CHINESE THE WRITE WAY: AN INNOVATIVE APPROACH TO TEACHING CHINESE CHARACTERS THROUGH STORIES BEHIND THE SCRIPTS

MQ. Zhao, A. Digh

Mercer University (UNITED STATES)

Abstract

The Chinese writing system is the only one in the world that has been used continuously for several thousands of years. The unique character set with thousands of distinct symbols (known as 字 zì) provides a common knowledge model for over a billion people speaking several major dialects, each of which is commonly recognized as a different language. However, throughout several major transformations, the pictorial features embedded in the scripts have changed to more abstract forms, which helped make Chinese the hardest language to learn. A new approach to teaching and learning Chinese, Chinese the WRITE Way, is proposed in this chapter, which emphasizes on conveying why a character is formed in its specific pattern and the compositional relationships between characters. A software application, CharActor, an ordinary reality that acts out language learning based on knowledge modeling, animation and text-to-speech is first demonstrated. The target enhanced reality (ER) system with GPT-type user interaction and personalized learning experience is proposed.

Keywords: Teaching Chinese the WRITE Way, Enhanced Reality, Knowledge Modeling, Personalized Learning.

1 INTRODUCTION

The Chinese writing system is the only one in the world that has been in use for several thousands of years. The shared writing system provides a common knowledge model for over a billion people speaking several major dialects, which are commonly recognized as different languages [Ghosh].

As compared to alphabetic languages with a few dozens of characters, the Chinese writing system (中文 zhōngwén) consists of tens of thousands distinct “characters” or 字 (zì). The “Standard Chinese Characters Table” (as supported by Word) includes some 8,015 distinct character forms, usually sufficient for general applications, while the Unicode count for mainland China alone is 65,941 [Wikipedia]. Each of these characters presents a unique two-dimensional pattern, which looks very different to people familiar with words spelt linearly with letters from a simple alphabet. No apparent hints for pronunciation of the symbol helps make foreign students clueless. All of these contributed to the recognition of Chinese being the most difficult to learn.

Typical approaches for teaching the Chinese language start from conversational contexts and then map the spoken words into their written forms. This can be referred to as the “speak way.” Used in China for as long as writing goes, this approach worked for school kids in that they would already have learned how to say their name and make daily conversations. What the teachers have to do is to show them “how” to write the words with characters. Without being given the pictorial hints (to be described in the next section) designed into the characters, students are asked to copy each character repeatedly, to learn the characters by rote. This inefficient way embodies its merits in training the hand-eye coordination to youngsters, as well as patience and habits to learn.

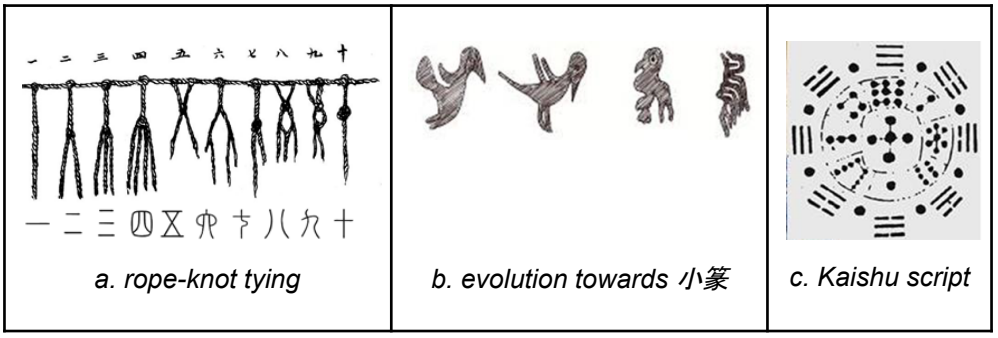
The nature of this chapter is to present the vision for an AI teaching tool specialized in the Chinese writing system. A survey on the evolution of the Chinese characters will be included along with the cultural aspects embedded in these symbols, so that the technical readers can understand the data modeling and management needs and the use cases to be implemented in an automated system with enjoyable user experiences.

2 BACKGROUND

The Chinese writing system is a major information management system used for thousands of years. Several changes have taken place, as illustrated in Fig. 1. However, documentations about the invention and transformation of the characters are at best incomplete.

Figure 1 – Transformation of 不 over time

Ancient people were said to have used rope-knot tying (Fig. 2a) to record events before a writing system was established. It is commonly accepted that the earliest systematic scripts, the Oracle Bone Scripts, matured in the Shang Dynasty. However, smaller amounts of symbols in similar shapes have been found in sites dated back up to 8000 years ago. As a logographic system, the scripts largely shared a pictorial tradition. Indeed, from the calligraphers' perspective, "Writing and painting are one" [Ouyang].



a. rope-knot tying | *b. evolution towards 小篆* | *c. Kaishu script*

Figure 2 – Origin of the Chinese Writing System

After the First Emperor of Qin unified China, he standardized the scripts into what is known as the small seal script (小篆 xiǎozhuàn). It was the work of Li Si (李斯) that created these new scripts, which still held the curving strokes and shapes of its predecessors (Fig. 2b). However, governing a country the size of China and making its laws and other written documents throughout required more clerks and an easier way to write.

The clerks drove the next major transformation and produced the clerical style, which became the standard in the Han dynasty, which followed the Qin dynasty. Strokes of the clerical style are more straightened, and later transformed into the so called Kaishu script (楷书 kǎishū) style (Fig. 2c), which is commonly known as the traditional script (since the 2nd century AD). To date, this character set is called 汉字(Hànzì), after the dynasty started to see this transformation into this more abstract style.

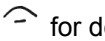
说文解字(shuō wén jiě zì) compiled by Xu Shen in the Han Dynasty (1st century) is the first dictionary of its kind in Chinese history. Under the title meaning "explain 文 (wén) and analyze 字 (zì)", [Xu] set up the standards for organizing the characters as used since the Han era with etymological roots in the ancient scripts. The differences between 文 (etymological roots) and 字 (Hanzi characters) can be seen from the ways the two characters are formed.

Table 1 – Singleton and Composite Characters

	A symbol formed with a number of strokes that can be used to convey knowledge.		Comes with two parts, 宀 house and 子 child: characters assembled with radicals to show new meanings.

In Shuowen Jiezi, six ways (六书) to create Chinese characters were identified as seen in Table 2.

Table 2 – Six Ways to Form Characters

Pictorial	pictograms like 木  depicting a tree with trunk, up-stretching branches, and roots	文 with single component
Indicative	self-explanatory symbols or ideograms: 上  indicating up, and 下  for down	
Associative	compounds (with radicals both representing meaning): 林 lín meaning a forest, with some trees. 林 = 木 + 木, both show meaning.	字, usually with two components, top-down, left-right, etc.
Picto-phonetic	one component representing meaning and the other indicating sound; accounted for some 95% of the characters: 妈 mā for mother, with 女 showing gender and 马 mǎ to show the sound.	
Borrowing	an existing character to represent new meanings	For vocab words here.
Extension	using another form to represent the same meaning	

Over 10,000 characters were organized into 540 sections, each listing one to a few hundred characters under a section header (also known as radical), like 木. All characters in a section are semantically related to the section header or radical, which is presented using a small seal script of Qin. For instance, the character 楷 (kǎi) is listed in the 木 section, meaning a kind of tree whose branches tend to be straight. As a photo-phenetic instance, its sound part or right-component, 皆 (jiē) indicates the sound, as it was pronounced back in the Tang Dynasty (gæi, which is remained in the Cantonese dialect gaai1 nowadays).

Shuowen Jiezi became the bridge between the Kaishu character set and the Jinwen/Seal scripts. It was still an incomplete “design document” due to the fact that the earlier Oracle Bone Scripts were not known until rediscovered in the turn of the 20th century. So Xu’s explanations demonstrated one consistent way to help readers understand the characters, while other “stories” with a current narrative may work even better.

To improve the literacy rate of the mass population, a systematic effort was made in 1956 to simplify the traditional character set, by drastically reducing the number of strokes used (e.g., 龟 vs. 龜 for turtle, with 7 vs. 16 strokes) and romanized phonetic symbols (known as pinyin) to denote the Mandarin pronunciation.

The simplified scripts further deviated from the pictorial features embedded in ancient scripts, while also basing on other variations, such as cursive stroking styles. Fig. 3a shows how the simplified character 东 can be traced back to the traditional form 東 through the cursive scripts. Characters using 東 as a phonetic component (such as 凍 dòng for freeze) are consequently simplified (such as 冻) (Fig.

3b). Consolidating characters sharing a “form” radical into one character can be seen in Fig. 3c, similar to the application of the “borrowing” rule from Table 3.

 *a. one way to simplify*	凍 ⇒ 冻 *b. simplified*	里 lǐ = 田+土　裏 and 裡 = 衣+里 Village field soil　inside inner-side cloth lǐ *c. simplifying to the shared radical*

Figure 3 - Tracing between Simplified and Traditional Characters

These practices have all made it harder to understand “why” a character is formed in a certain way. Although the traditional character set (as still used in places like Hong Kong, Taiwan, and overseas Chinese communities) carries more etymological tracings (though already limited), their simplified counterparts made the writing accessible to hundreds of millions and became the dominant form. Now with modern digitalization, characters are now strings of code to be processed by computers like any other kind of information. Following a long period of anxiety over digitizing and inputting Chinese characters, users can now input these symbols on a computer by either typing with a keyboard or drawing them out with a stylus, just like they would with their pencil or brush on paper. The original etymology is lost and continues to be forgotten over the years.

For new learners of the Chinese language, this does little to help them understand the language better. The purpose of the CharActER system is to show why a character is written in a certain form through compositional features described in “stories”; additional contents need to be produced and managed in its knowledge base. The stories will be based on “design documents” as provided in Shuowen [Xu] and the like, and can be generated at runtime by AI when taking into account the effectiveness learned from data in the registered user’s history. Though a search for any single character can return dozens or thousands of ideograph variations, the explanation on the why is still limited to copy the original wording from Shuowen, which is hard to grasp even by a general Chinese audience. For instance, a random search picked the character 臣 (“original meaning” as minister, https://hanziyuan.net/#臣), with a total 115 images returned. Without translating the Shuowen entry (牽也事君也象屈服之形) or linking the wording to the Seal script from Shuowen (), the reader cannot easily make the correlation. An analogy may best illustrate this concept: “a subject kneeling before the master with both hands supporting him to the sides (象屈服之形)”. The story, when rewritten in a language like English, can be *augmented* with animation such as highlighting the pen strokes for an ancient writing form or a more modern one, as well as visuals, to create the ultimate learning experience. Animating this scene in combination with some storytelling will make it easier to picture.

This change in the composition will not only benefit learners of the language. The unicode encoding system assigns a numerical value to the characters, regardless of the number of strokes. Hence, some modern software developers assume all Chinese characters are three bytes in UTF-8, and when going through a string of Chinese characters, one would simply skip forward by three bytes. Unfortunately, this did not hold true for all Chinese characters, such as those whose code points required 4-byte representations in UTF-8. If Chinese texts are misinterpreted by uniformly advancing the offset by three bytes, it could lead to security flaws [unicode.org].

Alternatively, as a conceptual modeling tool, ER models [Chen-1977] are widely used in database and knowledge base design. In [Chen-1997], the correspondence between the Chinese character construction and the ER modeling principles was discussed. Unfortunately, research on modeling and managing knowledge about the composition and recognition of Chinese characters as suggested by Chen were not followed. Consequently, even the leading AI, ChatGPT, failed to answer a simple “same-or-different” question about three commonly used Chinese characters.

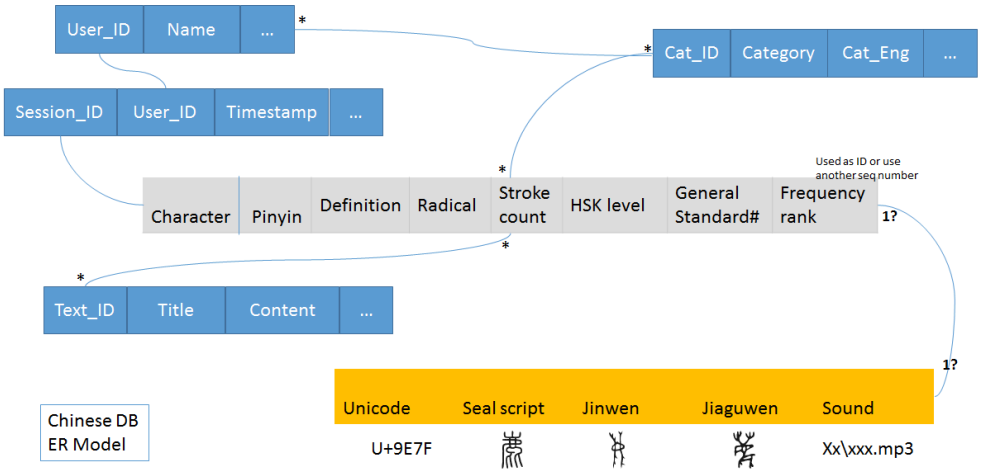


Figure 4 - Sample ER Model

As is commonly known when using ER modeling [Zhao], entity sets (the E) turn to become tables that store business data, relationship sets (the R) capture the mapping between entity instances which is at least as critical. In the case of modeling knowledge about the correspondence of form and meaning of Chinese characters, the entity types include stroke, radical, and character, etc. The relationship types that need to be used to infer the meaning of a character by its composition include:

- At the character level, for those with two components: need to track whether it is categorized as Associative or Picto-phonetic. This is referred to as the “knife” tool in the CharActER.
- At the character level, some characters are regarded as with only one component but actually has with “hidden” components (such as 东): need to trace the semantic hints in earlier scripts such as the traditional (東) or the more pictorial script types. This is referred to as the “telescope” or “lens” tool. “Stories” that convey the knowledge of character meaning (such as 東=日+木 and related explanations) need to be stored. Tracing as in Fig. 3a can be helpful to show the reason behind the simplification rule.
- At the sub-character level, such as for the composition of 土 (Fig. 5). Meaning embedded in each stroke may vary based on the relative position in the form (such as the upper and bottom 一 in 土)

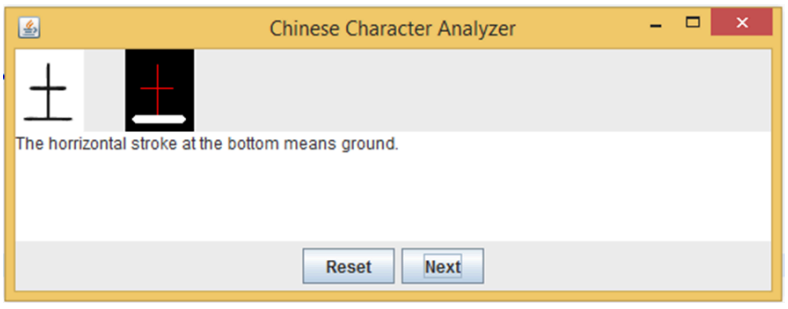


Figure 5 - Illustration for Contents Needed for Animation and Explanation

The 土 (tǔ for earth, soil) Episode: show both data modeling needs and user experience (UX) needs (Section IV).

- The bottom stroke indicates the land, with the vertical stroke depicting a plant.
- The upper horizontal stroke indicates top soil, which is what is needed for the plant to grow.
- Animation is used to show strokes in different colors: land=white, plant=green, and soil=yellow.

The “stories” are adapted from the explanations by [Xu] when applicable. It needs to be translated from ancient phrasing to modern Mandarin and the targeted foreign language, such as English. Versions based on a modern context will produce more retention.

In addition to modeling the knowledge embedded in the characters, relationships between characters also need to be managed. Typically, characters are first introduced and taught based on conversational contexts in the traditional “speak” way approach. Without following the dependencies between characters, it is difficult to explain the reasons why each character’s form would render the particular meaning. CharActER is designed to model the Chinese language through understanding its etymology.

3 CONCLUSIONS

This chapter explores the "Chinese the WRITE Way" approach, which treats Chinese writing as a system of knowledge and characters as its foundation. It involves teaching characters through modern artificial intelligence and information technology techniques. It shows how to create exercises based on characters and vocabulary to help users develop a greater understanding of both language and culture. The model is designed to establish connections between components within characters, between characters, and their original meanings, making it easier to appreciate cultural aspects.

Mastering a new language takes effort and dedication, with no shortcuts to fluency. It involves understanding and processing a set of intricate rules and regulations over time. It is something that the human brain has been perfecting since the dawn of our species. Machines which can help humans learn a second language are now a reality using a variety of tools driven by artificial intelligence, enhanced realities, and genetic algorithms. In particular, the AI chatbot ChatGPT is capable of simulating a real Chinese tutor or conversation partner in helping you with both pronunciation and communication skills. As Chomsky writes in 2023, “These programs have been hailed as the first glimmers on the horizon of artificial general intelligence — that long-prophesied moment when mechanical minds surpass human brains not only quantitatively in terms of processing speed and memory size but also qualitatively in terms of intellectual insight, artistic creativity and every other distinctively human faculty.” In the years to come, it will be very important to conduct research into the results of using ChatGPT software in learning Chinese both by university students as well as children. The way a child's brain is wired is not at all similar to that of an adult nor the workings of a machine learning program.

Possible future research may also involve utilizing graph and/or fuzzy database modeling to manage uncertainty, individualized narrative content production, crafting metrics for quantifying student proficiency as well as evaluating productivity, and optimizing user experiences with the help of virtual personas (known as metaverse avatars) along with extended reality.

ACKNOWLEDGEMENTS

The authors are grateful to Adam Griggs of Mercer University’s Tarver Library for his help in finding resources. An extra thank-you goes to Beth Stewart, who suggested the WRITE Way (in place of the Right Way); to Maria Cristina Petruso, who suggested ER in CharActER is also related with ER as in ER modeling. We also appreciate Jean J. Cheng for her insights and discussions; Gang Ding for his conversations about ChatGPT; and Becky Pirkle for inspiring the teaching plan made for the ACE school (in Appendix C). Teaching my GenEd course section subtitled “Understanding the Chinese Culture within a Global Context” at Mercer, along with several presentations at Mercer and Wesleyan College, contributed to much of what is covered in this chapter.

REFERENCES

[1] Chinese Text Project (ctext.org) dictionary [Retrieved: May 2023].

[2] Chinese character classification on Wikipedia (https://en.wikipedia.org/wiki/Chinese_character_classification), [Retrieved: May 2023].

[3] J. Biggs, “ChineseCubes Are Cubes That Teach You Chinese,” 2014, https://techcrunch.com/2014/04/21/chinesecubes-are-cubes-that-teach-you-chinese/amp/, [Retrieved: May 2023].

[4] L. Booth, “China’s ChatGPT Black Market Is Thriving,” 2023, https://vervetimes.com/chinas-chatgpt-black-market-is-thriving/, [Retrieved: May 2023].

[5] T. Chamorro-Premuzic, “I, Human,”, Harvard Business Review Press: Boston, MA, USA 2023.

[6] G. Chen, “Fuzzy Logic in Data Modeling,” Kluwer Academic Publishers, 1998.

[7] P. P. Chen, “The Entity-Relationship Model: toward a unified view of data,” ACM TODS 1 (1976), pp 9-36.

[8] P. Chen, “English, Chinese and ER Diagrams,” Data & Knowledge Engineering 23 (1997), pp 5-16.

[9] N. Chomsky, I. Roberts, J. Watumull, “Noam Chomsky: The False Promise of ChatGPT,” *New York Times,* 8 Mar. 2023, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html?smid=url-share, [Retrieved: May 2023].

[10] P. Ennen, P. Hsu, C. Hsu, C. Liu, Y. Wu, Y. Liao, C. Lin, D. Shiu, W. Ma, “Extending the Pre-Training of BLOOM for Improved Support of Traditional Chinese: Models, Methods, and Results,” 2023,

[11] https://arxiv.org/pdf/2303.04715.pdf, [Retrieved: May 2023].

[12] J. Gao, T. Zhang, C. Xu, “A Unified Personalized Video Recommendation via Dynamic Recurrent Neural Networks,” 2017, [https://dl.acm.org/doi/pdf/10.1145/3123266.3123433], [Retrieved: May 2023].

[13] I. Ghosh, Ranked: “The 100 Most Spoken Languages around the World,” 2020, https://www.visualcapitalist.com/ 100-most-spoken-languages/, [Retrieved: May 2023].

[14] K. Hao, “A new immersive classroom uses AI and VR to teach Mandarin Chinese,” 2019, https://www.technologyreview.com/2019/07/16/65550/ai-vr-education-immersive-classroom-chinese-ibm/amp/, [Retrieved: May 2023].

[15] R. Ishida, “Character encoding for beginners,” 2015, (https://www.w3.org/International/questions/qa-what-is-encoding#:~:text=UTF%2D8%20is%20the%20most,ways%20of%20encoding%20Unicode%20characters), [Retrieved: May 2023].

[16] L. Jiang, HSK Standard Course 1 SET - Textbook + Workbook (Chinese and English Edition) Beijing Language & Culture University Press, 2018.

[17] Y. Liu, T. Yao, N.-P. Bi, L. Ge, Y. Shi, Integrated Chinese 4th Edition, Volume 1 Textbook (Simplified Chinese) (English and Chinese Edition) 4th Edition, Cheng & Tsui, 2016

[18] J. Moore, “What Exactly Is Mixed Reality?”, https://www.bairesdev.com/blog/what-exactly-is-mixed-reality/, [Retrieved: May 2023].

[19] Z. Ouyang, W. C. Fong, “Chinese Calligraphy”, Yale University Press, New Haven, CT, USA; Foreign Languages Press, Beijing, China, 2008.

[20] R. Sears, “Chinese Etymology” (https://hanziyuan.net/). [Retrieved: May 2023].

[21] T. Shih, “Distributed Multimedia Databases,” Idea Group Publishing: Hershey, PA, USA 2002.

[22] T. Shih, W. Gunarathne, A. Orchirbat, H. Su, “Grouping Peers Based on Complementary Degree and Social Relationship using Genetic Algorithm,” 2018, https://dl.acm.org/doi/pdf/10.1145/3193180 Retrieved: May 2023]

[23] “Unicode Security Considerations,” 2014, https://unicode.org/reports/tr36/, [Retrieved: May 2023].

[24] Wikipedia, CJK Unified Ideographshttps://en.wikipedia.org/wiki/CJK_Unified_Ideographs, [Retrieved: May 2023].

[25] S. Xu, Shuowen Jiezi (说文解字 shuō wén jiě zì, Analyze scripts and explain characters, by 许慎, also known as Hsu Shen, flourished in the reign of Andi of Han 汉安帝, A.D. 107-125.). [Explanations used in this chapter were quoted from the Dictionary page on ctext.org. Retrieved: May 2023]

[26] S. Yin, J. Xu, L. Ge, “Explore on Online English Listening Evaluation System Using the Genetic Algorithm,” 2022, https://www.hindawi.com/journals/misy/2022/2497365/, [Retrieved: May 2023].

[27] M. Q. Zhao, “A First Course in Database Systems Using SQL Server,” 2018.

[28] A. Zvieli, Entity -- Relationship modeling and fuzzy databases, International Conference on Data Engineering, February 5-7, 1986.