems, with random slopes for each writing system. If the full model has lower AIC values than the null model (with the conventional threshold being $\Delta\text{AIC} > 2$), that means that the former is more informative than the latter. For perimetric complexity, the ΔAIC value is equal to 172.8, and for algorithmic complexity, this value corresponds to 146.1, meaning that they are both more informative than their respective null models. The β coefficients for relative frequency in the perimetric complexity (-2.40, 95% CI: [-1.76, -3.07]) and in the algorithmic complexity models (-26.6 95%CI: [-34.15, -19.17])) are both negative. These values of the coefficients indicate that the higher the frequency is, the less complex is the character, as illustrated in Fig. 2:

[1] We used the lme4 R-package to fit our models (Bates, Mächler, Bolker, & Walker, 2014)

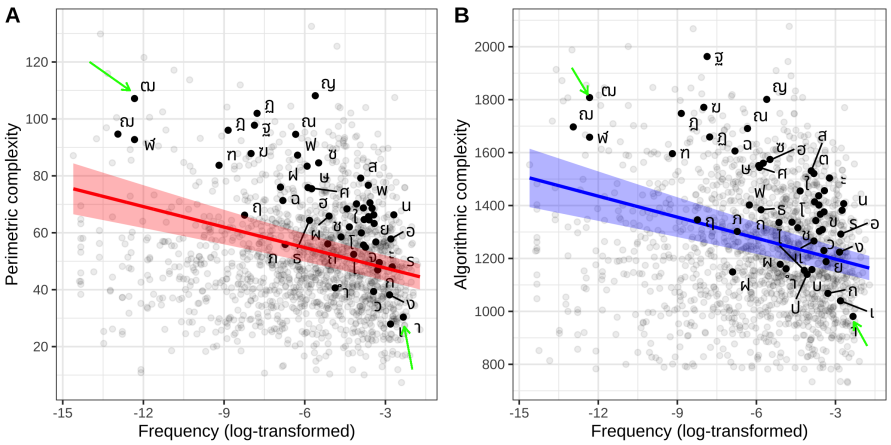

Figure 2. Predictions from perimetric complexity (A) and algorithmic complexity (B) mixed-effect linear regression models for all the scripts combined. Each point corresponds to a unique character. Shaded areas represent the 95% confidence interval for the predictions.

In Fig. 2, each dot represents an individual character. We added Thai characters to each plot for illustrative purposes. The arrows point to the most complex and less complex characters. Additionally, our results suggest that the effects hold for each writing system and are not an artifact from the aggregated data, see Fig. 3:

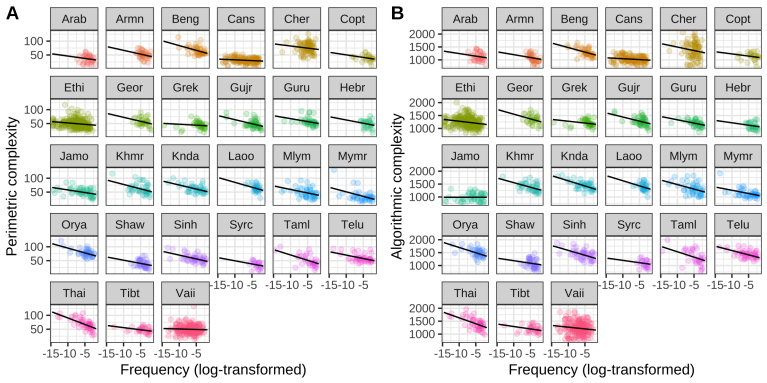

Figure 3. Predictions from perimetric complexity (A) and algorithmic complexity (B) models for individual writing systems. Each point represents an individual character, and each subplot corresponds to an individual writing system (annotated by its ISO 15924 code).

4. Conclusions

Using mixed effect linear regression models, we show that Zipf's law of abbreviation holds on all of the individual writing systems in our dataset, not just on the aggregated data taken as a whole, validating our preregistered predictions. These results hold for both our complexity measures and suggest that the law of abbreviation holds in a large variety of writing systems at the level of characters.

Zipf's law of abbreviation was also attested in iterative learning experiments (Kanwal et al., 2017), where the authors suggested that the need for efficient and accurate communication yields the inverse relationship between word frequency and its length. Since our results support the presence of Zipf's law of abbreviation in written communication, this suggests that the results from (Kanwal et al., 2017) can be expanded to graphic communication. Overall, such evidence for Zipf's law of abbreviation in writing supports the idea that optimizing production and reception costs would have been an important factor in the evolution of spoken and graphic communication.

Supplementary Materials

The preregistration for this study can be found here: `https://osf.io/ydr3n`

A Git-Hub repository will all of the data and code can be found here: `https://github.com/alexeykosh/2021-slojnost-project`

Acknowledgments

We thank Christian Bentz and Ramon Ferrer-i-Cancho for their helpful advice on this project, and for sharing their data.

References

Attneave, F., & Arnoult, M. D. (1956). The quantitative study of shape and pattern perception. *Psychological bulletin*, *53*(6), 452.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the leiden workshop on capturing phylogenetic algorithms for linguistics* (pp. 1–4).

Chang, L.-Y., Chen, Y.-C., & Perfetti, C. A. (2018). Graphcom: A multidimensional measure of graphic complexity applied to 131 written languages. *Behavior research methods*, *50*(1), 427–449.

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1560), 267–275.

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

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive science*, *31*(6), 961–987.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kelly, P., Winters, J., Miton, H., & Morin, O. (2021). The predictable evolution of letter shapes: An emergent script of west africa recapitulates historical change in writing systems. *Current Anthropology*, *62*(6), 000–000.

Miton, H., & Morin, O. (2021). Graphic complexity in writing systems. *Cognition*, *214*, 104771.

Morin, O., Kelly, P., & Winters, J. (2020). Writing, graphic codes, and asynchronous communication. *Topics in cognitive science*, *12*(2), 727–743.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision research*, *46*(28), 4646–4674.

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

Rovenchak, A., Mačutek, J., & Riley, C. (2008). *Distribution of complexities in the vai script.*

Rovenchak, A. A., & Vydrin, V. (2010). *Quantitative properties of the nko writing system.*

Semple, S., Hsu, M. J., & Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters*, *6*(4), 469–471.

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school chinese: Implications for learning to read. *Child development*, *74*(1), 27–47.

Tamariz, M., & Kirby, S. (2015). Culture: copying, compression, and conventionality. *Cognitive science*, *39*(1), 171–183.

Watson, A. B. (2012). *Perimetric complexity of binary digital images: Notes on calculation and relation to visual complexity* (Vol. 14; Tech. Rep.).

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books.