Simplified Chinese Character Distance Based on Ideographic Description Sequences

Yixia Wang and Emmanuel Keuleers
Tilburg University
Warandelaan 2, 5037 AB Tilburg
y.wang_1@tilburguniversity.edu, e.a.keuleers@tilburguniversity.edu

Abstract

Character encoding systems have long overlooked the internal structure of characters. Ideographic Description Sequences, which explicitly represent spatial relations between character components, are a potential solution to this problem. In this paper, we illustrate the utility of Ideographic Description Sequences in computing edit distance and finding orthographic neighbors for Simplified Chinese characters. In addition, we explore the possibility of using Ideographic Description Sequences to encode spatial relations between components in other scripts.

Keywords: Ideographic Description Sequences, Character distance, Character Neighbors

1. Introduction

Storage and communication of written text using digital computers requires conventions for encoding characters. Early efforts at establishing encoding standards were driven by practicality and economy of space. Developed in 1963, the American Standard Code for Information Interchange (ASCII; American National Standards Institute, 1995) lies at the basis of most character encoding systems in use today. ASCII uses a 7-bit encoding, with 32 of the 128 positions allocated to communication control characters and the other 96 reserved for numbers, upper- and lowercase letters of the English alphabet, and punctuation. As computer technology spread, ASCII was succeeded by ISO-8859 (ISO/IEC, 1987) which, with 8-bit encoding and language specific versions, enabled the encoding of characters for a wider variety of alphabetic writing systems (e.g., ISO 8859-5 for Cyrillic, ISO-8859-11 for Thai). Accommodation for storing the many characters used in CJK (Chinese, Japanese, Korean) writing systems, came with the Unicode Standard, with different variants allowing for up-to 32-bit encoding (>4 billion characters).

The legacy of ASCII lead to these successive standards allocating more and more space for individual characters instead of incorporating *compositionality*, which is a design feature of most writing systems (English writing being a notable exception). For instance, for writers of French it is understood that most vowels can be accented, yet ISO-8859-1 has different slots for â, ê, î, ô, and û; á, é, í, ó, and ú; etc. For other writing systems, such as Chinese, compositionality is the norm, rather than the exception.

Recognizing that it was necessary to represent characters that do not have a dedicated slot, such as rare or novel Chinese characters, Unicode 15.1 (Unicode Consortium, 2023) introduced Ideographic Description Sequences (IDSs) as a principled approach to encoding characters compositionally.[1] Figure 1 shows the rendition of a rare character using an IDS.

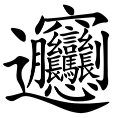

Figure 1: Chinese character *biang1* ’the sound of slapping and kneading noodles during noodle-making’. The ideographic description sequence for this character is ⿺⻌⿱穴⿲月⿱⿲幺言幺⿲長馬長 刂心.

Because of their ability to encode and represent characters compositionally, IDSs have a wide range of applications. In this paper, we will focus on a novel application, namely the use of IDSs to compute distance between Chinese characters. As section 2 will show, psycholinguistic literature has demonstrated that the identification of written words is influenced by their orthographic neighbors. Determining the orthographic neighbors of a word requires the ability to compute distances between any pairs of words. This is relatively straightforward to do in a language such as English, because most words do not have a hierarchical structure and characters are not compositional. It is far more difficult to do for Chinese characters.

Section 2 introduces related work on how diacritics and spatial relations influence word processing in various writing systems, providing theoretical background to support their explicit representations. Section 3 demonstrates a practical appli-

[1] Although the term *ideographic* is widespread, it is inaccurate. Chinese writing, specifically, is considered morphosyllabic (DeFrancis, 1989; Gorman and Sproat, 2023).

cation of IDSs in measuring distance between Chinese characters.

2. Related Work

2.1. Visual Processing of Diacritics

The use of diacritics is probably the best known application of compositionality in writing characters. A diacritic is usually defined as a glyph added to a character for pronunciation modification (Daniels and Bright, 1996). Evidence suggests that the processing of characters with diacritics depends on language features (Labusch et al., 2023). Ayçiçeği and Harris (2002) conducted a rapid serial visual presentation (RSVP) experiment in Turkish, showing more repetition blindness for words differing in a diacritic (işim- isim) as opposed to orthographic neighbours (ilim - isim), suggesting that characters with and without diacritics share the same mental representation. Perea et al. (2016) demonstrated that diacritic marks were quickly processed by the cognitive system during the early stages of word processing in Arabic, a script that is characterized by diacritical marks, position-dependent allography, and its cursive nature. Chetail and Boursain (2019), on the other hand, found that diacritic letters did not share the same abstract representations with their pure counterparts in French, where diacritic marks are predominantly observed on vowels. Marcet et al. (2022) found similar evidence for diverging abstract representations in Catalan, a language with complex grapheme-to-phoneme mappings.

2.2. Modeling Compositionality in Visual Processing

The Recognition by Components model (Biederman, 1987), asserted the significance of structural representations in object recognition. According to the model, the visual system recognizes an object by analyzing spatial arrangements of basic geometric shapes, such as cubes and cones. Transferring this to the domain of character recognition (Grainger et al., 2008), the relative positioning of components in a character is an important indicator of visual characteristics, helping to distinguish between characters (Lu et al., 2002). For example, the Chinese character 音 *yin1* 'sound' is distinguishable from another 昱 *yu4* 'bright' only by the relative positioning of components. The same is true for the diacritic letters ṡ and ṣ in the orthography of Yoruba in Nigeria. Arguably, even letters of Ъ and Б in Cyrillic script are compositionally similar, with a more nuanced difference in line orientation.

Other models focusing on the function of spatial relations in visual processing include Gestalt principles (Köhler, 1967; Todorovic, 2008) and feature integration theory (Treisman and Gelade, 1980).

3. Character Distance in Simplified Chinese Script

Ideographic description sequences were created to encode the spatial arrangement of components for CJK Unified Ideographs.[2] Unicode 15.1 (Unicode Consortium, 2023) defines eighteen ideographic description characters (IDCs), twelve of which are commonly used (Table 1).

An IDS consists of an IDC followed by its arguments, which can be either ideographs or another IDC. For instance, the IDS for the character 英 *ying1* 'blossom' is ⿱艹央, where ⿱ signifies top-down arrangement of the arguments 艹 and 央. Because the number of arguments to an IDC is always known, IDSs allow for nesting and concatenation. The ability to nest IDCs makes it possible to render complex spatial arrangements. For instance, the IDS for the character 嚻 *xiao1* 'a mythic beast' is ⿲⿱口口頁⿱口口.

When considering *distance* between the simplified Chinese characters 芍 *shao2* 'peony', 顶 *ding3* 'roof', and 英 *ying1* 'blossom', one approach would be to say that they are all different characters. Another approach could consist of noting that 芍 and 英 are both vertically arranged and have the element 艹 in common, whereas 顶 has no similarities to the other two characters, either in layout or components.

Existing methods for character similarity for Chinese characters can be divided in two main types: stroke-based and, more commonly, component-based. An abundance of literature defines the degree of character similarity based on shared component(s). In single-component comparison (Leck et al., 1995; Chen and Juola, 1982; Yeh and Li, 2002; Perfetti and Zhang, 1991), radicals (e.g., 蕉 *jiao1* 'banana' & 荐 *jian4* 'to recommend') or phonetic components (e.g., 煤 *mei2* 'coal' & 谋 *mou2* 'to plan') are used. Less often, smaller stroke patterns (e.g., 兑 *dui4* 'to exchange' & 分 *fen1* 'to divide'; Liu and Lin, 2008) and structural information (e.g., 啄 *zhuo2* 'to peck' & 偌 *ruo4* 'such'; Yeh et al., 1997) are used. Most of these methods define similarity in a binary way: a pair of characters is either similar or it is not. In the following sections, we propose an alternative method which is based on using edit distance on fully-decomposed IDSs and compare it to the existing approaches.

[2] CJK Unified Ideographs refers to a shared set of characters used in the writing systems of Chinese, Japanese, and Korean languages, all of which incorporate Han characters and their variations. CJKV extends the scope to include Vietnamese, which historically used Han characters.

IDC	Unicode	Name	Example	Example IDS
⿰	U+2FF0	Ideographic Description Character Left to Right	作	⿰亻乍
⿱	U+2FF1	Ideographic Description Character Above to Below	思	⿱田心
⿲	U+2FF2	Ideographic Description Character Left to Middle and Right	街	⿲彳圭亍
⿳	U+2FF3	Ideographic Description Character Above to Middle and Below	帚	⿳彐冖巾
⿴	U+2FF4	Ideographic Description Character Full Surround	回	⿴囗口
⿵	U+2FF5	Ideographic Description Character Surround from Above	网	⿵冂⿻㐅㐅
⿶	U+2FF6	Ideographic Description Character Surround from Below	凶	⿶凵㐅
⿷	U+2FF7	Ideographic Description Character Surround from Left	区	⿷匚㐅
⿸	U+2FF8	Ideographic Description Character Surround from Upper Left	庆	⿸广大
⿹	U+2FF9	Ideographic Description Character Surround from Upper Right	句	⿹勹口
⿺	U+2FFA	Ideographic Description Character Surround from Lower Left	这	⿺辶文
⿻	U+2FFB	Ideographic Description Character Overlaid	巫	⿻工从

Table 1: The table provides IDCs, their Unicode, names, example characters, and Ideographic description sequences for the character.

3.1. Character distance using fully decomposed IDSs

We retrieved a dataset with IDSs for Chinese characters from an online repository[3], which in turn was derived from the Character Information Service Environment (CHISE) IDS database[4] (Morioka and Wittern, 2002). Part of the open-source CHISE project to expand general-purpose coded character sets, the IDS database contains most of the CJKV Unified Ideographs of ISO/IEC 10646 (Morioka, 2015).

To limit the list to Chinese characters used in mainland China, we selected only the 20,830 characters documented in Xinhua Dictionary (http://xh.5156edu.com). Then, we normalized the selected IDSs by recursively replacing components that could be further decomposed by their corresponding IDS. The result was a set of fully decomposed IDSs. Inspection of the resulting IDSs showed that, in addition to the 12 IDCs, only 545 basic characters were required to encode the over 20,000 selected characters.

Levenshtein distance (LD) is defined as the number of insertions, deletions, and substitutions operated on a string to turn it into another string (Levenshtein, 1966). Inspired by Kruskal (1983), we gave the substitution a cost of 2 and the other two operations a cost of 1 (see Figure 2).

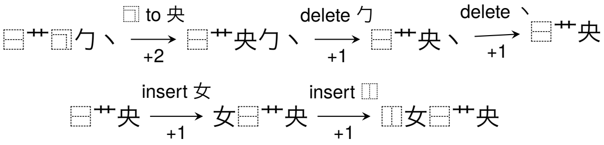

Figure 2: The first example illustrates substitution and deletion. Converting 芍 *shao2* ’peony’ (IDS: ⿱艹⿹勺丶) to 英 *ying1* ’blossom’ (IDS: ⿱艹央) involves one substitution and two deletions, resulting in an edit distance of 4. The second example illustrates insertion: Transforming 英 to 媖 *ying1* beauty (IDS: ⿰女⿱艹央) requires the insertion of ⿰ and 女, resulting in an edit distance of 2.

3.1.1. IDS distance vs methods based on shared components

Some Chinese characters incorporate the same radicals and residuals: 案 *an4* ’instance’ & 桉 *an1* ’the eucalyptus tree’, 召 *zhao4* ’to summon’ & 叨 *dao1* ’to chatter’, and 峯 *feng1* ’peak’ & 峰 *feng1* ’summit’. When similarity is based on shared components as in the examples (i.e., 吃 *chi1* ’to eat’, 员 *yuan2* ’member’, 哲 *zhe2* ’philosophical’, and 加 *jia1* ’to add’) provided in the work of Yeh and Li (2002), these pairs are identical because they all have the same components. This method falls short with respect to structural differences.

Figure 3 shows the IDS distance for the same characters. The IDS distance between 案 and 桉 is 4, equal to that between 召 & 叨, whereas closer are 峯 & 峰, which are only different in layout and have a distance of 2.

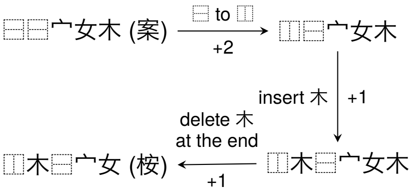

Figure 3: The edit distance of 案 & 桉 is 4, summing up 1 substitution, 1 insertion and 1 deletion.

[3] https://github.com/cjkvi/cjkvi-ids

[4] https://gitlab.chise.org/CHISE/ids

3.1.2. IDS distance vs methods based on radical-level shared components and character structures

The spatial arrangements of components in Chinese characters are highly correlated with their functions (semantic or phonetic). In various applications, characters that have identical structure and shared components are considered similar and selected as stimuli (Leck et al., 1995; Chen and Juola, 1982). Hence, these methods do not identify similarity among characters sharing both structure and components, but not the function of components. For instance, 杏 *xing4* ’apricot’ and 呆 *dai1* ’dull’ have different component order and semantic radical, but are otherwise identical. Figure 4 shows how, in comparison, IDS-based distance addresses this.

$$\text{⿱木口 (杏)} \xrightarrow[+1]{\text{delete 木}} \text{⿱口} \xrightarrow[+1]{\substack{\text{insert 木}\\ \text{at the end}}} \text{⿱口木 (呆)}$$

Figure 4: The edit distance of 杏 & 呆 is 2 via one deletion and one insertion.

3.1.3. IDS distance vs methods based on sub-radical shared components and character structures

Liu and Lin (2008) go beyond the radical - residual level and explore smaller stroke patterns in computing similarity between Chinese characters. They decompose a character into a set of 24 basic elements defined in the Cangjie code by Chu (1979). A character is represented by its structure (one of nine layout patterns, encoded as a real value) followed by up to three components. For example, 相 *xiang4* ’appearance’ has a representation of ’2 (layout code) - 木 (part 1) - 月山 (part 2)’. Although this approach goes a long way toward addressing the compositionality of characters in a principled manner, the limited basic components do not allow for an unambiguous specification of characters. In other words, in many cases the decomposition does not allow for recomposition of the original characters. Using only nine layout patterns is also insufficient, as Simplified Chinese characters can be as complex as encompassing up to 32 strokes (e.g, 龘 *da2* ’depicting the majestic soaring of a dragon’). Instead, using IDS, we encode character structures by explaining structural information between just two or three components predetermined by IDCs. This granularity is also a reason why our method produce faithful results.

The IDS representation does not have structural ambiguity between sequences. In the few cases where we have found characters to share the same IDS representation (56 out 20,830), this concerned historical variants of the same character with slightly different stroke variants. Figure 5 shows the IDS distance between 相 *xiang4* ’appearance’, 箱 *xiang1* ’case’, and 想 *xiang3* ’to miss’, which according to Liu and Lin (2008) would be dissimilar. It may seem to be surprising that the IDS distance between 箱 and 想 is so small, but, in addition to overlapping components, these two are both vertical characters hierarchically enclosing a horizontal character.

$$\text{⿰木目 (相)} \xrightarrow[+1]{\text{insert 竹}} \text{竹⿰木目} \xrightarrow[+1]{\text{insert ⿱}} \text{⿱竹⿰木目 (箱)}$$

$$\text{⿰木目 (相)} \xrightarrow[+1]{\text{insert ⿱}} \text{⿱⿰木目} \xrightarrow[+1]{\text{insert 心}} \text{⿱⿰木目心 (想)}$$

$$\text{⿱竹⿰木目 (箱)} \xrightarrow[+1]{\text{delete 竹}} \text{⿱⿰木目} \xrightarrow[+1]{\text{insert 心}} \text{⿱⿰木目心 (想)}$$

Figure 5: The figure shows the process to get edit distance among characters 相, 箱, and 想. We turn 相 into 箱 via two insertions, as is the case for 相 & 想. Converting 箱 to 想 requires one deletion and one insertion.

3.1.4. Three basic elements to compute character similarity

It seems that, from the above-mentioned examples, it is next to impossible to evaluate the degree of character similarity without the integration of three basic elements:

- components (stroke patterns or smaller), as opposed to single-component comparison.
- layouts governing components, in comparison with character structure possibly as a result of single component comparison.
- relative positions of components.

We suggest that if we want to claim that characters are similar, we need to make these three elements explicit.

One of the limitations of using IDS to compute edit distance is its comprised ability to differentiate relative positions of components in some cases. For example, characters 呆 *dai1* ’dull’, 宋 *song4* ’a surname’, and 告 *gao4* ’to sue’ have the same edit distance of 2, but while we rule in component position, 告 should be further away from 呆, as their shared component is located differently. In order to differentiate the effect of component order, one possible method is to increase the weight of insertion. However, this would lead to an asymmetry of pairwise distances, which requires further modification.

杳		鬱	
Neighbors	Distance	Neighbors	Distance
杳	0	鬱	0
查	2	鬰	6
査	2	爩	10
杏	2	鬯	12
𣎵	2	楙	18
杢	2	傻	19
李	2	糉	20
旮	2	鍐	20
木	2	菡	20
杰	2	涵	20
桼	2	燓	20
𣅀	2	椷	20
呆	2	樊	20
杲	2	鐢	20
妟	2	函	20
柰	2	乘	21
早	2	兇	21
旯	2	冟	21
日	2	爻	21
旦	2	壱	21

Table 2: Neighbors and their pairwise distance to 杳 (orthographically simple and dense) and to 鬱 (orthographically complex and distinct).

3.2. Character Neighbors

By computing character distance, it is possible to cluster orthographically similar characters by exhausting pairwise distances among all characters and sorting the result. Table 2 provides twenty nearest neighbors for 杳 *yao3* 'dim' and 鬱 *yu4* 'lush and growing abundantly'.

However, this could be problematic, as distance is modulated by sequence length. For example, character 鬱 is closer to 匕 *bi3* 'spoon' (distance: 23) than 礬 *fan2* 'alum' (distance: 24), although the latter may seem to be more similar due to identical structures and more common components. The reason is that 匕 (IDS: ⿺乚丿, length: 3) requires only addition to transform into the target character 鬱 (IDS: ⿳⿲木⿱㐅⿴丿乀木冖⿰⿱⿶凵⿴㐅⿱丶⿰丶丶丶⿺乚丿彡, length: 26), while 礬 (IDS: ⿱⿱⿲木⿱㐅⿴丿乀木⿴一人⿸⿱一丿口, length: 18), with longer sequences, requires additional deletion.

To address this, we normalize distance as follows: Let N_i be the length of IDS of character C_i and N_j be the length of IDS of character C_j:

$$max\ distance = \min(N_i, N_j) \times 2 + |N_i - N_j|$$

where *max distance* represents the upper bound of the cost from possible operations. The distance metric can then be normalized by calculating the relationship between the cost of operations actually used and the maximum possible costs. The normalized distance is calculated as:

$$normalized\ distance = \frac{edit\ distance}{max\ distance}$$

where *normalized distance* indicates a measure where lower values signify higher operational congruence and thus closer distance.

Twenty closest neighbors for 鬱 based on normalized distance are given in Table 3. Note that the normalized distance of 鬱 and and 礬 is 0.545, smaller than that of 鬱 and 匕, though not shown in the table, at 0.793. We can see that the adjusted distance reveals a cluster of character pattern that is closer to human intuition.

4. IDS as a general approach to expressing component relations in any script

There are some advantages of the ideographic description sequences. First, they have the potential to be used to describe attested compositional characters in any script. For instance, French *café* could be represented as (c, a, f, ⿱, ´, e), with the description character ⿱ indicating that the two subsequent elements are to be arranged top-down. Second, they provide the possibility to form new compositional characters. Finally, when words are represented using concatenated ideographic description sequences, they allow for more accurate measurement of word similarity. For instance, in French, the word pâte can be considered to differ by one character from both pate and pâté, but in the same way it can also be considered to differ by the absence or presence of a diacritic on one of the characters.

In practice, the arguments to a particular IDC are quite predictable. For example, the IDC ⿰ almost always has a semantic radical as its left component and a phonetic residual as its right component. The spatial rendering thus typically also encodes a specific relationship. Taking this idea further, we can consider an IDC as a way of connecting a specific type of linguistic relationship to a specific spatial rendering (e.g., morphological, semantic, ontological). For instance, Table 4 shows how one could consider the components of compound words as arguments to a relationship operator which horizontally concatenates the components. This would allow distinguishing compounds from non-compounds, at least in the underlying sequence. But instead of horizontal arrangement, we could also replace the horizontal IDC with an equivalent vertical IDC to achieve a different kind of representation. In some applications, English text could then be rendered as in Figure 6. On top of this, there are also a wide range of other possible applications for IDS: creating novel

Neighbors	鬱	鬰	爩	鬯	鏤	僂	樓	蓾	滷	⿰木鹵
Normalized Distance	0	0.125	0.192	0.3	0.442	0.463	0.463	0.5	0.5	0.5
Neighbors	磠	塷	鏀	燓	樊	鐢	鹵	棥	[illegible]	鹶
Normalized Distance	0.5	0.524	0.524	0.526	0.526	0.526	0.526	0.529	0.545	0.545

Table 3: Twenty neighbors to 鬱 based on normalized distance. Note that after adjusting for the complexity level of the characters, the result is closer to human intuition.

Compound word	IDS horizontal	IDS vertical	Representation vertical
red dwarf	⿲, red, , dwarf	⿱, red, dwarf	red dwarf
red-blooded	⿲, red, -, blooded	⿱, red, blooded	red blooded
redhead	⿰, red, head	⿱, red, head	red head

Table 4: Table shows three compound words, IDS for their original forms, IDS for vertical placements, and resulting vertical renditions.

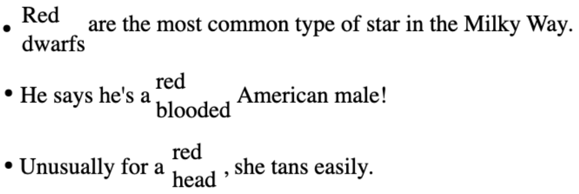
• Red dwarfs are the most common type of star in the Milky Way.
• He says he's a red blooded American male!
• Unusually for a red head , she tans easily.

Figure 6: A demonstration of rendering compound words in vertical layouts. Example sentences were retrieved from online Cambridge Dictionary (https://dictionary.cambridge.org).

Character	Script / Language	IDS
𞤱̌	Adlam	⿱, ◌̌, 𞤱
Ъ̀	Cyrillic	⿱, `, Ъ
님	Korean	⿱, ⿰, ㄴ, ㅣ, ㅁ
Ͽ	Greek	⿴ , ɔ, ·
Ŀ	Latin	⿺, L , ·
Đ	Latin	⿻, D, -
Ⱡ	Latin	⿻, L, ⿱, -, -
Æ	Latin / Cyrillic	⿰, A, E
Ǣ	Latin	⿱, -, ⿰, A, E
Ǟ	Latin	⿳, -, ◌̈, A
ஊ	Tamil	⿺, உ, ள

Table 5: Example characters represented as IDS in several scripts like Adlam, Cyrillic, Korean, Greek, Latin, and Tamil.

sequences; creating or adapting representations for under-resourced languages; rendering linguistic relationships spatially; substituting layouts, etc.

Examples of character IDS application in different scripts are shown in Table 5. While we use existing IDCs designed for Simplified Chinese characters in these examples, specific IDCs may need to be created to allow for script characteristics.

5. Conclusion

In this paper, we demonstrated that IDSs can be used to more precisely calculate edit distance and orthographic neighbors for Simplified Chinese characters. In addition, we explored the possibility of using IDSs to typographically represent morphological relationships. While Unicode currently only uses IDSs for CJK writing systems, the ability to represent characters compositionally gives IDSs a wide range of application beyond these scripts. In this way, representing characters and words using IDSs can offer methodological improvements in several areas.

References

American National Standards Institute. 1995. 7-bit American national standard code for information interchange. *Standards Action*. Retrieved February 6, 2024, from https://webstore.ansi.org/standards.

Ayse Ayçiçeği and Catherine L. Harris. 2002. How are letters containing diacritics represented? Repetition blindness for Turkish words. *European Journal of Cognitive Psychology*, 14(3):371–382.

Irving Biederman. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.

Hsuan-Chih Chen and James F Juola. 1982. Dimensions of lexical coding in Chinese and English. *Memory & Cognition*, 10:216–224.

Fabienne Chetail and Emeline Boursain. 2019. Shared or separated representations for letters with diacritics? *Psychonomic Bulletin & Review*, 26(1):347–352.

Bong-Foo Chu. 1979. Laboratory of Chu Bong-Foo. Retrieved Februrary 6, 2024, from http://www.cbflabs.com.

Peter T Daniels and William Bright. 1996. *The world's writing systems*. Oxford University Press.

John DeFrancis. 1989. *Visible speech: The diverse oneness of writing systems*. University of Hawaii Press.

Kyle Gorman and Richard Sproat. 2023. Myths about writing systems in speech & language technology. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 1–5.

Jonathan Grainger, Arnaud Rey, and Stéphane Dufau. 2008. Letter perception: from pixels to pandemonium. *Trends in Cognitive Sciences*, 12(10):381–387.

ISO/IEC. 1987. ISO/IEC 8859: 8-bit character encodings. International Organization for Standardization and International Electrotechnical Commission.

Wolfgang Köhler. 1967. Gestalt psychology. *Psychologische Forschung*, 31(1):XVIII–XXX.

Joseph B Kruskal. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237.

Melanie Labusch, Stéphanie Massol, Ana Marcet, and Manuel Perea. 2023. Are goats chèvres, chévres, chēvres, and chevres? Unveiling the orthographic code of diacritical vowels. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(2):301–319.

Kwong Joo Leck, Brendan S Weekes, and May Jane Chen. 1995. Visual and phonological pathways to the lexicon: Evidence from Chinese readers. *Memory & Cognition*, 23:468–476.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Yi-Chen Lin, Hsiang-Yu Chen, Yvonne C Lai, and Denise H Wu. 2015. Phonological similarity and orthographic similarity affect probed serial recall of Chinese characters. *Memory & Cognition*, 43:538–554.

Chao-Lin Liu and Jen-Hsiang Lin. 2008. Using structural information for identifying similar Chinese characters. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08*, page 93, Columbus, Ohio. Association for Computational Linguistics.

Qin Lu, Shiu Tong Chan, Yin Li, and Ngai Ling Li. 2002. Decomposition for ISO/IEC 10646 ideographic characters. In *COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization*.

Ana Marcet, María Fernández-López, Ana Baciero, Albert Sesé, and Manuel Perea. 2022. What are the letters e and é in a language with vowel reduction? The case of Catalan. *Applied Psycholinguistics*, 43(1):193–210.

Tomohiko Morioka. 2015. Multiple-policy character annotation based on chise. *Journal of the Japanese Association for Digital Humanities*, 1(1):86–106.

Tomohiko Morioka and Christian Wittern. 2002. Developping of character object technology with character databases. *IPA result report*.

Manuel Perea, Reem Abu Mallouh, Ahmed Mohammed, Batoul Khalifa, and Manuel Carreiras. 2016. Do Diacritical Marks Play a Role at the Early Stages of Word Recognition in Arabic? *Frontiers in Psychology*, 7.

Charles A Perfetti and Sulan Zhang. 1991. Phonological processes in reading Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4):633.

I-Fan Su, Sin-Ching Cassie Mak, Lai-Ying Milly Cheung, and Sam-Po Law. 2012. Taking a radical position: evidence for position-specific radical representations in Chinese character recognition using masked priming erp. *Frontiers in Psychology*, 3:333.

Dejan Todorovic. 2008. Gestalt principles. *Scholarpedia*, 3(12):5345.

Anne M Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Unicode Consortium. 2023. *The Unicode Standard, Version 15.1.0*. The Unicode Consortium, South San Francisco, CA. ISBN 978-1-936213-33-7.

Senlin Xu, Mingfan Zheng, and Xinran Li. 2020. String comparators for Chinese-characters-based record linkages. *IEEE Access*, 9:3735–3743.

Su-Ling Yeh and Jing-Ling Li. 2002. Role of structure and component in judgments of visual similarity of Chinese characters. *Journal of Experimental Psychology: Human Perception and Performance*, 28(4):933.

Su-Ling Yeh, Jing Ling Li, I Chen, et al. 1997. The perceptual dimensions underlying the classification of the shapes of Chinese characters. *Chinese Journal of Psychology*.

Pong Chi Yuen, Guo-Can Feng, and Yuan Yan Tang. 1998. Printed Chinese character similarity measurement using ring projection and distance transform. *International journal of pattern recognition and artificial intelligence*, 12(02):209–221.

Xiaochen Zhang, Siqin Yang, and Minghu Jiang. 2020. Rapid implicit extraction of abstract orthographic patterns of Chinese characters during reading. *Plos one*, 15(2):e0229590.