

Learning and Generalisation of Chinese Character Knowledge



ROYAL HOLLOWAY UNIVERSITY OF LONDON

Cheng-Yu Hsieh¹, Marco Marelli², Kathleen Rastle¹

¹ Department of Psychology, Royal Holloway, University of London, Egham, UK

² Department of Psychology, University of Milano–Bicocca, Milan, Italy



Introduction

- Chinese characters comprise a unit of meaning (*morpheme*) typically embedded in two-character compound words.
- Characters often convey multiple meanings (*morphemic ambiguity*).



花錢 spend-money 花心 fickle-heart 花生 flower-born



- Readers are sensitive to the statistical regularities of form-meaning relationships present in the writing system (Marelli et al., 2015).
- Morpheme knowledge may emerge from experience with whole words (Marelli et al., 2017; Tamminen et al., 2015)
- This study aims to investigate how Chinese readers learn and generalise character knowledge from experience with compound words.

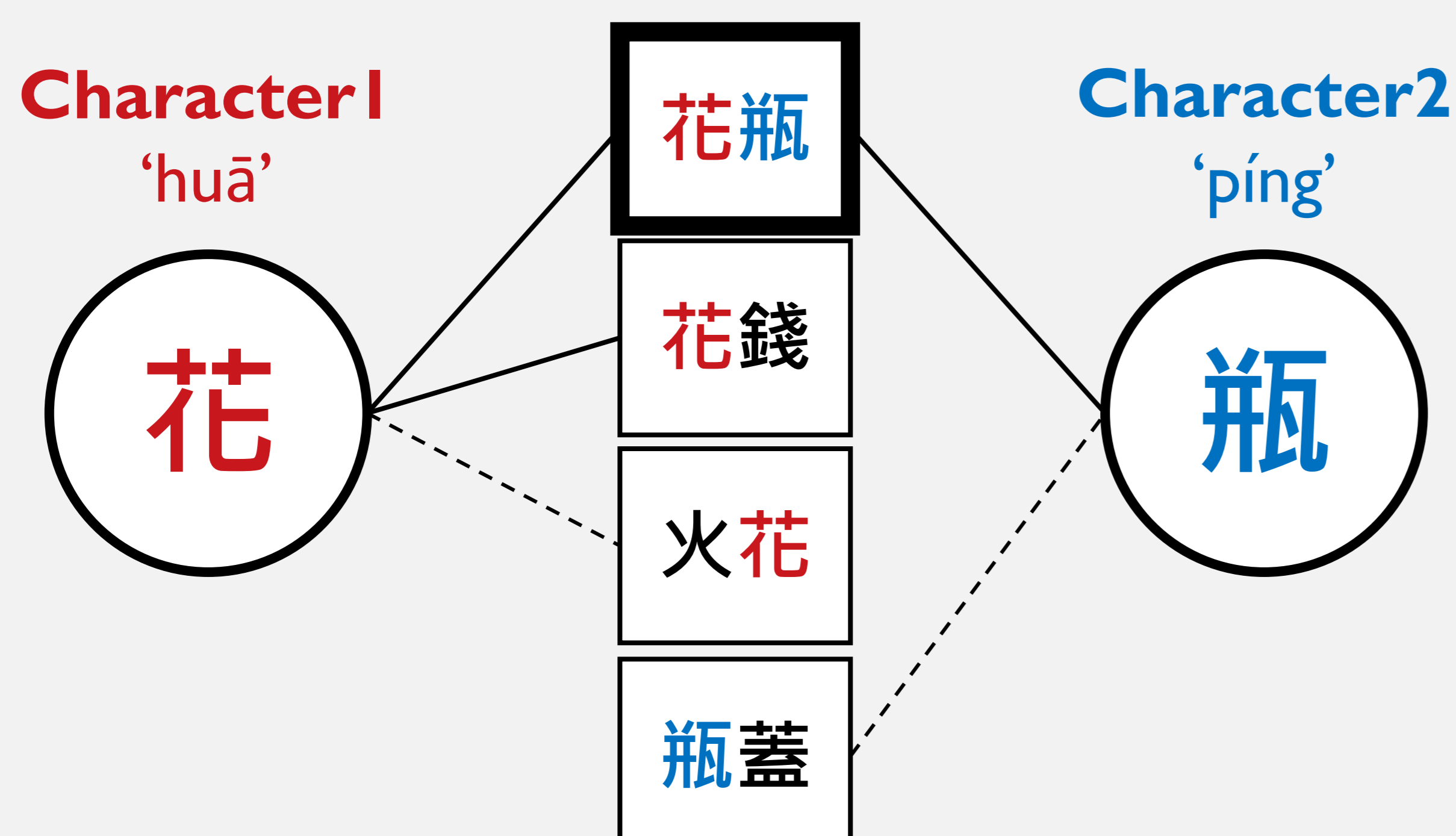
Methods

Lexical Decision Dataset on Two-Character Words (Tse et al., 2017)

- 33 adults from Hong Kong responded to 22,000 words & nonwords.

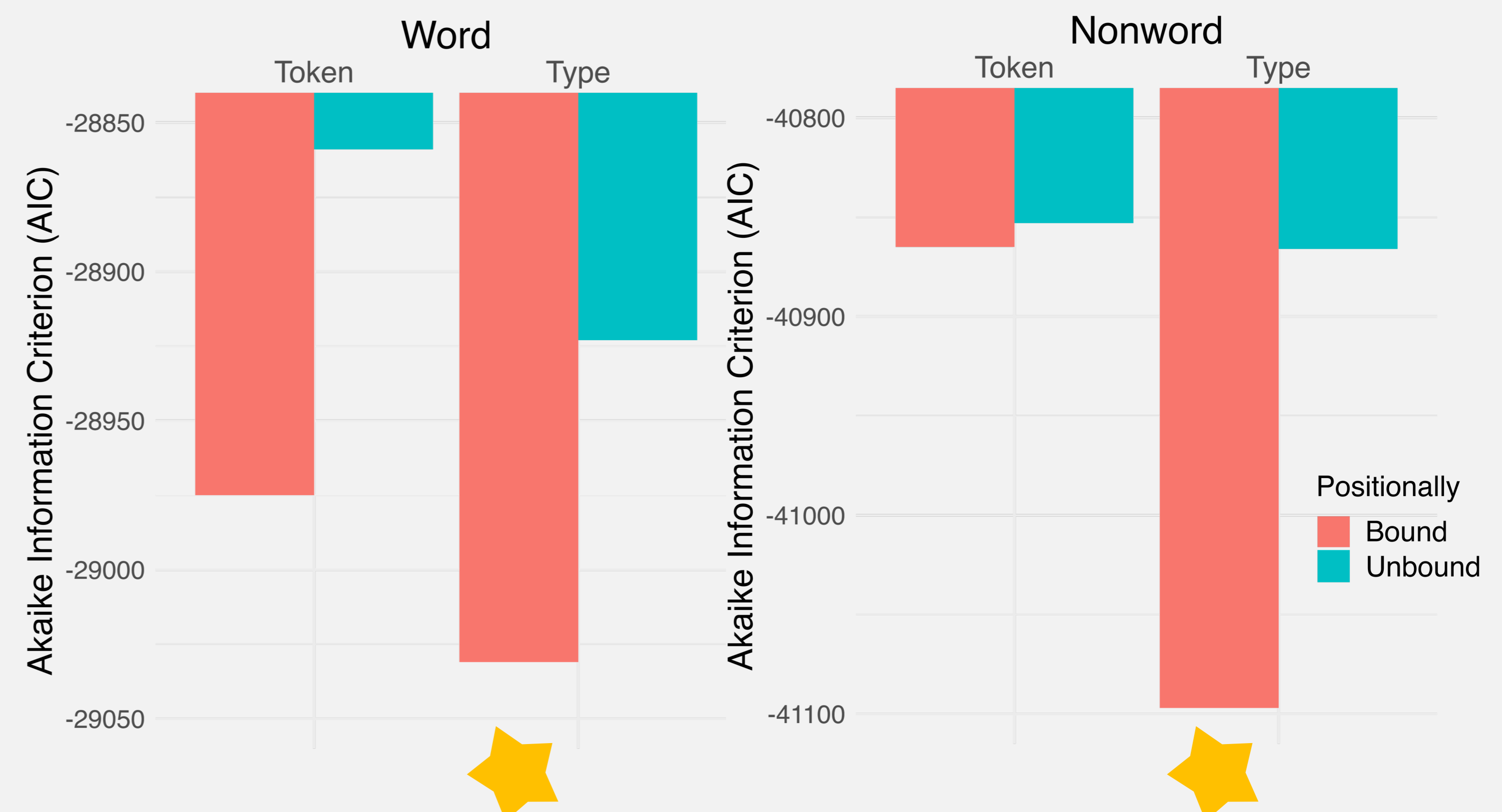
Family-related Metrics Under Different Definitions

- Family size:** the number of *family members* (words with the same character)
- Family semantic consistency:** average *semantic relatedness* (cosine similarity between word vectors derived from a Chinese word2vec model) between the meanings of the character and family members
- Both metrics can be defined along two dimensions:
 - by type frequency *vs.* by token frequency
 - positionally bound *vs.* unbound: whether the family members are constrained by the position (as demonstrated by the dot lines below).

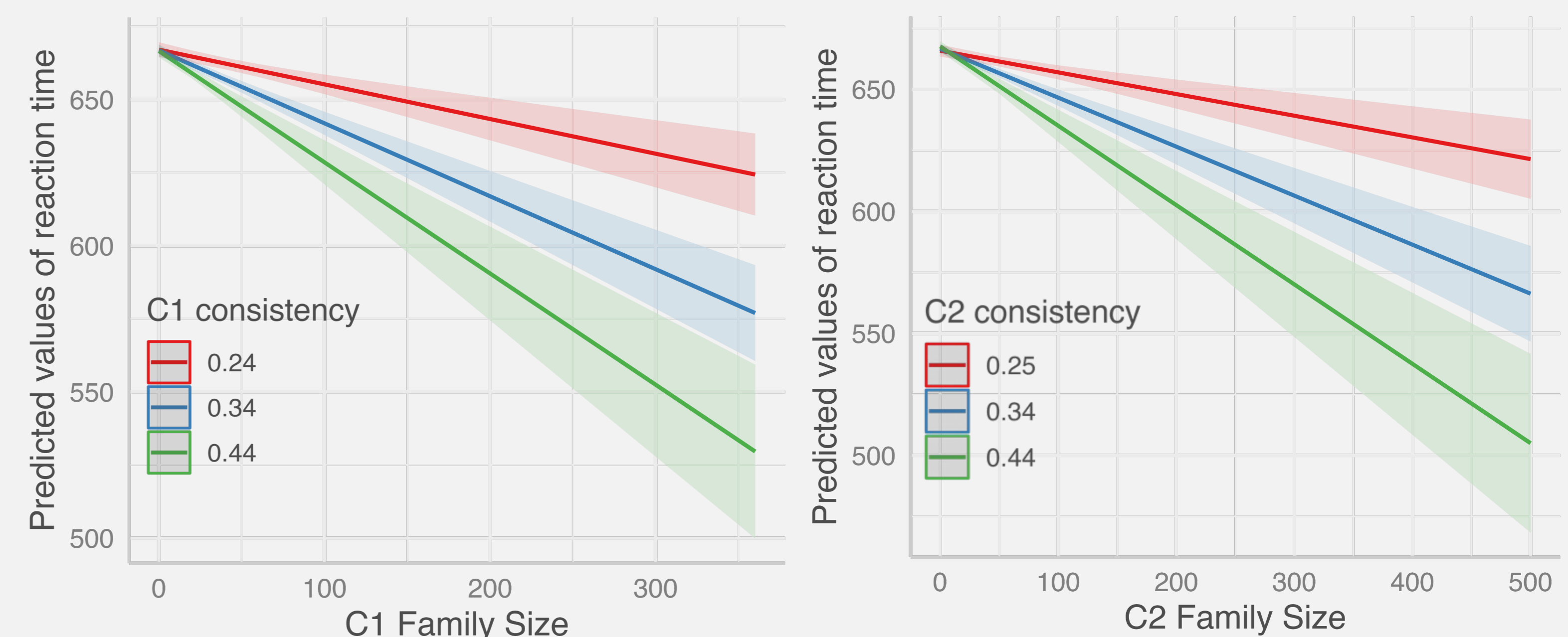


Results

Metric Selection: Type-based, positionally bound is the best

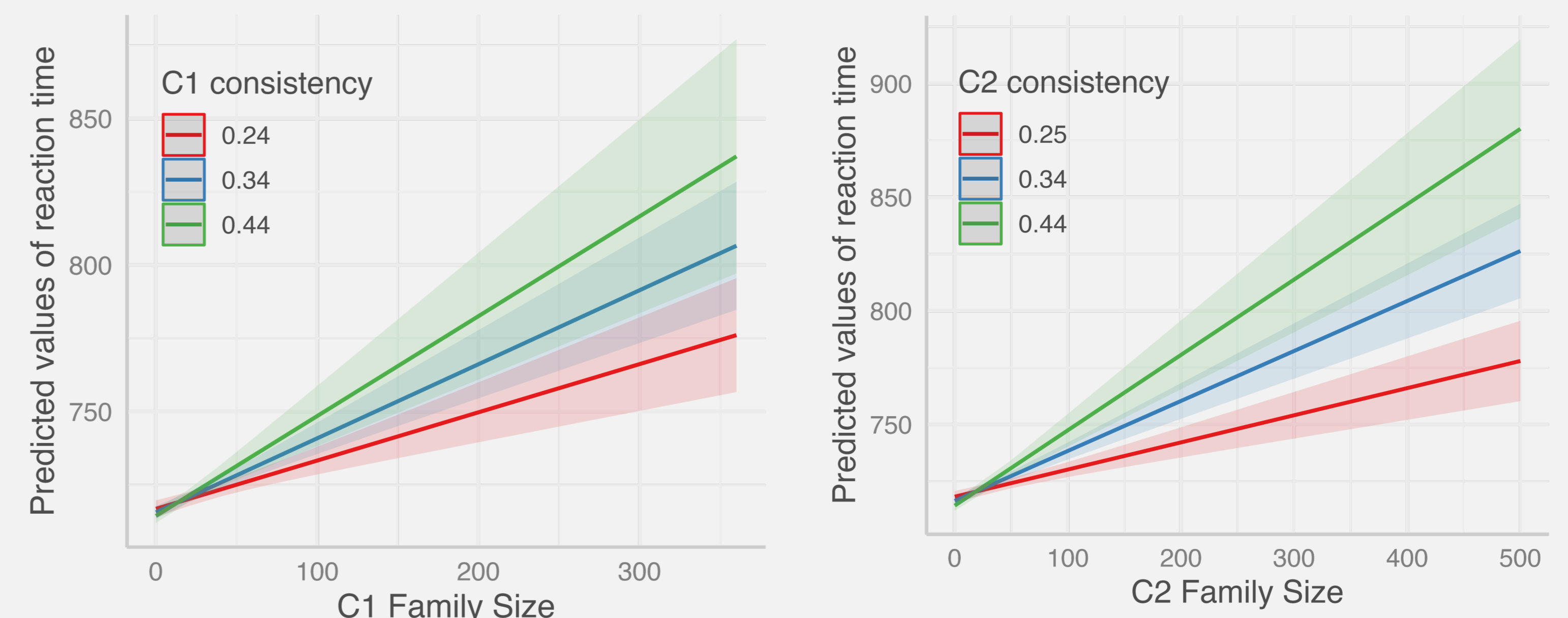


Model for Word Processing



- Words with more C1 and C2 family members are recognised faster.
- Words with greater semantic consistency within C1 and C2 families are recognised faster.

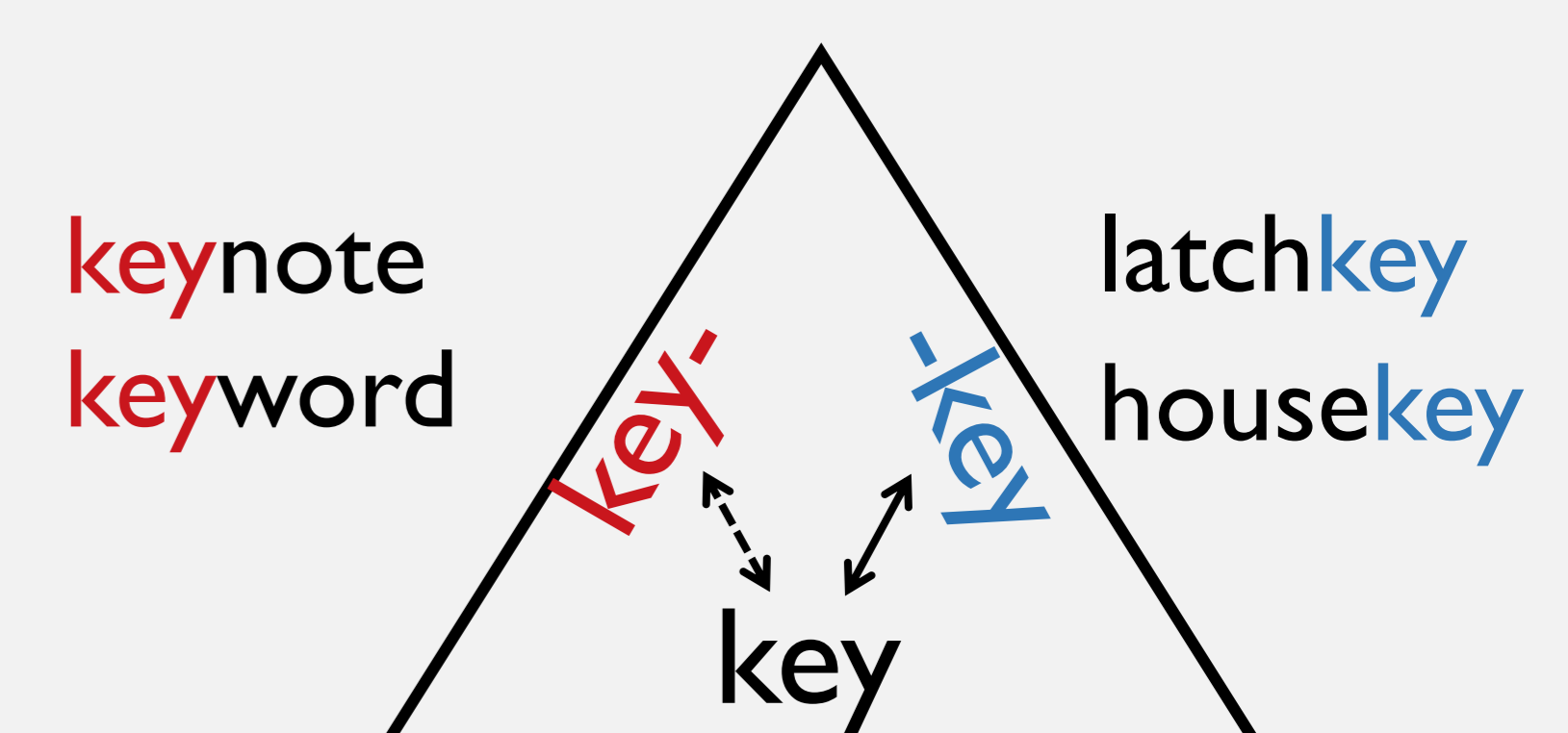
Model for Nonword Processing



- Nonwords with more C1 and C2 family members are harder to reject.
- Nonwords with greater semantic consistency within C1 family are harder to reject.

Conclusions

- The consistency with which individual Chinese characters contribute to the meanings of compound words influences processing of existing and novel compound words.
- Chinese readers pick up the statistical knowledge of characters through experience with compound words.
- The fact that the character occurs in a number of different words is important to character knowledge acquisition.
- Comprehension of compounds is supported by the knowledge of positionally bound constituents, rather than the meanings of the constituents as free words.



Reference

- Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, 68(8), 1571-1583.
- Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166, 207-224.
- Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, 79, 1-39.
- Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, 49(4), 1503-1519.