Efficiency in Writing Systems: Testing Zipf's Law of Abbreviation Across Letters, N-Grams, and Words

Ruimin Lyu (ruiminlyu@jiangnan.edu.cn)
School of Artificial Intelligence and Computer Science,Jiangnan University, 1800 Lihu Avenue, Binhu District Wuxi, Jiangsu 214122, China

Sihan Wang (762847914@qq.com)
School of Artificial Intelligence and Computer Science,Jiangnan University, 1800 Lihu Avenue, Binhu District Wuxi, Jiangsu 214122, China

Guoying Yang* (corresponding author, guoyingyang@jiangnan.edu.cn)
Jiangsu Key University Laboratory of Software and Media Technology under Human-Computer Cooperation, 1800 Lihu Avenue, Binhu District, Wuxi, Jiangsu 214122, China

Abstract

This study investigates Zipf's Law of Abbreviation (ZLA) across 155 writing systems, analyzing how visual complexity optimizes with information content at three linguistic levels: letters, n-grams, and words. Using perimetric and skeleton-length complexity metrics, we demonstrate that letters exhibit the strongest correlation ($\rho = 0.2 - 0.4$ in most languages), confirming their role as primary units of efficiency optimization. Larger units (n-grams/words) show weaker effects due to structural constraints. While alphabetic scripts (e.g., Latin-based) align robustly with ZLA, logographic (e.g., Chinese) and abugida (e.g., Kannada) systems reveal exceptions—some with near-zero or negative correlations—highlighting script-specific pressures like distinctiveness or historical preservation. Our findings refine ZLA by emphasizing visual (not just length-based) effort minimization and underscore letters as the fundamental locus of abbreviation effects. Limitations in script diversity and complexity metrics suggest future directions, including phylogenetic controls and perceptual complexity measures. This work advances the cross-linguistic study of writing system evolution under efficiency pressures.

Keywords: Zipf's Law of Abbreviation; perimetric complexity; skeleton length; information content; cross-linguistic analysis;

Introduction

Zipf's Law of Abbreviation (ZLA) posits that linguistic units with higher usage frequency tend to be shorter, reflecting an underlying efficiency principle in communication (Zipf, 1949). This principle has been extensively validated in word length distributions (Piantadosi et al., 2011; Kanwal, 2017&2018), phoneme systems (Bentz & Ferrer-i-Cancho, 2016; Tanida, 2023), and recently, in the visual complexity of individual letters (Koshevoy, Miton, & Morin, 2023). The latter study provided the first large-scale evidence that abbreviation pressures extend beyond spoken language to the structure of writing systems, demonstrating that more frequent letters tend to be visually simpler across 27 diverse scripts. This finding reinforced the notion that communicative efficiency shapes both phonological and visual aspects of language. However, the scope of this study left several open questions: **To what extent is this effect universal across languages? How does abbreviation operate at different levels of linguistic structure, such as individual letters, letter sequences (n-grams), and words?**

Building on these foundations, our study significantly extends previous work on writing system efficiency by introducing two key advancements. First, we expand the scope of analysis to over 150+ writing systems, providing a far broader empirical test of ZLA's applicability. Second, we systematically compare different linguistic units-letters, n-grams, and words-by quantifying the strength of their information-complexity relationships. While prior studies have confirmed an inverse correlation between frequency and complexity, we follow established information-theoretic approaches (Piantadosi et al., 2011; Ferrer-i-Cancho et al., 2013) by focusing on information content rather than raw frequency. Since information content, often measured as the negative logarithm of the probability of a information unit ($-\log_2(p)$) (Shannon, 1948), directly reflects the communicative "effect" of a unit, it provides a stronger test of ZLA's underlying efficiency principles than frequency alone.

A major challenge in studying abbreviation effects is the choice of linguistic units and effort metrics. Word length has been a central focus in past studies, but cross-linguistic variability in morphology, phonotactics, and orthographic depth complicates direct comparisons (Kuo et al., 2014). Letters, in contrast, provide a highly controlled unit of analysis, as they are purely visual symbols subject to direct constraints on recognition and production (Pelli et al., 2006). Moreover, empirical research has demonstrated that letter complexity directly affects reading speed (Abdelhadi et al., 2011; Ibrahim et al., 2002), perceptual span (Wang et al., 2014; Zhu et al., 2019), and handwriting efficiency (Chan & Lee, 2005; Zhang & Chen, 2017; Chow et al., 2003), reinforcing the importance of visual load in script optimization. However, writing systems do not operate solely at the level of individual letters-larger linguistic units such as letter sequences (n-grams) and words may also exhibit abbreviation effects. By systematically comparing information-complexity relationships at these three levels, we test whether letters provide the strongest empirical

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

support for ZLA, reinforcing the idea that abbreviation effects operate most clearly at the smallest functional unit of writing.

Our hypothesis predicts that across a diverse sample of writing systems, letters will exhibit the strongest positive linear correlation between information content and visual complexity, compared to n-grams and words. If confirmed, this would offer robust cross-linguistic support for the abbreviation principle, while also clarifying the linguistic level at which optimization pressures are most pronounced. Moreover, by explicitly defining effort in terms of visual complexity, our study contributes a clearer theoretical understanding of the cognitive and motoric constraints governing script evolution. In doing so, we refine Zipf's original formulation of abbreviation effects, providing a more precise interpretation of the "effort" minimized in communicative efficiency.

To test this, we analyze a large-scale cross-linguistic dataset spanning over 155 writing systems, covering alphabetic, abugida, and syllabic scripts. We measure information content based on Shannon entropy (Shannon 1948) and visual complexity using perimetric complexity (Attneave & Arnoult, 1956; Watson, 2012). By systematically comparing information-complexity correlations across letters, n-grams, and words, we aim to determine which linguistic unit best conforms to abbreviation effects and why.

Our study makes several contributions to the study of linguistic efficiency and script evolution. First, by significantly expanding the range of languages analyzed, we greatly extend the empirical generalizability of Koshevoy et al.'s (2023) findings, confirming whether abbreviation effects hold across a much broader set of writing systems. Second, by comparing multiple linguistic units, we provide a more precise empirical validation of ZLA, identifying letters as the optimal level of analysis for abbreviation effects in most writing systems. Together, these findings advance both the theoretical and empirical study of linguistic complexity, reinforcing the idea that both spoken and written communication are shaped by pressures to minimize encoding and decoding effort.

Dataset

To conduct a large-scale, cross-linguistic analysis of the relationship between information content and visual complexity, we selected a diverse set of languages from the Leipzig Corpora Collection (Goldhahn et al., 2012). This corpus offers text samples from a wide range of languages, though the available dataset sizes vary significantly across languages. To ensure both broad linguistic diversity and sufficient data for reliable statistical analysis, we applied two selection criteria:

1. Maximizing linguistic coverage—We prioritized the inclusion of as many languages as possible, covering diverse language families, scripts, and typological features.

Table 1: Languages Included in the Study: abbrevated name (Lan), fullname, native speaker population (Popul, millions)

Lan	Fullname	Popul	Lan	Fullname	Popul
afr	Afrikaans	7.2	mal	Malayalam	34
als	Albanian	7.5	mar	Marathi	83
amh	Amharic	32	mhr	Mari	0.4
ara	Arabic	310	min	Minangkabau	5.5
arg	Aragonese	0.01	mkd	Macedonian	1.3
arz	Egyptian Arabic	65	mlg	Malagasy	18
asm	Assamese	15	mlt	Maltese	0.5
ast	Asturian	0.1	mon	Mongolian	5.2
aze	Azerbaijani	23	mri-nz	Māori	0.05
azj-az	North Azerbaijani	23	msa	Malay	60
bak	Bashkir	1.2	mwl	Mirandese	0.01
ban-id	Balinese	3.3	mzn	Mazanderani	2.3
bar	Bavarian	14	nan	Min Nan	50
bel	Belarusian	6.5	nds	Low German	4.8
ben	Bengali	230	nep	Nepali	16
bos	Bosnian	2.5	new	Newar	1.3
bpy	Bishnupriya Manipuri	0.5	nld	Dutch	25
bre	Breton	0.2	nno	Norwegian Nynorsk	0.6
bul	Bulgarian	7	nob	Norwegian Bokmål	4.6
cat	Catalan	4.1	nor	Norwegian	5.3
ceb	Cebuano	20	oci	Occitan	0.6
ces	Czech	10.7	ori	Odia	38
che	Chechen	1.4	pan	Punjabi	125
chv	Chuvash	1.1	pap	Papiamento	0.3
cym	Welsh	0.5	pes	Western Persian	55
dan	Danish	6	plt	Plateau Malagasy	6
deu	German	76	pms	Piedmontese	1.6
diq	Zazaki	0.3	pnb	Western Punjabi	93
div	Dhivehi	0.3	pol	Polish	40
ekk	Standard Estonian	0.9	por	Portuguese	232
ell	Greek	13.5	pus	Pashto	40
eml	Emilian	1	roh	Romansh	0.04
eng	Simple English	370	ron	Romanian	24
epo	Esperanto	0	rus	Russian	154
est	Estonian	0.9	sah	Sakha (Yakut)	0.5
eus	Basque	0.75	san	Sanskrit	0.02
fao	Faroese	0.07	scn	Sicilian	4
fas	Persian	77	sco	Scots	1.5
fin	Finnish	5.4	sin	Sinhala	17
fra	French	80	slk	Slovak	5.2
fry	Frisian	0.5	slv	Slovenian	2.3
gle	Irish	0.17	sna-zw	Shona	7.5
glg	Galician	2.4	snd	Sindhi	32
gom	Goan Konkani	2.5	som	Somali	16
gsw	Swiss German	4.9	spa	Spanish	485
guj	Gujarati	56	sqi	Albanian	7.5
hat	Haitian Creole	10	srp	Serbian	8.2
hau	Hausa	50	sun	Sundanese	42
hbs	Serbo-Croatian	19	swa	Swahili	18
heb	Hebrew	5	swe	Swedish	10.5
hin	Hindi	345	swh	Swahili (variant)	18
hrv	Croatian	5.6	szl	Silesian	0.5
hsb	Upper Sorbian	0.02	tam	Tamil	75
hun	Hungarian	13	tat	Tatar	5.2
hye	Armenian	6.7	tel	Telugu	83
ido	Ido	0	tgk	Tajik	4.5
ina	Interlingua	0	tgl	Tagalog	28
ind	Indonesian	43	tuk	Turkmen	7.5
isl	Icelandic	0.32	tur	Turkish	88
ita	Italian	64	uig	Uyghur	10
jav	Javanese	82	ukr	Ukrainian	27
kal-gl	Greenlandic	0.056	urd	Urdu	69
kan	Kannada	44	uzb	Uzbek	32
kat	Georgian	3.7	uzn-uz	Northern Uzbek	30
kaz	Kazakh	13	vec	Venetian	3.9
kin	Kinyarwanda	12	vie	Vietnamese	86
kir	Kyrgyz	4.3	vls	West Flemish	1.1
kor	Korean	80	vol	Volapük	0
kur	Kurdish	16	war	Waray	3.5
lat	Latin	0	wln	Walloon	0.6
lav	Latvian	1.3	xho-za	Xhosa	8.2
lim	Limburgish	1.3	xmf	Mingrelian	0.5
lit	Lithuanian	2.8	yid	Yiddish	0.5
lmo	Lombard	3.5	zhs	Chinese	918
ltz	Luxembourgish	0.4	zsm	Standard Malay	18
lug	Luganda	8	zul-za	Zulu	12
lus-in	Mizo	0.83			
lvs	Latgalian	0.15		Total	5634

2. Ensuring a minimum corpus size—We required each language to have at least 30,000 sentences, as smaller

Figure 1: Dual-Metric Analysis of Linguistic Efficiency: (a) Perimetric and (b) Skeleton Length Complexity Correlations with Information Content Across 117 writing systems. Each subplot represents one writing system (labeled above), showing 5 linguistic units (left-to-right: letter, 2-gram, 3-gram, 4-gram, word) with three correlation coefficients (Pearson, Spearman, Kendall). Star(★) marks the strongest correlation per writing system.

corpora may introduce sampling biases or fail to provide robust estimates of letter- and word-level information measures.

Applying these criteria, we selected 155 languages, spanning multiple writing systems, including alphabetic, abugida, syllabic, and logographic scripts. These languages collectively represent a total native speaker population of approximately 5.63 billion, making this study one of the most comprehensive cross-linguistic analyses of writing system efficiency to date. A full list of the languages included, along with their geographic distribution and estimated native speaker populations, is provided in Table 1.

Method

Linguistic Units and Processing

We analyzed three levels of linguistic units across 155 writing systems: individual letters (1-grams), letter n-grams (2-4 grams), and complete words. For multi-character units, we focused on non-ligature scripts (Latin/Greek/Cyrillic) where characters maintain shape integrity. Cursive scripts (Arabic/Devanagari) were excluded from n-gram/word analyses due to positional shape variations.

Complexity Measures

We implemented two complementary complexity metrics:

1. **Perimetric Complexity (PC):**

For individual characters:

$$C = \frac{P^2}{4\pi A} \quad (1)$$

where P is contour perimeter and A is inked area, computed from standardized 100×100px character renderings.

For n-grams/words:

$$C = \frac{(\sum P_i)^2}{4\pi(\sum A_i)} \quad (2)$$

summing component perimeters and areas.

2. Skeleton Length Complexity (SLC) :

For individual characters:

$$SLC = \frac{L}{D} \quad (3)$$

where L is skeleton length (Zhang-Suen algorithm, 1984) and D is minimum enclosing circle diameter.

For n-grams/words:

$$SLC_{multi} = \sum \frac{L_i}{D_i} \quad (4)$$

summing normalized skeleton lengths.

Information Content Measure

For all linguistic units(letter, letter n-gram, word):

$$I = -\log_2 P \quad (5)$$

with probabilities P estimated from corpus frequencies.

Statistical Analysis

We employed three correlation measures:
1. Pearson's r (linear relationship)
2. Spearman's ρ (monotonic relationship)
3. Kendall's τ (ordinal association)

Analyses included:
- Hierarchical comparisons across units (letters, n-grams, words)
- Cross-metric validation (PC vs SLC)
- Script-type comparisons (alphabetic vs syllabic)

This multi-method approach enabled comprehensive evaluation of Zipf's Law while controlling for metric-specific biases and script variations.

Result

Comparing Letters, Letter N-Grams, and Words in Information-Complexity Correlation

Our comprehensive analysis writing systems reveals consistent patterns in how visual complexity optimizes with information content across different linguistic units. Using both perimetric complexity (Figure 1a) and skeleton length complexity (Figure 1b), we demonstrate that:

1. Letters (1-Gram) show strongest optimization: Both complexity measures (perimetric/skeleton) show highest correlations (★ in most of writing systems), confirming letters as primary optimization targets.
2. N-Gram decay effect: Correlations weaken from 2-Gram to 4-Gram, reflecting orthographic constraints.
3. Words occupy intermediate position: Word-level correlations typically surpass 3/4-Grams but remain below letters.
4. Metric-specific effects: (1) Perimetric complexity better captures optimization in alphabetic systems; (2)Skeleton length proves more sensitive to connected scripts.

These results extend Koshevoy et al. (2023) by: (1) Validating ZLA across two complexity dimensions; (2) Showing systematic decay with unit size; (3)Demonstrating stronger communication efficiency optimization at the letter level compared to words; (3) Revealing script-dependent optimization strategies.

The findings demonstrate writing systems prioritize letter-level efficiency, with larger units constrained by combinatorial and functional demands.

Figure 2: Letter-Level Linear Relationship Between Information Content and Perimetric Complexity

Evidence for the Stronger Optimization Pressure on Letters than Words

Our analysis of the top 20 most frequent letters and words across 155 languages reveals that letters appear ~1,000 × more frequently than words (mean ratio = 1,151.5, median = 1,228.4). This extreme frequency disparity (range: 661.2 –

1,380.2) explains the stronger optimization pressure onletters, as their pervasive use as writing system building blocks demands greater efficiency in visual processing. While words show weaker optimization due to morphological and orthographic constraints, their complexity remains influenced by constituent letter efficiencies.

Investigating the Letter-Level I-Complexity Relationship via Linear Regression

To further explore the linear relationship between information content and perimetric complexity at the letter level, we employed linear regression analysis, a classic method for quantifying systematic dependencies between variables. While the previous section demonstrated that letters generally exhibit the strongest correlation between I and C compared to larger linguistic units, the analysis was conducted on a subset of languages. Here, we extend our investigation to a broader dataset of 155 languages, including additional scripts that were omitted previously due to letter shape transformations when forming larger units (e.g., cursive and abugida scripts). By analyzing a more comprehensive dataset, we aim to reinforce the cross-linguistic generalizability of Zipf's Law of Abbreviation (ZLA) in writing systems.

Each subplot in Figure 2 represents a different language, with information content (I) on the x-axis and perimetric complexity (C) on the y-axis. Scatter points correspond to individual letters, and the fitted regression line is color-coded: red for significant positive correlations, green for significant negative correlations, and black for non-significant results. Pearson's correlation coefficient (ρ) and the coefficient of determination (r^2) are reported for each language, quantifying both correlation strength and explanatory power. Following are key findings:

1. **Most languages exhibit a significant positive correlation between I and C**: The fitted regression lines are predominantly red, with ρ values typically between 0.2 and 0.4, suggesting that letters with higher information content tend to have greater visual complexity.This pattern is particularly robust in Latin-based scripts (e.g., French (fra), Spanish (spa), English (eng)), reinforcing the notion that letter efficiency optimization is a widespread phenomenon, likely driven by communicative economy principles such as the Principle of Least Effort.
2. **Variation across languages-Strong, weak, and negative correlations**: Some languages exhibit particularly strong positive correlations, such as Xhosa (xho, ρ =0.61, r^2 = 0.38) and Zulu (zul, ρ =0.57, r^2=0.32), suggesting that information-driven complexity optimization is especially pronounced in these scripts. Chinese (zhs, ρ =0.14, r^2 =0.02) and Japanese (jpn, ρ =0.14, r^2 =0.02) show weaker correlations, likely due to their logographic writing systems, where complexity is influenced more by radical structure and morpho-phonemic constraints than by pure information-theoretic principles. A few languages, such as Kannada (kan, ρ = -0.18), display negative correlations, implying that higher-information letters in these scripts tend to be visually simpler. This might be linked to distinctiveness pressures, cognitive recognition efficiency, or unique script-specific constraints.
3. **Low r^2 values indicate additional influencing factors**: While a positive correlation is observed in most languages, r^2 values are generally below 0.15, suggesting that information content alone does not fully determine visual complexity. Factors such as letter distinctiveness, ease of writing, historical evolution, and cognitive processing demands likely contribute to shaping the observed complexity distributions. These findings indicate that ZLA operates within writing systems but is modulated by language-specific constraints, reinforcing the idea that communicative efficiency is not the sole determinant of script evolution.

By expanding the dataset to 155 languages, this analysis provides a more comprehensive empirical foundation for understanding the role of I-Complexity in writing systems. Our results confirm that letters are the primary unit where efficiency-driven optimization is observed, but they also reveal cross-linguistic variations that merit further investigation. Future work should examine how orthographic conventions, perceptual discriminability, and evolutionary pressures shape letter complexity across diverse writing systems.

Discussion

Our study provides strong empirical evidence that Zipf's Law of Abbreviation (ZLA) extends to the visual complexity of writing systems, particularly at the letter level. By analyzing letters, letter n-grams, and words across over 150 languages, we confirm that letters exhibit the strongest correlation between information content and perimetric complexity. This result aligns with prior research (Koshevoy et al., 2023) but significantly broadens its empirical scope, reinforcing the universality of abbreviation effects in writing.

Why Letters Exhibit the Strongest Optimization?

A key finding is that letters, rather than words or letter n-grams, exhibit the strongest abbreviation effects, likely due to several factors:

1. **Higher Frequency of Letters**–As shown in Table 2, letters appear hundreds to thousands of times more often than words, leading to stronger optimization pressure and clearer abbreviation effects.
2. **Letter Optimization Influences Words**-Words also show a moderate correlation between information content and complexity, likely due to their composition from already optimized letters. However, additional

linguistic constraints (e.g., morphology, orthographic conventions) weaken this effect.

3. **Weaker Effects in Larger Units**-The decline in correlation strength for n-grams suggests that as linguistic units grow larger, structural constraints increasingly outweigh efficiency pressures, making letters the primary site of abbreviation-driven optimization.

Cross-Linguistic Variability and Structural Constraints

While letter-level optimization is robust across many languages, our linear regression analyses (Figure 2) reveal significant variation:

1. Stronger Correlations in Alphabetic Scripts： Alphabetic languages (e.g., English, French, Spanish) show strong positive correlations, suggesting that letter-based systems prioritize efficiency in visual processing.
2. Weaker Correlations in Logographic and Syllabic Scripts： Logographic systems (e.g., Chinese) exhibit weaker correlations, as character complexity is driven by radical structures and morpho-semantic constraints rather than frequency alone.
3. Negative Correlations in Some Scripts： A few languages (e.g., Kannada) show negative correlations, possibly due to distinctiveness pressures, where high-information symbols must remain visually simple for recognition.

Anomalous Cases in Letter-Level Complexity Optimization

Our analysis reveals notable exceptions to the expected positive information-complexity correlation. Cebuano (ceb), Marathi (mar), Tamil (tam), and Volapük (vol) show near-zero correlations ($\rho \approx 0$), suggesting decoupled optimization. For Tamil, this may reflect historical preservation of ancient Dravidian forms, while Volapük's artificial origins prioritize systematic design over frequency-based evolution.

More strikingly, Bishnupriya (bpy) and Kannada (kan) exhibit negative correlations ($\rho < 0$), with higher-information letters being simpler. In Kannada (an abugida), frequent characters may simplify to maintain distinctiveness, while Bishnupriya's conjunct rules could favor simple base characters for combination. These cases demonstrate how script-specific requirements can override or reverse typical efficiency pressures.

Refining Zipf's Law in Writing Systems

Our findings refine the theoretical understanding of ZLA by emphasizing that effort minimization in writing is primarily visual, not just length-based. Unlike spoken language, where effort is articulatory, written systems minimize effort through visual simplicity. Furthermore, while abbreviation effects are observed in larger linguistic units (words, n-grams), they operate most efficiently at the letter level, reinforcing the idea that writing systems optimize their most fundamental units first.

Limitations

While our study provides broad empirical support for Zipf's Law of Abbreviation across 155 writing systems, several limitations warrant discussion. First, our dataset, though large, exhibits uneven script diversity, with disproportionate representation of Latin and Arabic-based systems. This clustering introduces potential phylogenetic dependencies, as shared script ancestry may correlate with shared linguistic features (e.g., phoneme-grapheme mappings). Future work could address this by incorporating phylogenetic regression or script-family random effects to disentangle evolutionary pressures from areal influences.

Second, our focus on perimetric and skeleton-length complexity, while theoretically grounded, does not capture all dimensions of visual processing. As noted by the reviewer, algorithmic complexity—which quantifies perceptual discriminability—might better reflect modern reading-optimized typography. Similarly, contextual predictability (cf. Piantadosi et al., 2011) could refine information-content estimates by accounting for letter redundancy in sequences.

These limitations, however, highlight rich opportunities for extension: (1) controlled script sampling to balance genetic and areal diversity, (2) multimodal complexity metrics (e.g., algorithmic, topological), and (3) diachronic analyses of script evolution. Such advances would further illuminate the interplay between efficiency, perception, and historical contingency in shaping writing systems.

Implications and Future Directions

Our findings highlight that writing systems evolve under efficiency pressures, balancing readability and production ease. Incorporating visual complexity into computational models of script evolution could enhance our understanding of how writing optimizes for cognitive processing.

Future research should explore alternative complexity metrics (e.g., stroke count, fractal dimension), investigate historical trends in script evolution, and conduct psycholinguistic studies (e.g., eye-tracking, reaction time) to assess how visual complexity affects reading fluency and cognitive load.

Conclusion

Our study confirms that Zipf's Law of Abbreviation applies most strongly at the letter level of writing systems, with letters showing the highest correlation between information content and visual complexity. This refines the theoretical understanding of effort minimization in writing, demonstrating that visual complexity plays a key role in script optimization. Future work should further examine how efficiency pressures interact with distinctiveness and cognitive constraints across diverse writing systems.

References

Abdelhadi, S., Ibrahim, R., & Eviatar, Z. (2011). Perceptual load in the reading of Arabic: Effects of orthographic visual complexity on detection. *Writing Systems Research,* 3(2), 117-127.

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

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). University of Tübingen.

Chan, A. H. S., & Lee, P. S. K. (2005). Effects of different task factors on speed and preferences in Chinese handwriting. *Ergonomics*, 48(1), 38-54.

Chow, S. M. K., Choy, S.-W., & Mui, S.-K. (2003). Assessing handwriting speed of children biliterate in English and Chinese. Perceptual and Motor Skills, 96(2), 685-694.

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

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 759–765.

Ibrahim, R., Eviatar, Z., & Aharon-Peretz, J. (2002). The characteristics of Arabic orthography slow its processing. Neuropsychology, 16(3), 322.

Kanwal, J., et al. (2017). Language-users choose short words in predictive contexts in an artificial language task. *CogSci 2017*.

Kanwal, J. K. (2018). Word length and the principle of least effort: Language as an evolving, efficient code for information transfer.

Koshevoy, A., Miton, H., & Morin, O. (2023). Zipf' s law of abbreviation holds for individual characters across a broad range of writing systems. Cognition, 238, 105527.

Kuo, L.-J., Li, Y., Sadoski, M., & Kim, T.-J. (2014). Acquisition of Chinese characters: The effects of character properties and individual differences among learners. *Contemporary Educational Psychology*, 39(4), 287-300.

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., Tily, H. J., & Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108, 3526-3529.

Tanida, Y. (2023).The Relationship Between Word Length and Average Information Content in Japanese. *Cognitive Science*, 47(6), e13302.

Zhang, Q., & Feng, C. (2017). The interaction between central and peripheral processing in Chinese handwritten production: Evidence from the effect of lexicality and radical complexity. Frontiers in Psychology, 8, 334.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

Wang, H., He, X., & Legge, G. E. (2014). Effect of pattern complexity on the visual span for Chinese and alphabet characters. Journal of Vision, 14(8), 6

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

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley.

Zhang, T. Y., & Suen, C. Y. (1984). A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 27(3), 236-239.

Zhu, Z., Yu, D., He, X., Wang, J., & Legge, G. E. (2019). Perceptual learning of visual span improves Chinese reading speed. Investigative Ophthalmology & Visual Science, 60(6), 2357-2368.