languages

MDPI

Article

Orthographic Visual Distinctiveness Shapes Written Lexicons: Cross-Linguistic Evidence from 131 Languages

Jiazheng Wang, Ruimin Lyu *, Hangyu Zhu and Zhenping Xie

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; 6233112042@stu.jiangnan.edu.cn (J.W.); 8202204005@jiangnan.edu.cn (H.Z.); xiezp@jiangnan.edu.cn (Z.X.)
* Correspondence: ruiminlyu@jiangnan.edu.cn

Abstract

Written language is a multimodal system that integrates visual, phonological, and semantic information. This study examines whether orthographic visual distinctiveness—the degree to which word forms differ visually—acts as a structural constraint across languages. Using standardized script renderings from 131 languages, we extracted visual features of words through a Vision Transformer (VIT) and compared visual distances between co-occurring word pairs from natural corpora and random word pairs from lexicons, controlling for word length and related factors. The results show that co-occurring words are visually more distinct than expected by chance, and this effect is consistent across diverse writing systems. These findings indicate that visual distinctiveness contributes independently to the organization of written language, reflecting an underlying pressure toward visual discriminability in lexical form. Beyond linguistic implications, the framework demonstrates how deep vision models can capture cognitively meaningful visual features of text, offering new perspectives for multimodal research on orthography, reading, and cross-lingual modeling.

Keywords: orthographic visual distinctiveness; writing systems; graphemic similarity; cross-linguistic variation; multimodal representation; orthographic distance; Vision Transformer (VIT)

Academic Editor: Anthony Pak-Hin Kong

Received: 21 October 2025
Revised: 6 December 2025
Accepted: 8 December 2025
Published: 11 December 2025

Citation: Wang, J., Lyu, R., Zhu, H., & Xie, Z. (2025). Orthographic Visual Distinctiveness Shapes Written Lexicons: Cross-Linguistic Evidence from 131 Languages. *Languages*, *10*(12), 301. https://doi.org/10.3390/languages10120301

1. Introduction

The structure of words in human language reflects a long history of adaptation to diverse functional and cognitive pressures. Far from being arbitrary, linguistic form evolves under constraints that promote communicative efficiency and processing ease. Prior research has demonstrated that phonological distinctiveness prevents confusion in speech perception (Wedel et al., 2018), communicative efficiency drives word-length optimization relative to predictability and frequency (Piantadosi et al., 2011), and learning and memory limitations shape lexical organization and transmission (Bybee, 2006; Kirby et al., 2015; Zuidema & De Boer, 2009). Together, these factors suggest that linguistic systems tend to balance multiple pressures that favor effective communication.

Yet, written language introduces an additional modality through which such pressures operate. Orthography is not merely a symbolic representation of sound but a multimodal system that integrates visual, phonological, and semantic cues. Reading recruits both sensory and linguistic processes: visual decoding activates phonological codes and semantic associations even during silent reading (C. A. Perfetti & Hart, 2008; Rayner, 1998). The perceptual appearance of text—its letterforms, contrast, and spacing—affects not only legibility but also reading fluency and word recognition (Beier & Larson, 2010; Tinker, 1963).

Despite this, the visual dimension of linguistic form has received relatively little systematic attention compared with the auditory domain.

A preliminary illustration of this phenomenon is presented in Figure 1, which compares representative examples of ancient and modern writing systems across several language families. The upper panels display early scripts such as Sumerian cuneiform, Egyptian hieroglyphs, and ancient Greek cursive, while the lower panels show modern counterparts including Devanagari, Arabic, and Latin. As writing systems evolved, the visual separation between words became increasingly pronounced: early scripts show continuous symbol sequences with limited spacing, whereas modern scripts exhibit clearer boundaries and stronger visual differentiation. This visible progression suggests that word-level visual distinctiveness may have gradually increased through cultural evolution, improving legibility and reducing perceptual confusability. Although the comparison is illustrative, it provides an intuitive motivation for examining whether such visual discriminability systematically shapes lexical form across languages.

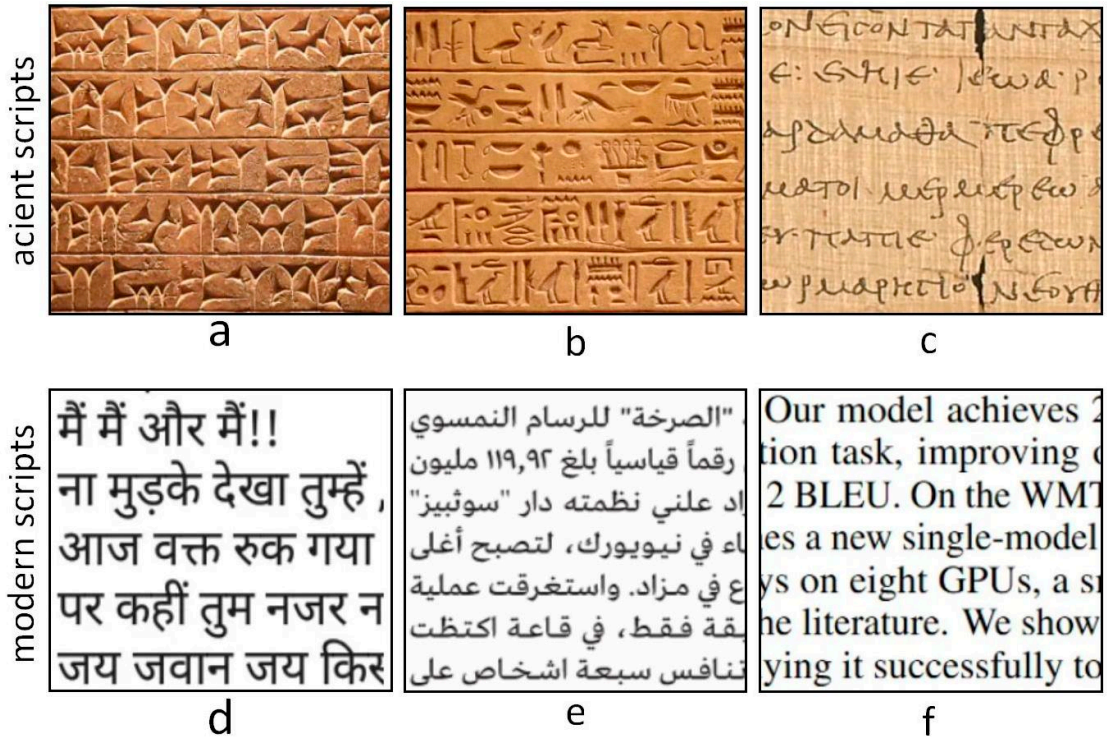

Figure 1. Ancient and modern scripts illustrating changes in inter-word visual distinctiveness. Panels (**a**–**c**) show ancient scripts—(**a**) Sumerian cuneiform, (**b**) Egyptian hieroglyphs, and (**c**) ancient Greek cursive; panels (**d**–**f**) present modern scripts—(**d**) Devanagari, (**e**) Arabic, and (**f**) Latin. A general trend toward greater visual separation and boundary clarity is observable from older to more recent writing systems.

Evidence from cognitive psychology and typography supports this interpretation. Experimental studies show that increased visual differentiation among graphemes enhances letter identification (Geyer, 1977), reduces error under noisy conditions (Bernard, 2016), and improves reading efficiency. Perceptually distinct graphemes lower confusability (Mueller & Weidemann, 2012), while even subtle typographic refinements—such as spacing adjustments or stroke width variation—can significantly affect lexical decision performance (Beier & Larson, 2010; Tinker, 1963). These findings underscore that orthographic visual distinctiveness (orthographic visual distinctiveness) is cognitively functional, yet prior

work has largely examined perception within fixed orthographies rather than its potential role as a shaping force in the evolution of written form.

Building on these insights, the present study explores whether orthographic visual distinctiveness systematically constrains lexical form across languages. A direct historical analysis would be ideal but remains infeasible due to limited diachronic corpora. As an alternative, we adopt a large-scale, cross-linguistic approach, leveraging data from 131 languages to test whether words that co-occur in natural corpora tend to be more visually distinct than expected by chance. This method assumes that, if written language has been shaped by long-term visual optimization, then word pairs that co-occur in natural language should display greater visual dissimilarity than randomly paired words drawn from the lexicon.

To operationalize visual structure, we employ Vision Transformers (VITs)—a class of deep neural networks that process images as sequences of visual patches that approximate human shape-based visual perception through self-attention mechanisms. Self-attention allows the model to globally weight and integrate information across all parts of an image, enabling sensitivity to overall shape and structural configuration rather than only local visual details. In contrast to traditional convolutional neural network (CNN)-based models, which primarily rely on local convolution filters and hierarchical receptive fields, VITs directly capture global visual relationships. VITs have demonstrated strong shape bias (Dehghani et al., 2023) and high cross-script generalizability, successfully modeling diverse character systems including Chinese (Dan et al., 2022), Persian numerals (Ardehkhani et al., 2024), Hangul (Shana et al., 2024), and Manchu (Zhou et al., 2024). Their ability to extract orthographic visual features enables a scalable and language-neutral comparison of visual distinctiveness across writing systems.

To isolate orthographic visual distinctiveness from simpler metrics, we include a control analysis based on word-length difference, which correlates with both visual complexity and lexical processing difficulty (Changizi et al., 2006; New et al., 2006; O'Regan et al., 1984). By comparing visual dissimilarity while accounting for word-length variation, we separate the contribution of structural contrast from mere size effects.

Our results show that, in most languages, co-occurring word pairs are visually more distinct than randomly paired words, and this pattern persists after controlling for length. These findings suggest that visual distinctiveness functions as an independent, modality-specific constraint in the organization of written lexicons. Beyond their linguistic implications, the results also demonstrate that computational vision models can reveal cognitively relevant regularities in text, bridging perceptual and linguistic structure. This perspective extends functional theories of language design by incorporating visual perceptual pressures into models of lexical and orthographic evolution.

2. Materials and Methods

The study aims to test whether visual dissimilarity between word forms is systematically greater in natural language use than would be expected by random lexical structure. Our analytical framework comprises four components:

1. Corpus selection;
2. Extraction of orthographic visual features using a Vision Transformer (VIT);
3. Sampling and comparison of word pairs;
4. Statistical evaluation of visual versus effects.

The overall workflow is summarized in Figure 2:

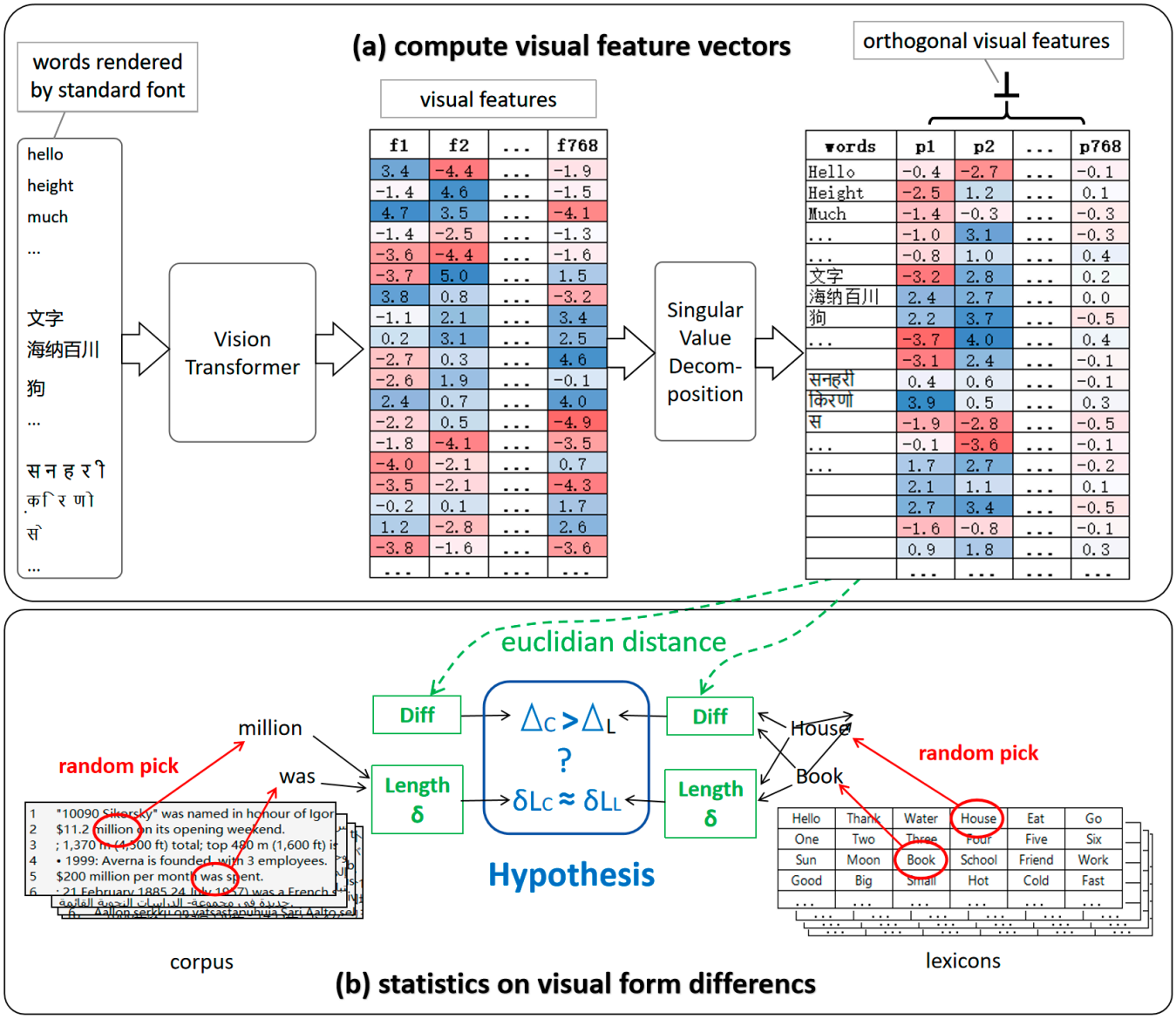

Figure 2. Overview of the analytical framework for measuring orthographic visual distinctiveness. (**a**) Word forms from 131 languages are rendered in standardized fonts and encoded into 768-dimensional visual features using a Vision Transformer (ViT). (**b**) The embeddings are orthogonalized using singular value decomposition (SVD), and Euclidean distances between word pairs are computed as measures of visual dissimilarity. Two conditions are compared: co-occurring word pairs from corpora and random word pairs from lexicons. Larger visual distance in corpus-based pairs reflects enhanced visual distinctiveness in natural usage. A synthetic cross-linguistic summary of effect sizes is provided in Figure 3.

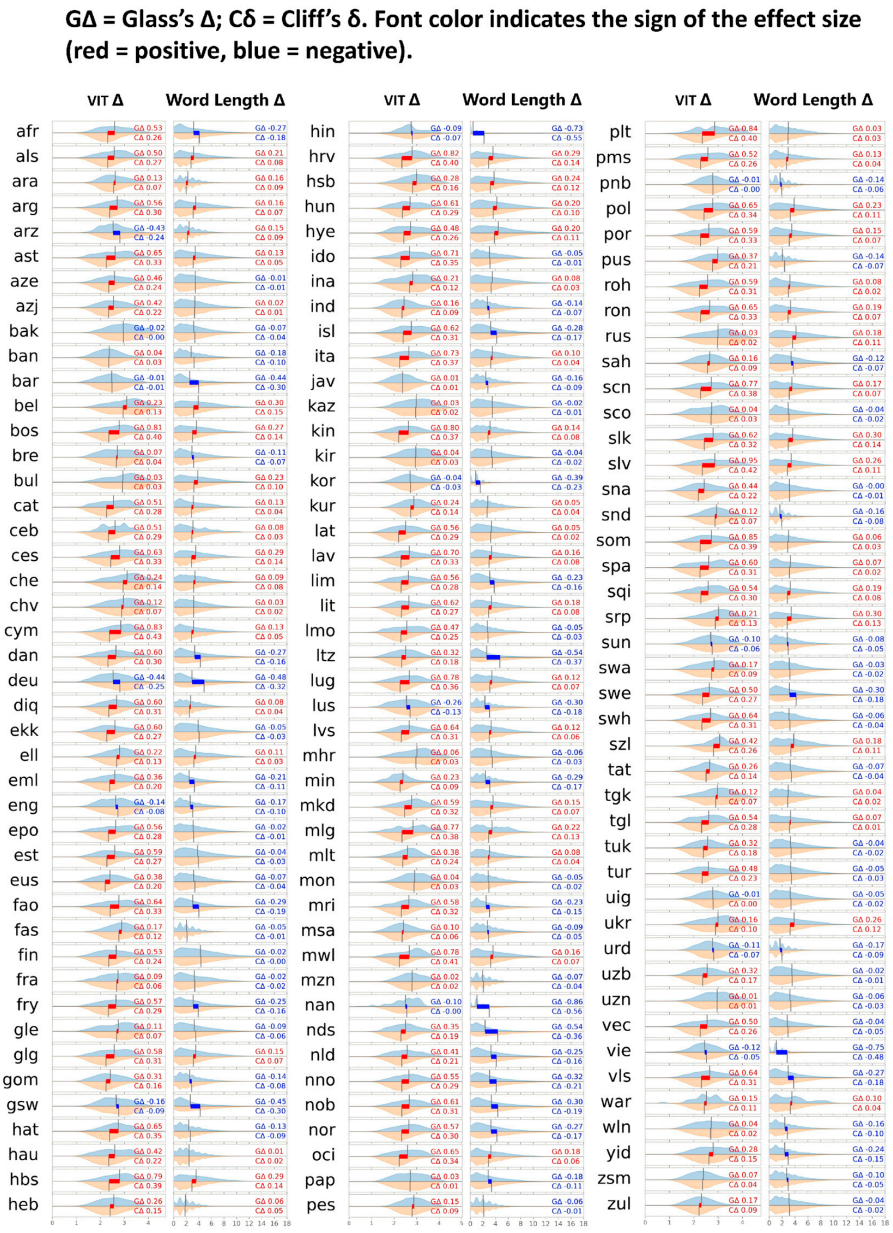

Figure 3. Cross-linguistic comparison of visual and structural differences between word pairs. Ridge plots show pairwise visual-form distances (**left**) and word-length differences (**right**) for 131 languages. Blue distributions correspond to corpus co-occurrences and orange to random lexicon pairs; gray lines mark means, with Glass's Δ and Cliff's δ indicating effect sizes. Most languages exhibit substantially higher visual distances in corpus pairs than in random pairs, while length differences remain weak or inconsistent. All three-letter language abbreviations and their corresponding full language names are listed in Appendix A (Table A1).

2.1. Corpus Selection

We selected a monolingual Wikipedia corpus from 131 languages in the Leipzig corpus collection (Hewavitharana & Vogel, 2013). The Wikipedia data source is particularly suitable for our task due to its cross-linguistic coverage, corpus scale, and consistent written-style documentation. As a predominantly formal written corpus, Wikipedia establishes effective control over linguistic variations across different language datasets. To ensure comparability and model stability, we excluded Chinese and Japanese because they lack explicit word boundary markers and rely on language-specific segmentation schemes, which introduce substantial methodological ambiguity for defining word-level units in

a cross-linguistic and computationally uniform manner (Sproat et al., 1996; Xue, 2003). Additionally, languages with very small corpora (<2 MB) or significant script-mixing noise were excluded. Each remaining corpus was subsampled to 30,000 sentences.

The dataset prepared for analysis has an average total word count of 478,046 and a vocabulary size (i.e., the number of unique word types) of 59,798 for each language. The smallest corpus contains 274,687 words with a vocabulary size of 7994, reflecting cross-linguistic differences in average sentence length and lexical diversity, rather than differences in the number of sentences, which was strictly fixed at 30,000 for all languages.

2.2. Visual Feature Extraction

Words were rendered in a standardized font (black-on-white background) using the most common fonts for each language from Google Fonts, with Arial as the default for smaller languages. Images were resized to a consistent 800 × 224 pixels—a size chosen based on pre-experiments that ensured over 99.99% of words were properly displayed. This size was selected to accommodate common tokenization issues, where unusually long tokens may result in display anomalies. The word images were then input to a pretrained Vision Transformer (VIT-B/16) model (Dosovitskiy, 2020), which has been shown to outperform CNN-based models in cross-script symbol recognition tasks (Ardehkhani et al., 2024; Dan et al., 2022). VIT's attention-based architecture enables sensitivity to global shape and structural contrast (Dehghani et al., 2023), making it a plausible proxy for human visual perception of orthographic forms.

To reduce noise and align feature dimensions across languages, we applied singular value decomposition (SVD) to the 768-dimensional VIT embeddings (i.e., numerical vector representations that encode the visual form of each word image), retaining all 768 orthogonal components. The purpose of this procedure is to remove correlations between feature dimensions and to place all features in a common orthogonal space, thereby improving the stability and comparability of subsequent distance measurements across languages. Visual dissimilarity between word pairs can then be computed as Euclidean distance in this orthogonal feature space. The dataset consists of an average vocabulary size of 59,798 words, corresponding to 59,798 word images per language, with each image being represented by a 768-dimensional visual feature. I also designed two sampling methods for testing, which are explained in detail in Section 2.3.

2.3. Sampling and Comparison Design

We compared two sets of word pairs:

1. Co-occurring word pairs (5000 pairs for each language) defined as sentence-internal adjacent word pairs (i.e., bigrams), randomly sampled from all such bigrams in the running corpus of each language (corpus condition);
2. Word pairs (5000 pairs for each language) randomly and uniformly drawn from the language-specific lexicon, without considering word frequency (lexicon condition), as an unstructured baseline.

For each pair, we computed both the visual distance (Δ) and the absolute difference in word length (δ). Crucially, we hypothesized that Δ is systematically larger in corpus pairs ($\Delta\_C > \Delta\_L$), while δ should remain similar ($\delta\_C \approx \delta\_L$), as an emergent outcome rather than a control variable.

2.4. Statistical Evaluation

To quantify effect sizes, we employed Glass's Δ (Glass, 1976) and Cliff's δ (Cliff, 1993), both robust to unequal sample sizes and non-normal distributions. These measures allowed

us to assess whether visual differences observed in corpus-based pairs are statistically and practically greater than expected from random lexical structure alone.

3. Results

To evaluate whether orthographic visual distinctiveness is systematically enhanced in natural language use, we analyzed pairwise visual distances between word forms across 131 languages.

The comparison between co-occurring words in natural corpora and randomly paired words from lexicons revealed a clear and consistent trend: in nearly all languages, naturally co-occurring words are visually more distinct than expected by chance. This effect persists after controlling for word length and is robust across writing systems, indicating that visual differentiation functions as an independent structural property of written language.

The same comparison was performed on word-length differences, which served as a structural baseline to ensure that visual effects were not merely reflections of size variation. Results from all 131 languages are summarized in Figure 3 and Table 1.

Table 1. Aggregate results across 131 languages.

Measure	Glass's Δ (Visual)	Cliff's δ (Visual)	Glass's Δ (Length)	Cliff's δ (Length)
Mean	0.361	0.186	−0.031	−0.029
Median	0.387	0.212	−0.015	−0.008
Positive num	119	119	59	60
Total num	131	131	131	131
% Positive	91%	91%	45%	46%

As illustrated in the ridge plots, for nearly every language the blue distributions (corpus pairs) show higher mean visual distances than the orange distributions (lexicon pairs). Effect sizes are consistently positive. More specific statistics can be observed in Table 1 below.

Combining the charts, we can observe mean and median effect sizes for visual and length-based measures (Glass's Δ and Cliff's δ). Visual effects show a strong positive bias across languages, whereas length effects remain small and variable. Glass's Δ exceeds 0.2 in 85 languages and is positive in 118, while Cliff's δ exceeds 0.1 in 88 and is positive in 118. The unimodal and symmetric shape of most distributions suggests that the enhancement reflects a genuine central tendency rather than a few outliers. These findings indicate that words that co-occur in natural usage are visually more distinct than would be expected by chance, supporting the hypothesis of enhanced orthographic contrast in real language data.

In contrast, word-length differences display weaker and more variable effects. The corresponding distributions are often skewed or multimodal, and effect sizes remain small: Glass's Δ exceeds 0.2 in only 16 languages (positive in 59), and Cliff's δ exceeds 0.1 in 15 (positive in 60). There is no systematic pattern showing that corpus pairs differ consistently in length from random pairs. This confirms that the observed visual distinctiveness cannot be explained by trivial structural variation.

Overall, across 131 languages, visual-form distances derived from VIT embeddings reveal a robust and widespread enhancement of orthographic visual distinctiveness in natural corpora. This large-scale pattern suggests that the visual form of written words is subject to independent structural pressures that promote visual discriminability—a theme further explored in the Discussion.

4. Discussion

This section discusses the broader implications of our findings on orthographic visual distinctiveness and its role in shaping written language.

We begin by summarizing the main empirical results and their theoretical significance and then relate them to prior research in orthography, perception, and multimodal language processing.

Subsequent sections address cross-linguistic and typological perspectives, methodological considerations, and potential limitations.

Finally, we discuss how these findings inform our understanding of human reading versus model-based "reading" and outline directions for future research and applications in linguistics and multimodal NLP.

4.1. Summary of Findings and Theoretical Implications

This study provides cross-linguistic evidence that orthographic visual distinctiveness in written word forms is systematically enhanced in natural language use. By comparing word pairs sampled from corpora and lexicons across 131 languages, we found that naturally co-occurring words exhibit significantly greater visual distinctiveness, even when controlling for word length. These results demonstrate that visual discriminability is a stable and pervasive property of written language, rather than a by-product of simple structural variation.

Across languages, the visual distance between corpus-based word pairs is consistently larger than that of random lexical pairings. The effect is robust: Glass's Δ exceeds 0.2 in 85 languages and is positive in 118, while Cliff's δ exceeds 0.1 in 88 and is positive in 119.

Moreover, both corpus and lexicon distributions are largely unimodal and symmetric, suggesting that the enhancement of visual distinctiveness represents a central tendency across languages rather than the influence of a few outliers.

In contrast, word-length differences show greater variability and lack systematic directionality, confirming that the observed visual effect is not reducible to size or orthographic density.

These findings extend the concept of distinctiveness-based selection, previously established in phonology (Wedel et al., 2018) and lexical semantics (Piantadosi et al., 2011), into the visual domain of written language. Just as spoken languages evolve toward greater phonological contrast to minimize auditory ambiguity, writing systems appear to evolve toward greater visual distinctiveness, reducing perceptual overlap and improving readability. This convergence implies that orthographic visual distinctiveness, like phonological contrast, serves as a functional mechanism optimizing language for communication.

In addition, the results support the integration of visual features into multimodal models of language representation (Kirby et al., 2015), suggesting that perceptual constraints form an essential part of the communicative design of linguistic systems.

4.2. Relation to Prior Work in Orthographic and Perceptual Research

This study builds upon a long tradition of research in perceptual psychology and orthographic design, which has demonstrated that the visual structure of writing systems directly affects reading performance and lexical processing. At the level of individual graphemes, letter distinctiveness has been shown to improve recognition accuracy and reading speed (Geyer, 1977), particularly under degraded or low-contrast conditions (Bernard, 2016). Similarly, graphemic similarity influences lexical access and confusability during word recognition (Mueller & Weidemann, 2012), while typographic variation—including stroke width, spacing, and font style—can modulate both readability and cognitive load even in familiar scripts (Beier & Larson, 2010; Tinker, 1963). Together, these findings underscore

that visual form is not a neutral carrier of linguistic information but a functional component of the reading process.

However, most previous studies have focused on within-script perception, employing controlled experiments or artificial stimuli to assess visual confusability and recognition speed. In contrast, our approach extends this inquiry to the language-system level, asking whether natural lexicons themselves show a cumulative bias toward visual differentiation. By analyzing thousands of real word pairs across 131 languages, we reveal a large-scale statistical tendency that complements experimental results: visual discriminability functions as an organizing pressure in written lexical form. This perspective introduces a bridge between micro-level perceptual effects and macro-level linguistic structure, suggesting that perceptual constraints observed in cognitive experiments may also shape the distributional design of orthographic systems over time.

Consequently, these findings provide a theoretical and empirical foundation for integrating perceptual principles into models of multimodal language representation and processing.

4.3. Cross-Linguistic and Typological Considerations

Our cross-language analysis offers insights into how orthographic visual distinctiveness interacts with script typology. While writing systems differ in mapping style (alphabetic vs. logographic), inventory size, directionality and segmentation conventions (Sampson, 2016), our findings suggest a common tendency across typologies: word forms that co-occur tend to be visually more distinct than random pairings. This universality indicates that perceptual separation may operate as a modality-independent constraint on written lexicons.

At the same time, typological factors may modulate the effect. For example, scripts with large grapheme inventories (e.g., Korean characters) or dense visual packing may rely more on shape differentiation (Miton & Morin, 2021), whereas alphabetic systems might exploit spacing, letter-contrast or boundary cues. Features such as cursive connectivity, diacritics, or script-mixing may impose additional visual constraints on word-form design (Meletis & Dürscheid, 2022).

From a typological perspective, our results invite a richer orientation: visual discriminability should be considered as an additional dimension of writing-system variation alongside traditional axes such as phonographic depth, inventory size and transparency. In other words, script design may reflect not only linguistic mapping decisions but also perceptual design pressures. Future research might examine how historical reforms, orthographic simplification, or script transitions correspond to changes in visual separability over time.

In sum, the cross-linguistic evidence supports a broad but nuanced picture: while script conventions vary widely, the drive toward enhanced orthographic visual distinctiveness between word forms appears to be a common structural tendency, thus contributing to our understanding of writing systems as perceptually grounded cultural artifacts.

4.4. Methodological Reflections: Multimodal Analysis

Our findings underline the need to regard written language not simply as a symbolic code but as a multimodal communicative system, where visual, phonological, and semantic information jointly shape linguistic form. Traditional corpus linguistics has focused primarily on lexical or syntactic structure, yet text is inherently visual—its layout, spacing, and letterforms affect readability and cognitive processing (C. A. Perfetti & Hart, 2008).

By representing words as visual objects, this study introduces a scalable way to quantify orthographic form. Using deep vision models such as the Vision Transformer, we can approximate human shape-based perception and thus measure visual distinctiveness

across scripts in a language-neutral manner. The key contribution is not technical but conceptual: it demonstrates how computational vision can serve as a methodological bridge between corpus linguistics and the study of orthography, enabling quantitative comparisons of visual structure that were previously unattainable.

In this sense, the multimodal perspective adopted here expands the analytical scope of linguistic research, offering new tools to investigate how perceptual constraints interact with lexical organization and script design.

4.5. Human Reading vs. Language Model "Reading": A Multimodal Gap

Human reading is a deeply multimodal cognitive process, integrating orthographic, phonological, and semantic information with visual properties such as font, spacing, and layout, and dynamically modulated by context, attention, and familiarity (C. Perfetti, 2007; Rayner, 1998; Scaltritti et al., 2019; Slattery & Rayner, 2013). Reading therefore engages both perceptual and linguistic systems: visual form serves not merely as a carrier of symbolic information but as a perceptual interface that shapes recognition and comprehension.

In contrast, current language models (LLMs) treat text as purely symbolic token sequences without perceptual grounding. They process strings of characters through statistical and attentional mechanisms yet remain blind to modality-specific cues such as orthographic visual distinctiveness, spacing, or graphemic structure. This difference reveals a fundamental multimodal gap between human reading and machine processing. While humans rely on the integration of perceptual and linguistic information, LLMs operate on abstracted symbols detached from sensory context.

Our findings highlight this gap: the demonstrated role of orthographic visual distinctiveness underscores how perception constrains linguistic structure in ways current models fail to capture. Incorporating visually informed representations—for instance, features derived from Vision Transformer encodings—may therefore help develop cognitively grounded multimodal language models that better approximate human reading (see Figure 4).

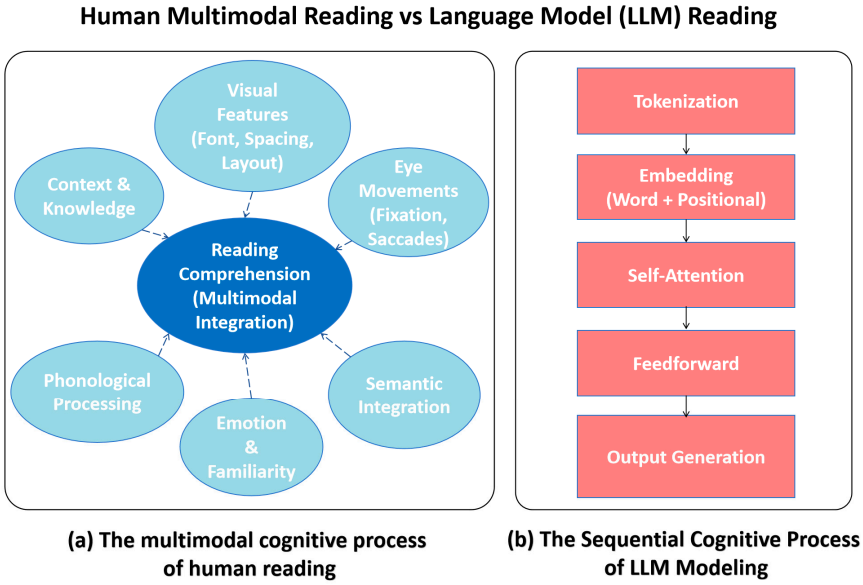

Figure 4. Human multimodal reading vs. symbolic processing in large language models. (**a**) Human reading integrates multiple perceptual and cognitive components—visual features, eye movements, phonological decoding, semantic integration, and contextual knowledge—into a unified comprehension process. (**b**) In contrast, large language models process tokenized symbol sequences through embedding and attention mechanisms without perceptual grounding, lacking modality-specific constraints such as orthographic visual distinctiveness.

4.6. Limitations and Future Directions

Despite its large cross-linguistic scope, this study has several limitations.

First, the VIT-based visual features used here approximate human shape perception but do not capture the full complexity of reading—such as eye-movement dynamics, familiarity effects, or contextual facilitation.

Second, although word length was explicitly controlled, other psycholinguistically relevant variables—including lexical frequency, predictability, and semantic association—were not directly modeled. These factors may indirectly influence visual distinctiveness and warrant further investigation in extended statistical frameworks.

Third, script-level properties—such as cursiveness, diacritics, or connected writing (e.g., in Arabic or Devanagari)—were not explicitly analysed and could moderate the observed effects.

Accordingly, the present study focuses on alphabetic and non-cursive writing systems in which word-level units are explicitly defined and visually separable. This design choice necessarily limits the typological coverage of the current analysis. In particular, the exclusion of logographic systems such as Chinese and mixed morpho-syllabic systems such as Japanese, as well as cursive writing systems with continuous character connections, means that the present findings cannot yet be generalized to all script types. Extending the present framework to segmentation-dependent scripts and to highly connected writing systems remains an important direction for future research.

Future work should therefore examine these factors within and across script families, integrating typological stratification to better isolate visual and structural influences. In addition, extending the analysis to historical scripts and diachronic corpora would provide direct evidence of how visual optimization unfolds over time.

The findings also point to several directions for future research and application. Theoretically, models of language evolution and orthographic change could incorporate visual discriminability as a central constraint, simulating how phonological, semantic, and visual pressures jointly shape lexical structure. Practically, integrating visual constraints into text design and AI-based generation could enhance readability and accessibility, particularly for a broad range of readers.

More broadly, the study underscores the importance of combining visual, phonological, and semantic modeling in multimodal frameworks, paving the way for cognitively grounded approaches to written-language processing and cross-cultural communication.

5. Conclusions

This study provides large-scale cross-linguistic evidence that orthographic visual distinctiveness in written word forms is not arbitrary but reflects systematic pressures toward perceptual distinctiveness. By combining deep visual representations with multilingual corpora, we show that co-occurring words are visually more distinct than random lexical combinations, a pattern that remains robust across 131 languages and independent of word length. These findings suggest that written languages, like spoken ones, evolve under functional pressures that promote discriminability and communicative clarity.

The results extend theories of distinctiveness-based selection to the visual domain, positioning orthographic design as part of a broader adaptive system linking perception, cognition, and communication. They also underscore a key difference between human multimodal reading—which integrates orthographic, phonological, semantic, and contextual information—and the symbolic text processing of current language models. Recognizing this gap invites new perspectives on how perceptual factors shape linguistic form and how future models might incorporate visually grounded representations to better reflect human language processing.

Overall, modeling language as a multimodal system highlights that perception and communication are inseparable dimensions of linguistic structure. Understanding written language through this integrated lens enriches both theoretical linguistics and applied language technologies, bridging cognitive, typological, and computational approaches to the study of human communication.

Author Contributions: Conceptualization, J.W.; Methodology, J.W.; Software, J.W.; Validation, J.W.; Formal analysis, J.W.; Investigation, J.W.; Resources, R.L. and Z.X.; Data curation, J.W.; Writing—review & editing, R.L. and H.Z.; Supervision, R.L., H.Z. and Z.X.; Project administration, J.W.; Funding acquisition, R.L. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are derived from publicly available sources in the Leipzig Corpus Collection (https://corpora.uni-leipzig.de). Processed data and analysis scripts are available from the corresponding author upon reasonable request.

Acknowledgments: The author thanks the Leipzig Corpus Collection team for providing publicly accessible multilingual data. The author also acknowledges the use of ChatGPT (GPT-5, OpenAI) for text refinement and confirms full responsibility for the final manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Language Abbreviations

This appendix lists all three-letter language abbreviations used in Figure 3 together with their corresponding full language names.

Table A1. Three-letter language abbreviations and corresponding full language names.

Code	Language_Name
afr	Afrikaans
sqi	Albanian
als	Alsatian
ara	Arabic
arg	Aragonese
hye	Armenian
ast	Asturian
aze	Azerbaijani
ban	Balinese
bak	Bashkir
eus	Basque
bar	Bavarian
bel	Belarusian
bos	Bosnian
bre	Breton
bul	Bulgarian
cat	Catalan
ceb	Cebuano
che	Chechen
chv	Chuvash
hrv	Croatian

Table A1. *Cont.*

Code	Language_Name
ces	Czech
dan	Danish
diq	Dimli
nld	Dutch
arz	Egyptian Arabic
eml	Emiliano-Romagnolo
eng	English
epo	Esperanto
ekk	Standard Estonian
est	Estonian (macrolanguage)
fao	Faroese
fin	Finnish
vls	Flemish
fra	French
fry	Frisian
glg	Galician
lug	Ganda
deu	German
gom	Goan Konkani
ell	Greek
hat	Haitian Creole
hau	Hausa
heb	Hebrew
hin	Hindi
hun	Hungarian
isl	Icelandic
ido	Ido
ind	Indonesian
ina	Interlingua
pes	Iranian Persian
gle	Irish
ita	Italian
jav	Javanese
kaz	Kazakh
kin	Kinyarwanda
kor	Korean
kur	Kurdish
kir	Kyrgyz
lat	Latin
lav	Latvian
lim	Limburgish
lit	Lithuanian
lmo	Lombard
nds	Low German
ltz	Luxembourgish
mkd	Macedonian
mlg	Malagasy
msa	Malay
mlt	Maltese
mri	Maori
mzn	Mazanderani
mhr	Meadow Mari
nan	Min Nan
min	Minangkabau
mwl	Mirandese

Table A1. *Cont.*

Code	Language_Name
lus	Mizo
mon	Mongolian
azj	North Azerbaijani
uzn	Northern Uzbek
nor	Norwegian
nob	Norwegian Bokmal
nno	Norwegian Nynorsk
oci	Occitan
pap	Papiamento
pus	Pashto
fas	Persian
pms	Piemontese
plt	Plateau Malagasy
pol	Polish
por	Portuguese
ron	Romanian
roh	Romansh
rus	Russian
sco	Scots
srp	Serbian
hbs	Serbo-Croatian
sna	Shona
scn	Sicilian
szl	Silesian
snd	Sindhi
slk	Slovak
slv	Slovenian
som	Somali
spa	Spanish
lvs	Standard Latvian
zsm	Standard Malay
sun	Sundanese
swa	Swahili (macrolanguage)
swh	Standard Swahili
swe	Swedish
gsw	Swiss German
tgl	Tagalog
tgk	Tajik
tat	Tatar
tur	Turkish
tuk	Turkmen
ukr	Ukrainian
hsb	Upper Sorbian
urd	Urdu
uig	Uyghur
uzb	Uzbek
vec	Venetian
vie	Vietnamese
wln	Walloon
war	Waray
cym	Welsh
pnb	Western Panjabi
sah	Yakut
yid	Yiddish
zul	Zulu

References

Ardehkhani, P., Ardehkhani, P., & Hooshmand, H. (2024, February 21–22). *ViT-pmn: A vision transformer approach for persian numeral recognition*. 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), Babol, Iran.

Beier, S., & Larson, K. (2010). Design improvements for frequently misrecognized letters. *Information Design Journal (IDJ)*, *18*(2), 118–137. [CrossRef]

Bernard, B. (2016). *Continuous-time repeated games with imperfect information: Folk theorems and explicit results* [Ph.D. thesis, University of Alberta].

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. [CrossRef]

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *The American Naturalist*, *167*(5), E117–E139. [CrossRef] [PubMed]

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494. [CrossRef]

Dan, Y., Zhu, Z., Jin, W., & Li, Z. (2022). PF-ViT: Parallel and fast vision transformer for offline handwritten chinese character recognition. *Computational Intelligence and Neuroscience*, *2022*(1), 8255763. [CrossRef]

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., & Alabdulmohsin, I. (2023, July 23–29). *Scaling vision transformers to 22 billion parameters*. International Conference on Machine Learning, Honolulu, HI, USA.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, arXiv:2010.11929.

Geyer, L. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*(5), 487–490. [CrossRef]

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. [CrossRef]

Hewavitharana, S., & Vogel, S. (2013). Extracting parallel phrases from comparable data. In *Building and using comparable corpora* (pp. 191–204). Springer.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. [CrossRef]

Meletis, D., & Dürscheid, C. (2022). *Writing systems and their use: An overview of grapholinguistics*. de Gruyter.

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

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, *139*(1), 19–37. [CrossRef]

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. [CrossRef] [PubMed]

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillére, B. é. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 250.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. [CrossRef]

Perfetti, C. A., & Hart, L. (2008). The lexical quality hypothesis. In *Precursors of functional literacy* (pp. 189–213). John Benjamins Publishing Company.

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. [CrossRef]

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372. [CrossRef]

Sampson, G. (2016). Typology and the study of writing systems. *Linguistic Typology*, *20*(3), 561–567. [CrossRef]

Scaltritti, M., Miniukovich, A., Venuti, P., Job, R., De Angeli, A., & Sulpizio, S. (2019). Investigating effects of typographic variables on webpage reading through eye movements. *Scientific Reports*, *9*(1), 12711. [CrossRef]

Shana, A., Putu, S. N., & Putri, D. P. S. (2024). Hangul character recognition of a new hangul dataset with vision transformers model. *SINTECH (Science and Information Technology) Journal*, *7*(3), 190–202. [CrossRef]

Slattery, T. J., & Rayner, K. (2013). Effects of intraword and interword spacing on eye movements during reading: Exploring the optimal use of space in a line of text. *Attention, Perception, & Psychophysics*, *75*(6), 1275–1292. [CrossRef] [PubMed]

Sproat, R., Shih, C., Gale, W. A., & Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational linguistics*, *22*(3), 377–404.

Tinker, M. A. (1963). Influence of simultaneous variation in size of type, width of line, and leading for newspaper type. *Journal of Applied Psychology*, *47*(6), 380. [CrossRef]

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61–88. [CrossRef]

Xue, N. (2003). Chinese word segmentation as character tagging. *International Journal of Computational Linguistics & Chinese Language Processing*, *8*(1), 29–48.

Zhou, Z., Xu, H.-M., Shu, Y., & Liu, L. (2024, June 16–22). *Unlocking the potential of pre-trained vision transformers for few-shot semantic segmentation through relationship descriptors*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA.
Zuidema, W., & De Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*(2), 125–144. [CrossRef]