Grapho-Syllabic Systematicity in Chinese: Chinese Pictographs Have a Non-Arbitrary relation with their Pronunciations

Fiona Du (dushuhanpku@126.com)
School of Philosophy, Psychology and Language Sciences,
The University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

Hana Jee (h.jee@yorksj.ac.uk)
School of Education, Language and Psychology,
York St John University, York YO31 7EX, United Kingdom

Monica Tamariz (M.Tamariz@hw.ac.uk)
Psychology, Hariot-Watt University,
Edinburgh EH14 4AS, United Kingdom

Richard Shillcock (rcs@inf.ed.ac.uk)
School of Philosophy, Psychology and Language Sciences,
The University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

Abstract

It was recently found that letter-shapes have a non-arbitrary relation with their canonical pronunciations, in multiple orthographies and quantified across the whole of each orthography: letters that look similar tend to have similar pronunciations. Similarly, there is phonosemantic systematicity at the word level: words that sound similar tend to have similar meanings. We investigated for the first time whether a similar systematicity exists in Chinese characters. We measured all the pairwise phonological distances and all the pairwise orthographic distances of the 58 Chinese pictographic characters that are taught to year 1 and 2 in Chinese primary schools. The correlation was tested between the two lists of distances and verified by a Mantel test. We found a significant *negative* correlation between characters and their segmental pronunciations: characters that look similar tend to have dissimilar segmental pronunciations. This contrasts with the positive correlations found in previous similar research with alphabetic writing. We conclude, first, that questions of systematicity in the Chinese writing system are tractable in the same terms and by the same methodology as that applied to alphabetic writing systems. Second, segment-based processing requires to be augmented by tones for there to be systematicity that is comparable to that found in alphabetic writing systems. Any non-arbitrary relation between letter shapes and sounds may help bootstrap the acquisition of literacy.

Keywords: Chinese pictographs, orthography, phonology, grapho-phonemic systematicity

Background

A recent study found that letter shapes are non-arbitrarily related to their canonical pronunciations across the whole of the relevant alphabet (Jee, Tamariz, & Shillcock 2020; Jee, Tamariz, & Shillcock, 2021). Multiple writing systems, regardless of whether they were naturally developed (e.g. English) or consciously designed (e.g. Korean), showed significant *grapho-phonemic systematicity*: similar letters tend to have similar pronunciations.

Chinese characters are known to have three separate dimensions: *Wenzi* (characters), *Yinyun* (historical phonology) and *Xungu* (semantics). These three aspects are related to each other in a way that from every aspect one can infer some information concerning the other two (Wang, 1980). Recent studies have begun to discover that Chinese written forms can trigger a sense of how the character sounds (Chan, Minett, & Li, 2016; Reich, Chou, & Patterson, 2003).

As in other research, Chinese characters may be uniquely valuable in informing us about the envelope of possible relationships between the spoken and written forms of human language. Any potential systematicity in the relationship between characters and their sounds will help in processing the language. Investigating such a systematicity may answer questions regarding understanding the 'logosyllabic system' (Shen & Guo, 2014), acquiring linguistic signs, and approaching lexical meanings.

This study examines the relationship between the phonological system and the orthographic system in Chinese pictographs. Pictographs are not only culturally important but they also have experienced a long process of orthographical evolution as the earliest type of Chinese characters (Qiu, 1988). In all the pictographs we selected for this study, there is no difference between their traditional forms and their simplified forms. We chose to study characters that are acquired early by children. We assumed that if cultural evolution has instilled any systematicity it would be most likely to exist in culturally central, basic parts of the orthography where it can help to bootstrap learning (cf. Monaghan, et al., 2014). Note that our selected subset of pictographic characters can serve as the component radicals of a larger set of semantic-phonetic compound characters. The typical semantic-phonetic compound character is made up of two radicals, with a semantic radical on the left and a phonetic radical on the right. The pictographic characters we studied and any attendant systematicity are therefore of considerable applied interest in Education.

Method

Applying the experimental paradigm developed to calculate meaning-form systematicity (Jee, Tamariz, & Shillcock, 2022; Monaghan, et al., 2014; Tamariz, 2008), we measured all the pairwise distances between phonological forms of syllables and the corresponding pairwise visual distances between the pictographic characters. The correlation between these two lists of distances is defined as the grapho-syllabic systematicity. Arbitrariness will be represented as a non-significant correlation coefficient. The whole process was conducted in Python 3.7.

Sample

We collected 58 characters from Chinese textbooks designed for primary school students in Grade 1 (28 characters) and Grade 2 (30 characters). We measured the phonological and orthographic distances within the 28 Year 1 characters and within the 58 Year 1+Year 2 characters).

Measuring Phonological and Orthographical Distances

Each phoneme of a syllable was encoded into a vector based on its articulatory features. We employed two separate sets of phonological features: *Modern Standard Mandarin* and *Middle Chinese*. Then the distances between the vectors were calculated as Euclidean distances (cf. Monaghan et al., 2010).

For contemporary Mandarin, we referred to Zhao and Li (2009a). Their phonology template includes six slots: CVVVCT (C: consonant; V: vowel; T: tone). We collected all the required phonological representations from the *Phonological Representation Database for Chinese Characters, PRDCC* (Li & MacWhinney, 2002; Zhao & Li, 2009b).

As a representative phonological system of Middle Chinese, we chose the 切韵 *Qieyun* system, approximately from the Northern and Southern dynasties (420~589 A.D) to the *Tang* dynasty (618~907 A.D; Wang, 1980), which was recorded in *Qieyun*, a Chinese rime dictionary. With reference to the binary-value representation in Zhao and Li (2009a), we used a different 6-slot template to represent Middle Chinese phonology: CDHVCT (C: consonant; D: *deng*; H: *hu*; V: vowel; T: tone). We omitted tones, in this initial study.

In line with Jee, et al. (2020; 2021), we measured orthographical distances using Hausdorff distance (Huttenlocher, Klanderman & Rucklidge, 1993). It determines the directional resemblance between two images in terms of superimposition. Given that different fonts might lead to different results, we examined 12 representative fonts (Fig 1). They are historically classified in Table 1.

Figure 1: Examples of the characters in 12 fonts

Table 1: The Chinese fonts.

Type of script	Example	Font name
Seal script	好久不见	Xiaozhuan
Clerical script	好久不见	Lishu
	好久不见	HuawenLishu
Regular script	好久不见	Kaiti
	好久不见	HuawenKaiti
	好久不见	HuawenXingkai
	好久不见	HuawenXinwei
Modern font	好久不见	MicrosoftYahei
	好久不见	FangzhengYaoti
	好久不见	Songti
	好久不见	HuawenZhongsong
	好久不见	Fangsong

Results

We applied *Spearman's rho* (r_s) due to the possibility of non-normal distributions in the data. We verified the significance tests by conducting Mantel tests.

Table 2 and 3 shows the grapho-phonemic systematicity of Chinese pictographs when the phonological features were based on Modern Standard Mandarin. When Year 1 and Year 2 are combined (58 characters), we found an overall negative correlation between the phonological distances and the orthographical distances (Table 2). The characters with similar syllabic sounds tend to have distinct visual forms.

We observed the same pattern with just the Year 1 data (Table 3), although the coefficients and significance levels were slightly reduced. There was a decline in the absolute correlations from Year 1 to Year (1+2), which means that the negative systematicity has declined after entering Year 2.

Table 2: Grapho-phonemic systematicity (Year 1+2).

Font	Example	r_s	p	
Xingkai	好久不见	-0.12	< .001	***
Xinwei	好久不见	-0.10	< .001	***
HWkaiti	好久不见	-0.10	< .001	***
HWlishu	好久不见	-0.07	0.003	**
Lishu	好久不见	-0.07	0.004	**
MS Yahei	好久不见	-0.07	0.005	**
Kaiti	好久不见	-0.05	0.04	*
Songti	好久不见	-0.03	0.22	
Zhongsong	好久不见	-0.03	0.24	
Yaoti	好久不见	-0.01	0.76	
Fangsong	好久不见	0.02	0.44	
Xiaozhuan	好久不见	0.03	0.17	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The results illustrated in tables are sorted by r_s values. The modern fonts are shaded.

Table 3: Grapho-phonemic systematicity (Year 1 only).

Font	Example	r_s	p	
Lishu	好久不见	-0.17	0.001	***
HWkaiti	好久不见	-0.14	0.008	**
Xingkai	好久不见	-0.13	0.013	*
HWlishu	好久不见	-0.11	0.04	*
Xinwei	好久不见	-0.10	0.06	.
MS Yahei	好久不见	-0.06	0.21	
Kaiti	好久不见	-0.02	0.71	
Songti	好久不见	-0.001	0.99	
Zhongsong	好久不见	0.003	0.96	
Fangsong	好久不见	0.05	0.39	
Yaoti	好久不见	0.07	0.2	
Xiaozhuan	好久不见	0.15	0.004	**

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The results illustrated in tables are sorted by r_s values. The modern fonts are shaded.

Mantel tests confirmed the correlations (Fig 2).

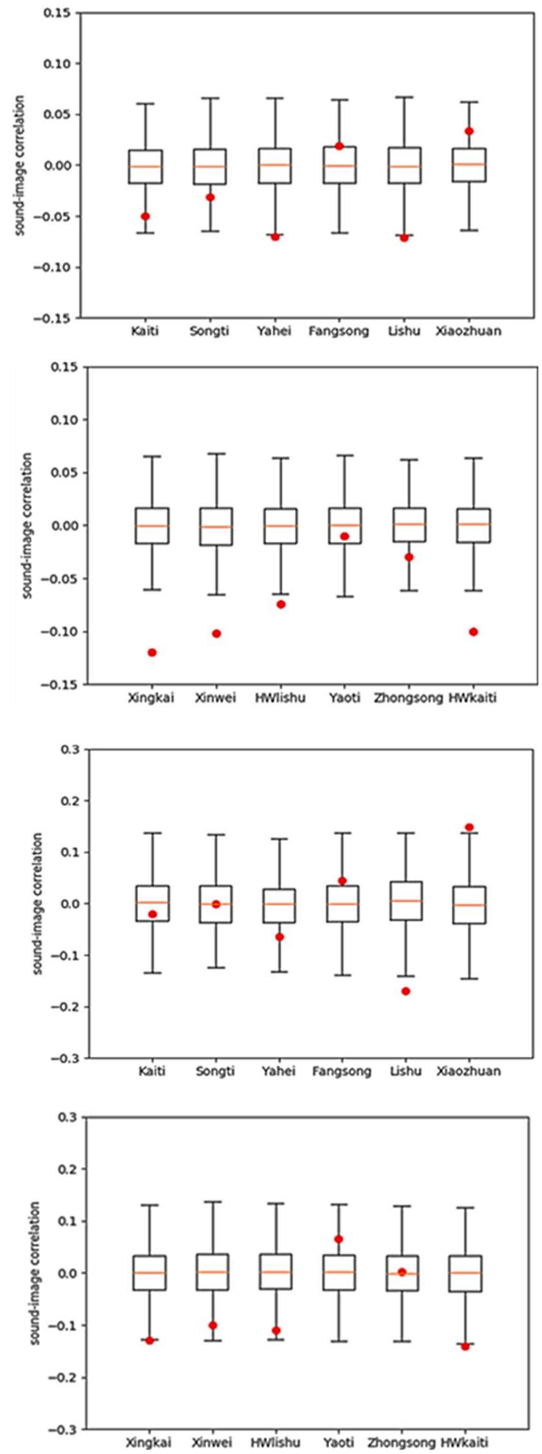

Figure 2: Mantel tests results. The veridical correlation coefficients are represented as red dots in the distributions of correlation coefficients obtained by chance. Year 1 and 2 (upper two), Year 1 only (below two).

Table 4 shows the results when the phonological features followed Middle Chinese phonological representations. Several fonts returned significant and marginally significant negative correlations between spoken and written forms.

Table 4: Grapho-phonemic systematicity of Middle Chinese.

Font	Example	r_s	p	
Zhongsong	好久不见	-0.145	0.005	**
Songti	好久不见	-0.114	0.027	**
Yahei	好久不见	-0.104	0.043	.
Yaoti	好久不见	-0.100	0.053	.
Xiaozhuan	好久不见	-0.089	0.083	.
HWlishu	好久不见	-0.088	0.087	.
Fangsong	好久不见	-0.085	0.098	
Lishu	好久不见	-0.051	0.321	
Kaiti	好久不见	-0.042	0.411	
Xingkai	好久不见	-0.041	0.424	
Xinwei	好久不见	0.023	0.659	
HWkaiti	好久不见	-0.004	0.942	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The results illustrated in tables are sorted by r_s values. The modern fonts are shaded.

Discussion

The written forms of Chinese are cognitively intriguing. Not only are they one of very few logographic orthographies, they have also been through multiple—at times institutionalised—adjustments. Their phonological systematicity has been overlooked by researchers, mainly due to an excessive focus on the relation between meaning and visual form.

We present what we believe is the first report of the systematic negative correlation between a key, motivated sample of Chinese characters and their pronunciations. This resembles the analysis reported by Jee, et al. (2020; 2021) only with an opposite, negative direction. Negative systematicity is not arbitrariness. It shows a tendency to maximize the distinctiveness between characters. In other words, the early developers of Chinese orthography tended to choose distinctively different syllabic units to refer to similar sounds. In a few studies of *meaning-form* systematicity (Monaghan, et al., 2014; Jee, et al., 2022), some categories of the subdivided sample did in fact return the opposite systematicity to the whole systematicity due to the unique characteristic of the category.)

We chose to focus on the simple pictographic characters taught in the first two years of school. We hypothesised that any systematicity may be strongest in these characters, which educators had intuited should be acquired first. Pictographs make up less than 7% of the 2906 most commonly used characters involved in Outline of Chinese Words and Chinese Characters Level (Li, 2003). However, these simple pictographs turn out to exhibit a level of distinctiveness.

Children learn their first characters from scratch. Although they do not indicate shared semantic/phonetic information as compounds words do, pictographs exploit visual iconicity by symbolizing objects or relations (Xiao & Treiman, 2012). Indeed, Chinese grade 1 textbooks teach many pictographs together with the original objects they depict; systematicity from iconicity may suit children's cognition better than regularity based on semantic/phonetic categorization.

Such iconicity goes from a real-world object to a simplified stroke-based caricature of that object. Our findings are concerned with the next step—from the character to its designated syllable. We have discovered a small but significant level of grapho-syllabic systematicity in some fonts, as a result of our study. We suggest that children may benefit from this systematicity between written Chinese characters and their designated syllables. For instance, it is now possible to decide algorithmically the best order in which to expose grade 1 learners to pictograms, such that distinctiveness is maximised at every point.

The next level of structure beyond pictographs and their corresponding syllables is compound characters in which such pictographs feature as semantic-phonetic compounds. These semantic-phonetic compound characters account for the largest proportion of characters (Li, 2003; Hsiao & Shillcock, 2006). Some 90% of them have the phonetic radical on the right of a left-right structure. For example, the pictograph 马 (HORSE) is the radical of a series of semantic-phonetic compound words, such as 妈 mā (mǎ MOTHER) and 码 mǎ (CODE); they sound the same without tones. Overall, the phonetic radical's pronunciation can sound the same (apart from the tone) as the semantic compound, as in 妈 mā (mǎ MOTHER), or the phonetic radical can be a partial match—or even no match—with the pronunciation of the semantic compound. At this level we are seeing a grapho-semantic systematicity that is *compositional* and can be understood from its parts, as opposed to the *distributed* and *comprehensive* (i.e. calculated over the whole sample of characters) systematicity we have investigated here.

The non-modern fonts tend to have higher negative correlations than the modern fonts. Specifically, all the variants of regular script showed significant negative correlations in the year (1+2) data, which means they are more distinctive than modern fonts. Considering these fonts resemble brush strokes (Table 1) this finding resembles the conclusions of Jee, et al., (2021) where the highest grapho-phonemic systematicity was found in the 궁서 font (e.g. 다람쥐 헌 쳇바퀴에 타고파). Further research is required regarding the relation between the ink area of characters and the exact method of calculating the visual distance between characters.

Can we understand the systematicity we have discovered for our sample of pictograms in certain fonts by comparing Chinese with an alphabetic language like English? The processing of both languages can be thought of in terms of the constraint satisfaction triangle: semantics-phonology-graphics.

In English, the graphics-phonology relation is a positive correlation: there is positive grapho-phonemic systematicity, in that similar letters tend to have similar canonical pronunciations. In Chinese, we have reported a *negative* grapho-syllabic systematicity, in that similar pictograms tend to have *dissimilar* syllabic pronunciations.

In English, the grapho-semantic relation exists, although it is commonly overlooked in theorising. An example is the "psy" in "psychology", where the /s/ sound might otherwise be written as "s". This is the graphics relating directly to the semantics. In Chinese, the grapho-semantic relation is explicit in the pictograms.

In English, the phono-semantic relation exists as a positive correlation between phonological distances and semantic[1] distances between words. In Chinese, there are no studies, as far as we are aware, quantifying semantic distances between Chinese words and using the distances to estimate systematicity.

The process of constraint satisfaction—whereby consistent semantic, phonological and graphical dimensions align—is therefore different in English and Chinese. The three arms of the triangle have different strengths in each language.

Our study is an initial attempt to make the issue of systematicity tractable in Chinese. We conclude that systematicity can be detected in the limited sample of pictographs that we studied. The approach used in previous studies can be successfully transferred to Chinese.

We further conclude that syllabic phonological form expressed purely in terms of segmental phonology produces systematicity that is potentially beneficial to learners, but that it involves a negative correlation between syllabic phonological distances and graphical distinctions. We predict that bringing tone into the calculations will increase phonological distances and reduce the negative correlation between syllabic form distances and graphical distances. We hypothesise that this addition will bring the picture of Chinese processing more into line with the existing theorising that it is *cognitive effort* that mediates the observed systematicity in other languages and other writing systems (Jee, et al., 2022).

Conclusion

We report the grapho-syllabic systematicity for the first time in Chinese, confirming that systematicity is not unique to alphabetic orthographies (cf. Jee, et al., 2020; Jee et al., 2021). It appears to be a universal feature across many orthographies of different types, but the surface nature of the systematicity may vary between languages, orthographies, fonts, and sub-samples of the lexicon.

[1] Semantic distances may be measured by Latent Semantic Analysis or by inter-word distances in Wordnet.

References

Chan, S., Minett, J., & Li Wing Yee, F. (2016). *The Routledge encyclopedia of the Chinese language.* London; New York: Routledge, Taylor & Francis Group.

Hsiao, J. H. W., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research*, *35*(5), 405-426.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9), 850-863.

Jee, H., Tamariz, M., & Shillcock, R. (2020). Quantifying sound-graphic systematicity and application on multiple phonographs. In *CogSci*.

Jee, H, Tamariz, M. & Shillcock, R. Quantifying sound-graphic systematicity; Application to multiple phonographic orthographies. Haralambous, Y. (2021, February). *Proceedings of Grapholinguistics in the 21st Century*, June 17-19, 2020. In G21C 2020: Grapholinguistics in the 21st Century (Vol. 4). Fluxus

Jee, H., Tamariz, M., & Shillcock, R. (2022). Exploring meaning-sound systematicity in Korean. *Journal of East Asian Linguistics*, *31*(1), 45-71.

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers*, 34, 408-415.

Li, D. (2003). *Concise and Practical Study of Chinese Characters*. Beijing: Peking University Press.

Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2010). Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon*, *5*(3), 281-299.

Monaghan, P., Shillcock, R., Christiansen, H. M., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1651), 20130299.

Reich, S., Chou, T. L., & Patterson, K. (2003). Acquired dysgraphia in Chinese: Further evidence on the links between phonology and orthography. *Aphasiology*, *17*(6-7), 585-604.

Shen, Y. & Guo, R. (2014). *Modern Chinese*. Beijing: Higher Education Press.

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2), 259-278.

Wang, li. (1980) *Hanyu Shigao* [Historical Sketch of the Chinese Language], Beijing: Zhonghua Shuju.

Xiao, W., & Treiman, R. (2012). Iconicity of simple Chinese characters. *Behavior research methods*, *44*(4), 954-960.

Zhao, X., & Li, P. (2009a). *Phonological Representation Database for Chinese Characters* [database]. Retrieved from http://www.personal.psu.edu/pul8/chi_rep/chinese_patphon.html

Zhao, X., & Li, P. (2009b). An online database of phonological representations for Mandarin Chinese. In *Behavior Research Methods*, 41(2), 575-583.

Filename:	chinese_pictographs_final
Directory:	C:\Users\hanaj\OneDrive\Documents
Template:	C:\Users\hanaj\OneDrive - York St John University\Research\2. word_cogsci_fullpaper\cogsci_full_paper_template_word.dotx
Title:	Pre Test Excerpt
Subject:	
Author:	Hana Jee
Keywords:	
Comments:	
Creation Date:	5/9/2022 12:45:00 PM
Change Number:	44
Last Saved On:	5/9/2022 5:18:00 PM
Last Saved By:	Hana Jee
Total Editing Time:	162 Minutes
Last Printed On:	5/9/2022 5:19:00 PM
As of Last Complete Printing	
Number of Pages:	6
Number of Words:	3,111 (approx.)
Number of Characters:	17,735 (approx.)