LETTERS AND THEIR SOUNDS ARE NOT PERFECTLY ARBITRARY: EXPLORING GRAPHO-PHONEMIC SYSTEMATICITY IN MULTIPLE ORTHOGRAPHY SYSTEMS

HANA JEE [1], FIONA DU [1], MONICA TAMARIZ [2], and RICHARD SHILLCOCK [1]

[*]Corresponding Author: h.jee@yorksj.ac.uk
[1]Psychology, The University of Edinburgh, Edinburgh, UK
[2]Psychology, Heriot-Watt University, Edinburgh, UK

Language, as a complex system, suggests coordination between subsystems. Recent studies demonstrated that semantically similar words tend to have similar pronunciation (Blasi et al., 2016; Dautrich et al., 2017; Jee, Tamariz, & Shillcock, 2022; Monaghan et al., 2014; Tamariz, 2008). The current research, for the first time, quantified mapping between letters and their canonical pronunciations, or *grapho-phonemic systematicity*.

We examined naturally developed phonograms (Arabic, English, Greek, and Hebrew), consciously designed phonograms (Korean, Shavian alphabet, and Pitman's shorthand), a logographic orthography (Chinese) and fictitious orthography systems (Aurebesh and Klingon).

We measured all the pairwise phonological distances between phonemes in the respective alphabet system, and the corresponding pairwise orthographical distances between letters. We then tested Pearson's *r* between these two lists of pairwise distances. The positive correlation coefficient means that similar letter-shapes have similar canonical pronunciation. In contrast, the negative correlation means that similar letter-shapes have more distinct sounds, or vice versa. We verified the significance of the correlations by conducting Monte-Carlo permutation tests.

For the phonological distance, phonemes were encoded into vectors according to the articulatory features and the distance between the vectors were calculated in various ways. We applied three methods to measure the pairwise distances between letter-shapes. *Pixel count* simply defines the distances as the difference in the number of pixels between two characters. *Perimetric complexity* is defined

as ink area divided by perimeter of the character, thus the distance means the difference in complexity. *Hausdorff distance* (Huttenlocher et al., 1993) quantifies the difference between two images. Since each letter was saved as an image file (PNG), we were able to compare the contribution of the font to the grapho-phonemic systematicity.

We found the significant grapho-phonemic systematicity for all conventional writing systems and two English shorthand systems. Those fictitious alphabets did not show any systematicity. Considering each orthographic distance measure focuses on distinct aspect of the letter-shapes, the fact that a certain method maximised the systematicity of the writing system implies how it evolved.

Semitic orthography systems (Arabic, English, Greek and Hebrew) showed highest grapho-phonemic systematicity when measured by pixel count (e.g. English upper-cases $r = .22$, $p < .001$; English lower-cases $r = .14$, $p = .02$), which indicates that more articulatorily complicated phonemes take up more space in written forms. Effort in writing is easily understood as a letter's elaborateness—how long it takes to reproduce a character. Elaborateness is typically proportional to the number of pixels. Korean, Shavian alphabet and Pitman's shorthand were all intentionally designed to exploit the systematicity between letters and sounds. For instance, voiced-voiceless phoneme pairs share the identical visual features with slight variations. This topological difference was well-captured by Hausdorff distance (e.g. Korean KCC Eun-young $r = .39$, $p <. 001$).

Although limited in number (N = 58), we found the significant grapho-phonemic systematicity in the Chinese characters that are acquired in the first and second year of the primary school. We found the negative correlation coefficient ($r = -.12$, $p < .001$), indicating that Chinese was influenced by an evolutional force that *distinguishes* linguistic symbols. The finding implies that grapho-phonemic systematicity may exist to facilitate language learning and orthography acquisition.

Our analyses are first a proof of concept: it is possible to quantify grapho-phonemic systematicity across a whole alphabet, for particular fonts and for different languages. We also have confirmed and quantified the systematicity intended by the authors of Korean writing system and English shorthand systems. Our future research can shed more lights on sub-structure of grapho-phonemic systematicity: the contribution of each phoneme/letter to the whole systematicity; whether the more frequent phoneme/letter contributes more to the whole systematicity; and most importantly, how this grapho-phonemic systematicity bootstraps infants' learning orthography.

References

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818-10823.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, *41*(8), 2149-2169.

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. (2022). Exploring meaning-sound systematicity in Korean. *Journal of East Asian Linguistics*, *31*(1),

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

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