Zipf's Law of Abbreviation holds for individual characters across a broad range of writing systems

Alexey Koshevoy, Helena Miton, Olivier Morin

Cognition 238 (2023) 105527

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short communication

Zipf's Law of Abbreviation holds for individual characters across a broad range of writing systems

Alexey Koshevoy [a,b,c,*], Helena Miton [d], Olivier Morin [b,c]

[a] Laboratoire de Psychologie Cognitive, Aix-Marseille Université, CNRS, 13003 Marseille, France
[b] Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, UMR 8129, 75005 Paris, France
[c] Minds and Traditions Group, Max Planck Institute for Geoanthropology, 07745 Jena, Germany
[d] Santa Fe Institute, Santa Fe, NM 87501, USA

ARTICLE INFO

Dataset link: https://osf.io/h8mqk/?view_only=e8782e472b9b4af4994900cd74666685

Keywords:
Writing systems
Zipf's Law of Abbreviation
Complexity
Frequency effects

ABSTRACT

Zipf's Law of Abbreviation – the idea that more frequent symbols in a code are simpler than less frequent ones – has been shown to hold at the level of words in many languages. We tested whether it holds at the level of individual written characters. Character complexity is similar to word length in that it requires more cognitive and motor effort for producing and processing more complex symbols. We built a dataset of character complexity and frequency measures covering 27 different writing systems. According to our data, Zipf's Law of Abbreviation holds for every writing 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

1.1. Zipf's Law of Abbreviation

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 (or Zipf's Law of Brevity) (Zipf, 1949). This principle has since been found in data from many different languages. For instance, Bentz and Ferrer-i-Cancho (2016) showed that Zipf's Law of Abbreviation holds on written words based on data from 986 different languages, accounting for 13% of the world's languages (see also Petrini, Casas-i Muñoz, Cluet-i Martinell, Wang, Bentz et al. (2022) for spoken language data). Piantadosi, Tily, and Gibson (2011) demonstrated that this law holds even when information content[1] is taken instead of frequency. This law has also been found in other species' communication systems (Favaro et al., 2020; Heesen, Hobaiter, Ferrer-i Cancho, & Semple, 2019; Huang, Ma, Ma, Garber, & Fan, 2020; Semple, Hsu, & Agoramoorthy, 2010). Additionally, Zipf's Law of Abbreviation has been shown to spontaneously arise in communication games involving artificial languages (Kanwal, Smith, Culbertson, & Kirby, 2017; Krauss & Weinheimer, 1964). The omnipresence of this principle resulted in claims of it being an essential property of communication systems (Ferrer-i-Cancho, Hernández-Fernández, Lusseau, Agoramoorthy, Hsu et al., 2013; Zipf, 1949).

According to Zipf (1949), languages are subject to two opposing pressures: "speaker's economy" and "auditor's economy". The former refers to the speaker's desire to decrease the size of the lexicon and unify words to reduce production effort (unification), while the latter refers to the auditor's need for a large vocabulary giving each individual meaning a corresponding word, leading to a decrease in the effort required to identify the correct meaning. (diversification) (Zipf, 1949, p. 21). These opposing pressures result in some words becoming more frequent than others. Additionally, Zipf showed that more frequent words tend to be shorter than less frequent words. This observation follows from the Principle of Least Effort, which suggests that living organisms tend to minimize their effort on average. It implies reducing their average articulation effort by pronouncing fewer sounds overall, resulting in a reduction in the number of sounds pronounced (minimization of the cumulative production cost). For the sake of brevity, we will refer to "speaker's economy" as pressure for **efficiency**, and to "auditor's economy" as pressure for **communicative accuracy** (following Kanwal et al. (2017), Kemp and Regier (2012)). This approach is embedded into all variable-length coding algorithms, such as Huffman coding (Huffman, 1952) or Morse code. Morse code can be thought of as an example of a purposeful minimization of the cumulative production cost. S. Morse and A. Vail chose the length of each signal inversely proportional to the frequency of the corresponding English letter (Gleick, 2011).

* Correspondence to: Institut Jean Nicod; UMR 8129 Pavillon Jardin, Ecole Normale Supérieure; 29, rue d'Ulm, F-75230 Paris cedex 05, France.
E-mail address: alexey.koshevoy@univ-amu.fr (A. Koshevoy).

[1] Predictability of a word given the context.

https://doi.org/10.1016/j.cognition.2023.105527
Received 12 October 2022; Received in revised form 12 June 2023; Accepted 14 June 2023
Available online 24 June 2023

However, several results have recently challenged this idea. For example, Clink, Ahmad, and Klinck (2020) found no evidence of Zipf's Law of Abbreviation in gibbon calls, and Bezerra, Souto, Radford, and Jones (2011) reported that it is not present in the calls of golden-backed uakaris either. Furthermore, Miton and Morin (2019) have found no evidence of Zipf's Law of Abbreviation in European heraldry. They argued that one of the preconditions for a graphic code to obey this law is that it should lack iconicity, which does not hold for heraldry. Overall, these conflicting results show that more communicative systems should be examined for the presence of the Law of Abbreviation to address the considerations of it being universal.

1.2. Writing systems

As writing systems can be thought of as communication system, which map written characters to phonemes, syllables, or morphemes (Coulmas, 2003), we may expect that the same effect will hold for individual characters. Characters do not have length, unlike words. Nevertheless, the visual complexity of characters shares several relevant properties with spoken word length. Complex characters take more effort to write (Lin, Chao, Hsu, Hsu, Chen et al., 2019) and read, just like long words are more effortful for speakers and hearers. (Tamaoka & Kiyama, 2013) show that in the low frequency band, Kanji characters take more time to process depending on their visual complexity, as well as the accuracy of identification is inversely proportional to visual complexity. Compare the Greek letters σ and ψ. ψ takes at least two strokes to be written, while only one is required for σ. Characters in writing systems are under similar pressures as words in spoken languages (Miton & Morin, 2021). To make an analogy with Zipf's reasoning, writing systems can be thought of as being subject to a pressure for efficiency aimed at reducing the effort required to produce and process individual characters, and a pressure for communicative accuracy, which aims at increasing the ease, for readers, of retrieving the linguistic units corresponding to individual characters. Therefore, more frequent characters are expected to become less complex than less frequent characters, while still preserving sufficient complexity to ensure distinguishability (see Han, Kelly, Winters, and Kemp (2022)). Additionally, writing systems follow the requirement of lacking iconicity (Miton & Morin, 2019), indicating that writing systems should follow Zipf's Law of Abbreviation.

Zipf's law of Abbreviation has been found in several individual writing systems. For instance, in the Nko writing system (West Africa), there is a negative correlation between the complexity of characters and their frequency (Rovenchak & Vydrin, 2010). Similar results were reported for the Vai writing system (Rovenchak, Mačutek, & Riley, 2008), and Mandarin Chinese characters (Shu, Chen, Anderson, Wu, & Xuan, 2003). The few studies that have tested this hypothesis show a negative correlation 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 approach differs from the idiosyncratic methods primarily based on stroke counts used in previous studies (see Changizi and Shimojo (2005) for an example of such methodology).

1.3. Hypothesis

Since a clear parallel can be drawn between writing systems and other communicative systems that show the Law of Abbreviation, we can hypothesize that writing systems are subjected to 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 for minimization of graphic complexity in the evolution of writing systems (Kelly, Winters, Miton, & Morin, 2021) and in interactive graphical communication experiments (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Tamariz & Kirby, 2015), suggesting that, when graphic shapes are highly complex, a trend towards simplification can be expected on grounds of efficiency as long as it does not conflict with the distinctiveness of shapes (which would hinder communicative accuracy). Given this, we expect that frequency should negatively correlate with complexity, i.e., more frequent characters should have become simpler visually due to the trade-off between the pressures for efficiency and communicative accuracy.

2. Materials

2.1. Complexity measures

The dataset used in this study combines complexity measures from Miton and 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 and Arnoult (1956), and is defined as follows:

$$C = \frac{P^2}{4\pi A} \tag{1}$$

In (1), C is perimetric complexity, P is the sum of the inside and outside perimeter of the inked surface, and A is the total area. Miton and Morin (2021) computed this complexity measure using an implementation proposed in Watson (2012). Several studies have indicated that perimetric complexity correlates with human visual processing and production effort since it is linked to the stroke length required to draw a character (Chang, Plaut, & Perfetti, 2016; Pelli, Burns, Farell, & Moore-Page, 2006).

The second complexity measure used in this study is algorithmic complexity. Algorithmic complexity corresponds to the number of bytes needed to store a compressed version of the character. This measure has been previously used in Tamariz and Kirby (2015) and Han et al. (2022) for visual complexity. Algorithmic complexity can be interpreted as the length of the shortest computer program needed to restore the initial image. For instance, perimetric complexity for σ is 21.01 and the perimetric complexity for ψ is 75.6. The algorithmic complexity for these characters corresponds to 997 and 1295, respectively. Algorithmic and perimetric complexity measures are strongly positively correlated in our data ($r(1560) = 0.797$, $p < .001$).

2.2. Data sources

The frequencies of individual characters were obtained from biblical texts extracted from www.bible.com. If data on the desired writing system was not available on www.bible.com, we used data from Bentz and Ferrer-i-Cancho (2016), which is based on the Parallel Bible Corpus (Mayer & Cysouw, 2014). Additionally, for Shavian, we extracted the data from www.shavian.info/books/. The texts were preprocessed to remove the punctuation, numbers, and characters that do not belong to the writing system of interest. 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. Although Piantadosi et al. (2011) claims that predictability in context is a better predictor for the word length than frequency, several researchers have since challenged these results (see, for example, Koplenig, Kupietz, and Wolfer (2022), Levshina (2022), Meylan and Griffiths (2021)). This, together with the small sizes of the corpora used in our study, which influences the accuracy of measures like predictability, influences the decision to include frequency as the main predictor in the study instead of predictability in context.

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

2.3. Inclusion criteria

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

- It had available Unicode-encoded text files.
- It was 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 is substantially transformed to encode these languages.
- The writing system was 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 example, Korean writing today is overwhelmingly based on the Hangul writing system, with only occasional use of Hanja (Chinese characters). We focused on analyzing Hangul and disregarded Hanja.
- Writing systems with less than a hundred thousand characters of available text were excluded.

2.4. Dataset description

The resulting dataset includes 1560 characters from 27 writing systems. Our dataset consists of four abjads, fourteen abugidas, five alphabets, one featural system, and four syllabaries. This dataset covers all existing types of 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 per writing system is 42; the writing system with the lowest number of characters is Syriac (22 characters), and the largest writing system is Ethiopic (251 characters). Family is a category based on each script's geography and ancestry which is determined following Daniels and Bright (1996), Miton and Morin (2021). The geographic distribution of the writing systems in the dataset is shown in Fig. 1:

3. Analysis

The proposed hypothesis was tested using a mixed-effect linear regression predicting character's complexity from its relative frequency (fixed effect FREQUENCY) and the writing system to which the character belongs (random effect WRITING SYSTEM). This model has both random slopes and random intercepts for each writing system and was run on algorithmic complexity and perimetric complexity data separately, resulting in two separate models for each corresponding measure. In every analysis below, the results come from the two models associated with their respective complexity measures. We used the lme4 R-package to fit our models (Bates, Mächler, Bolker, & Walker, 2014).

3.1. Initial models

First, we measured the null model's Akaike information criterion (AIC). The null model included only the random effect of WRITING SYSTEM. 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 slopes and intercepts for WRITING SYSTEM. If the full model has lower AIC values than the null model (with the conventional threshold being $\Delta AIC > 2$), this indicates that the full model is more informative than the null model. For perimetric complexity, the ΔAIC value is equal to 172.8. For algorithmic complexity, this value corresponds to 152.6, meaning that they are both more informative than their respective null models. The conditional.[2] R^2 for the perimetric complexity model is equal to 0.53, and the R^2 for the algorithmic complexity model is equal to 0.48 The β coefficients for relative frequency in the perimetric complexity (−2.4, 95% CI: [−3.07, −1.76]) and in the algorithmic complexity models (−28.05 95% CI: [−35.28, −21.2]) are both negative.[3] These values of the coefficients indicate that with higher frequencies, corresponding characters become less complex, as illustrated in Fig. 2.

3.2. Model with nested family

In addition to the analysis above, we have also controlled for FAMILY by nesting each writing system inside its respective family. When comparing the AIC values for the null model (a model only containing the random effect of WRITING SYSTEM nested in FAMILY) with the full model's AIC (a model containing the complexity measure as the fixed effect), the ΔAIC is equal to 172.8 for perimetric complexity, and 152.59 for algorithmic complexity models. The β coefficients in both models are also negative: (−2.4, 95% CI: [−3.07, −1.76]) for perimetric complexity and (−28.05 95% CI: [−35.28, −21.2]) for algorithmic complexity. Controlling for FAMILY does not affect our predictions.

3.3. Individual scripts

Overall, our results suggest that the effects hold for each writing system and are not an artifact from the aggregated data, see Fig. 3.

Additionally, the random slope values showing the effect of relative frequency on character complexity as it varies for each script support our claim about the effect holding for each writing system in the sample, see Fig. 4.

[2] Variance explained by both fixed and random factors.

[3] To provide more evidence for the robustness of our result, we included the Spearman's correlation coefficients for each script in the supplementary materials.

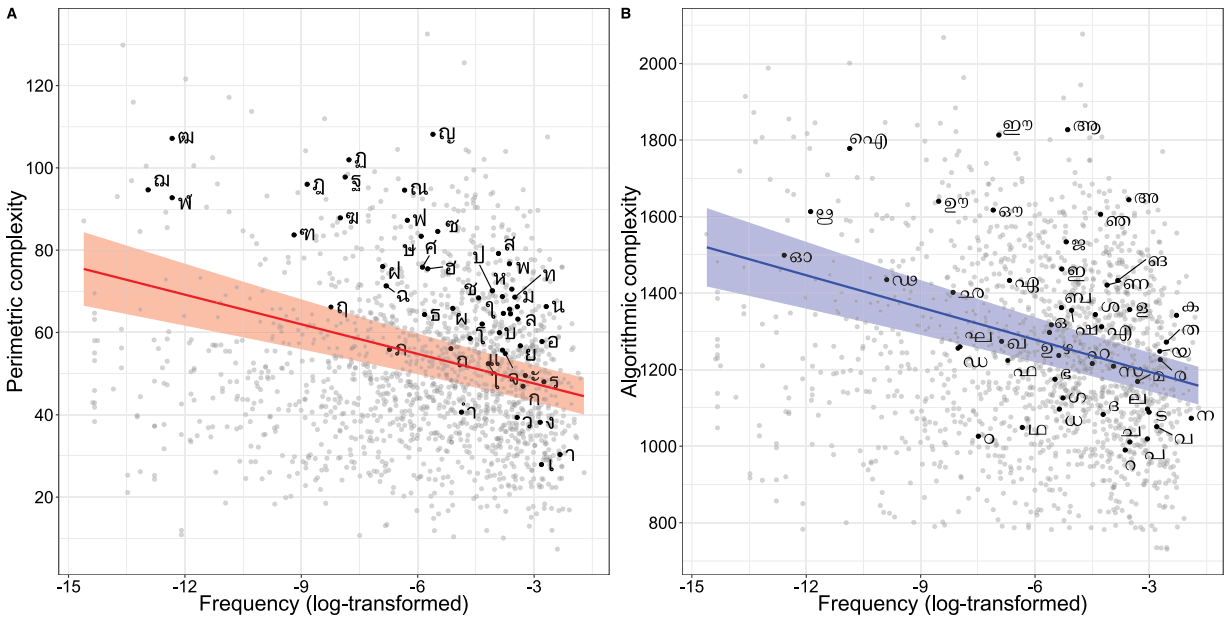

Fig. 2. Frequency and complexity measures for all the scripts combined. Each dot represents an individual character ($n = 1560$ letters from 27 scripts). The colored lines represent the averaged predictions from the mixed-effect linear regression models. Each point corresponds to a unique character measured for perimetric complexity (A) and algorithmic complexity (B). Red and blue shaded areas represent the 95% confidence interval for the predictions. We added Thai (A) and Burmese characters (B) for illustrative purposes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

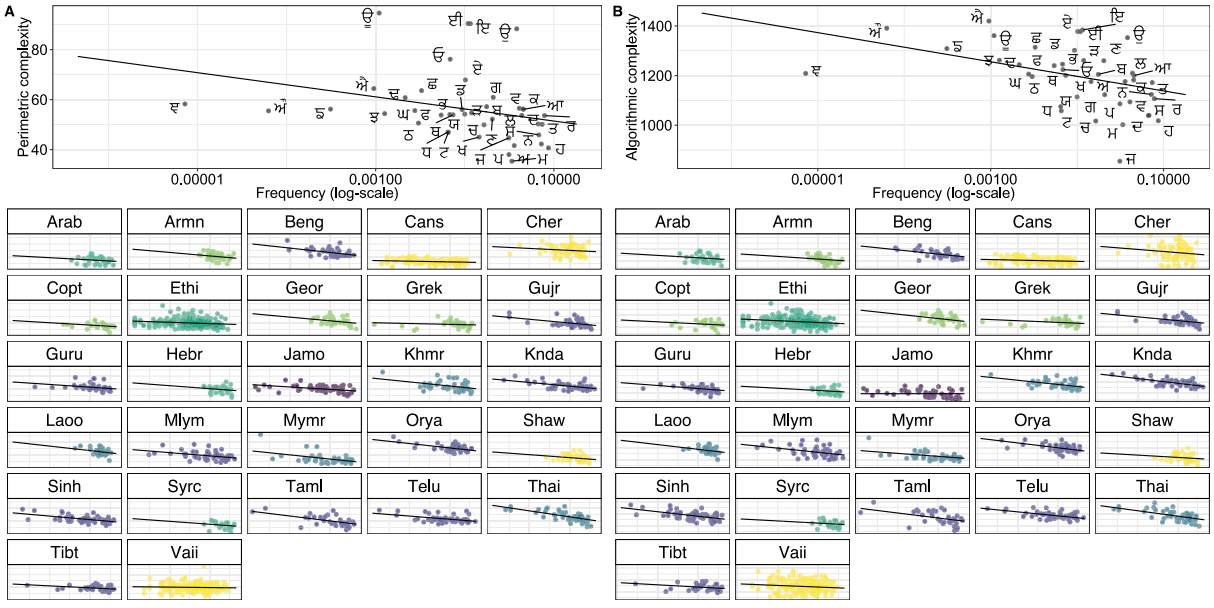

Fig. 3. Frequency and complexity measures for each script extracted from the initial models. Black lines represent the predictions from perimetric complexity (A) and algorithmic complexity (B) models. On the top plane, the Gurumukhi writing system is used for illustrative purposes. On the bottom plane, each point represents an individual character, and each subplot corresponds to an individual writing system (annotated by its ISO 15924 code). Colors indicate individual writing systems. All of the x and y axes of the bottom plots are identical to the axes of the respective plots on the top part of the figure. Colors correspond to the family attribution of the script (see legend in Fig. 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 4, every script has a negative random slope value in both perimetric (A) and algorithmic complexity (B) models. Altogether, these results support our hypothesis, indicating the presence of Zipf's law of Abbreviation in each of the 27 writing systems that we have included in our dataset.

4. Discussion

Zipf's Law of Abbreviation is believed to be an essential property of communication systems. However, it has seldom been tested for graphic communication systems. The length of written words reflects their phonological length, and is thus widely used as a proxy for it. The complexity of individual letters, on the other hand, is decoupled from phonological complexity. Using mixed effect linear regression models, we show that Zipf's Law of Abbreviation holds for all of the individual writing systems in our dataset, not just on the aggregated data taken as a whole, validating our preregistered predictions. This result hold for both of our complexity measures and suggest that the law of Abbreviation holds at the level of individual characters in a large variety of writing systems.

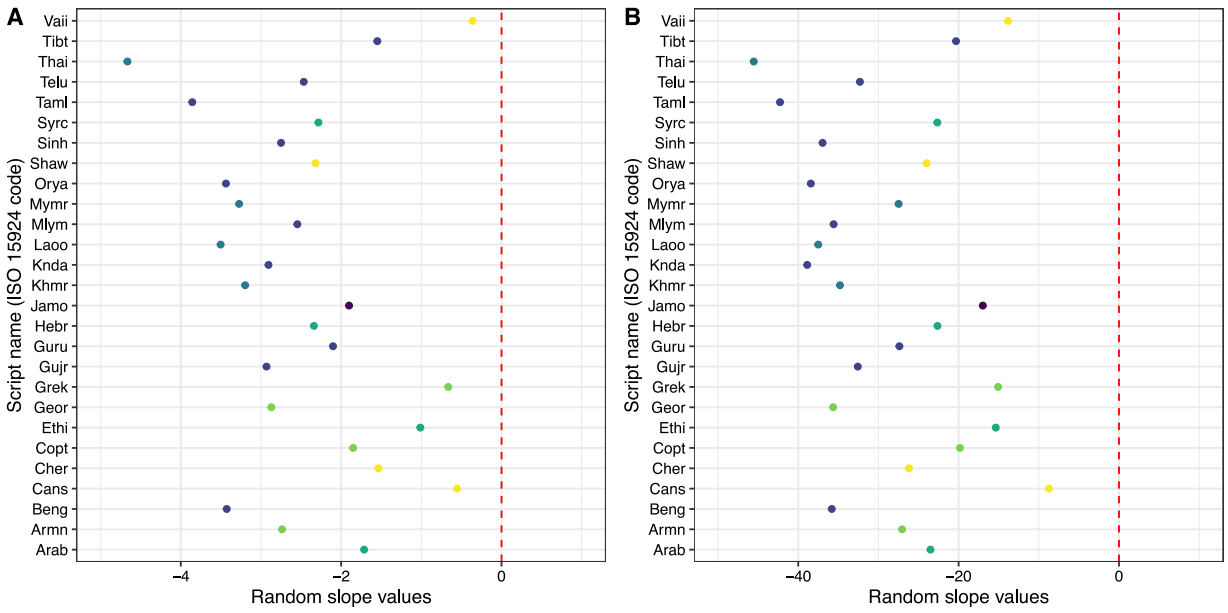

Fig. 4. Random slope values obtained from the mixed-effects model for each writing system in the database. Points correspond to the slope coefficients of the effect of relative frequency on perimetric (A) and algorithmic complexity (B) for each script. Red dotted lines correspond to the slope value of zero. Colors correspond to the family attribution of the script (see legend in Fig. 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Zipf's Law of Abbreviation as a general property of writing systems

In this study, we have used automated, computational, and replicable complexity measures, as compared to previous literature that mostly relied on idiosyncratic or manual measures. Moreover, our predictions were tested on a broad range of writing system types (abjads, abugidas, alphabets, featural systems, and syllabaries). The only major typological exception is logo-syllabic systems, but other studies showed the Law of Abbreviation to apply there as well — see Shu et al. (2003) for Chinese. Since Zipf's Law of Abbreviation has been found across all of the writing systems in our dataset, and it holds for both complexity measures, it hints at the possible universality of this law for written communication.

These results further support the idea that this law arises from a trade-off between pressures for efficiency and communicative accuracy. As there is a clear parallel between a combined length of strokes needed to produce an individual character and word length, a minimization of the cumulative production cost is expected to be at play. The same holds for perception — more complex characters take more visual processing effort. Additionally, characters also need to be distinguished from each other; therefore, a degree of complexity is required (Han et al., 2022; Miton & Morin, 2021). For instance, Tamaoka and Kiyama (2013) showed the importance of visual complexity in the processing of low-frequency Kanji characters as compared to high-frequency ones, suggesting that pressure for communicative accuracy is present in scripts. Since efficiency and communicative accuracy are both identifiable in writing systems, and we have found Zipf's Law of Abbreviation to be present, this result further supports Zipf's idea that this law is a result of a trade-off between these two forces.

4.2. Implications for the study of communication systems

The Law of Abbreviation is attested not only in spoken language but also in the communicative systems of other species and in writing systems, as shown in this study. This possibly implies that the efficiency and communicative accuracy trade-offs are essential properties that shape every communication system that satisfies certain conditions. In Ferrer-i-Cancho et al. (2013), the authors suggested that minimization of cumulative production effort is a central property of human behavior in general and communication systems in particular. Writing fulfills one of the conditions previously identified for respecting ZLA: it lacks iconicity (Morin, 2022). On the other hand, writing has historically been a costly and prestigious cultural practice, an occasion to display virtuosity and skill through intricate shapes. The relative scarcity of literate people and the inertia of institutions could also have stood in the way of the simplification processes necessary for ZLA to occur. It is all the more remarkable that ZLA is as clearly evident for individual written letters as for spoken words.

CRediT authorship contribution statement

Alexey Koshevoy: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Helena Miton:** Conceptualization, Methodology, Software, Writing – review & editing, Supervision. **Olivier Morin:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code are stored here: https://osf.io/h8mqk/?view_only=e8782e472b9b4af4994900cd74666685.

Acknowledgments

We thank Christian Bentz and Ramon Ferrer-i-Cancho for their helpful advice on this project and for sharing their data. This study was supported by the EUR FrontCog grant ANR-17-EURE-0017 and ANR-10-IDEX-0001-02 to PSL. HM acknowledges the support of a Santa Fe Institute Complexity Fellowship.

Appendix A. Supplementary data

We preregistered the predictions and the data collection protocols. The pre-registration and the code for replicating the results are available at https://osf.io/h8mqk/?view_only=e8782e472b9b4af4994900cd74666685.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105527.

References

Attneave, Fred, & Arnoult, Malcolm D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin*, *53*(6), 452.

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bentz, Chris, & Ferrer-i-Cancho, Ramon (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). University of Tübingen.

Bezerra, Bruna M., Souto, Antonio S., Radford, Andrew N., & Jones, Gareth (2011). Brevity is not always a virtue in primate communication. *Biology Letters*, *7*(1), 23–25.

Chang, Li-Yun, Plaut, David C., & Perfetti, Charles A. (2016). Visual complexity in orthographic learning: Modeling learning across writing system variations. *Scientific Studies of Reading*, *20*(1), 64–85.

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

Clink, Dena J., Ahmad, Abdul Hamid, & Klinck, Holger (2020). Brevity is not a universal in animal communication: Evidence for compression depends on the unit of analysis in small ape vocalizations. *Royal Society Open Science*, *7*(4), Article 200151.

Coulmas, Florian (2003). *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press.

Daniels, Peter T., & Bright, William (1996). *The world's writing systems*. Oxford University Press on Demand.

Favaro, Livio, Gamba, Marco, Cresta, Eleonora, Fumagalli, Elena, Bandoli, Francesca, Pilenga, Cristina, Isaja, Valentina, Mathevon, Nicolas, & Reby, David (2020). Do penguins' vocal sequences conform to linguistic laws? *Biology Letters*, *16*(2), Article 20190589.

Ferrer-i-Cancho, Ramon, Hernández-Fernández, Antoni, Lusseau, David, Agoramoorthy, Govindasamy, Hsu, Minna J., & Semple, Stuart (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578.

Garrod, Simon, Fay, Nicolas, Lee, John, Oberlander, Jon, & MacLeod, Tracy (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Gleick, James (2011). *The information: A history, a theory, a flood*. Vintage.

Han, Simon J., Kelly, Piers, Winters, James, & Kemp, Charles (2022). Simplification is not dominant in the evolution of Chinese characters. *Open Mind*, *6*, 264–279.

Heesen, Raphaela, Hobaiter, Catherine, Ferrer-i Cancho, Ramon, & Semple, Stuart (2019). Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1896), Article 20182900.

Huang, Mingpan, Ma, Haigang, Ma, Changyong, Garber, Paul A., & Fan, Pengfei (2020). Male gibbon loud morning calls conform to Zipf's law of brevity and Menzerath's law: Insights into the origin of human language. *Animal Behaviour*, *160*, 145–155.

Huffman, David A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

Kanwal, Jasmeen, Smith, Kenny, Culbertson, Jennifer, & Kirby, Simon (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, Piers, Winters, James, Miton, Helena, & Morin, Olivier (2021). The predictable evolution of letter shapes: An emergent script of West Africa recapitulates historical change in writing systems. *Current Anthropology*, *62*(6), 000.

Kemp, Charles, & Regier, Terry (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Koplenig, Alexander, Kupietz, Marc, & Wolfer, Sascha (2022). Testing the relationship between word length, frequency, and predictability based on the german reference corpus. *Cognitive Science*, *46*(6), Article e13090.

Krauss, Robert M., & Weinheimer, Sidney (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1), 113–114.

Levshina, Natalia (2022). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, *24*(2), 280.

Lin, Yu-Chen, Chao, Yen-Li, Hsu, Chieh-Hsiang, Hsu, Hsiao-Man, Chen, Po-Tsun, & Kuo, Li-Chieh (2019). The effect of task complexity on handwriting kinetics. *Canadian Journal of Occupational Therapy*, *86*(2), 158–168.

Mayer, Thomas, & Cysouw, Michael (2014). Creating a massively parallel Bible corpus. *Oceania*, *135*(273), 40.

Meylan, Stephan C., & Griffiths, Thomas L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, *45*(6), Article e12983.

Miton, Helena, & Morin, Olivier (2019). When iconicity stands in the way of abbreviation: No Zipfian effect for figurative signals. *Plos One*, *14*(8), Article e0220793.

Miton, Helena, & Morin, Olivier (2021). Graphic complexity in writing systems. *Cognition*, *214*, Article 104771.

Morin, Olivier (2022). The puzzle of ideography. *Behavioral and Brain Sciences*, 1–69.

Pelli, Denis G., Burns, Catherine W., Farell, Bart, & Moore-Page, Deborah C. (2006). Feature detection and letter identification. *Vision Research*, *46*(28), 4646–4674.

Petrini, Sonia, Casas-i Muñoz, Antoni, Cluet-i Martinell, Jordi, Wang, Mengxue, Bentz, Christian, & Ferrer-i Cancho, Ramon (2022). The optimality of word lengths. Theoretical foundations and an empirical study. arXiv preprint arXiv:2208.10384.

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

Rovenchak, Andrij A., Mačutek, Ján, & Riley, Charles (2008). Distribution of complexities in the vai script.

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

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

Shu, Hua, Chen, Xi, Anderson, Richard C., Wu, Ningning, & Xuan, Yue (2003). Properties of school Chinese: Implications for learning to read. *Child Development*, *74*(1), 27–47.

Tamaoka, Katsuo, & Kiyama, Sachiko (2013). The effects of visual complexity for Japanese kanji processing with high and low frequencies. *Reading and Writing*, *26*(2), 205–223.

Tamariz, Mónica, & Kirby, Simon (2015). Culture: Copying, compression, and conventionality. *Cognitive Science*, *39*(1), 171–183.

Watson, Andrew B. (2012). Perimetric complexity of binary digital images: Notes on calculation and relation to visual complexity. *Mathematica Journal*, *14*.

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