Word size in spoken and written Mandarin Chinese

James Myers
(National Chung Cheng University)

Although Chinese characters represent monosyllabic morphemes and word boundaries are not marked orthographically, Chinese words actually tend to be disyllabic. Since this is the size of a metrical foot, a prosodic explanation seems plausible (Duanmu 2007). The prosody hypothesis thus predicts a modality effect (speech vs. writing) on word size, both for word coinage and for word selection. These predictions were tested in the spoken (half million word tokens) and written (ten million word tokens) subcorpora of the Academia Sinica Balanced Corpus (Chen et al. 1996).

Disyllabic (two-character) words predominate both in speech (13,669 types) and in writing (97,899 types). However, as shown in Figures 1 and 2, modality affects potential word coinage, as extrapolated by Generalized Inverse Gauss-Poisson (GIGP) Large-Number-of-Rare-Events (LNRE) modeling (Evert and Baroni 2007): while in speech the predicted increase in types as a function of larger token samples is steepest for disyllables, in writing it is steepest for three-character words.

As for word selection, thousands of Chinese lemmas (syntactic/semantic lexical entries) are "elastic" in size, with speaker/writers free to realize them as either mono- or disyllabic (e.g., *zhuō(zi)* 'table', *dì(di)* 'younger brother', *(dà)gē* '(big) elder brother', *dōng(fāng)* 'east(ern direction)', *(xī)guā* '(water)melon'). In a picture naming experiment, Perry and Zhuang (2005) found that the probability of choosing the disyllabic form of elastic lemmas increased when the picture set included objects with non-elastic disyllabic names. To see if prosodic priming also occurs naturally, we selected elastic nouns (with the help of Duanmu and Dong forthcoming) used consistently as nouns, in both forms, in the spoken (146 lemmas) and written (990 lemmas) subcorpora.

We then computed the log ratio of disyllabic (two-character) to monosyllabic (one-character) words within a ten-word window preceding the elastic word target, used effect coding to indicate the absence/presence of this same lemma (in either form) in this window, and crossed these predictors (rescaled to *z* scores) in a by-lemma mixed-effects logistic regression model predicting disyllabic form choice for the target elastic words. As shown in Figures 3 and 4, the modalities showed similar patterns: the greater the disyllabic ratio in the preceding words, the more likely speakers were to choose the disyllabic/two-character form (β_{speech} = 0.37, $\beta_{writing}$ = 0.39), and while there was an interaction with lemma repetition (β_{speech} = 0.22, $\beta_{writing}$ = 0.14), word length priming was also significant without repetition (β_{speech} = 0.30, $\beta_{writing}$ = 0.35) (all *p*s < .0001).

Since modality matters, and LNRE models take Zipf's law into effect, the preference for disyllabic words in speech seems genuinely prosodic; the preference for longer words in written Chinese may reflect the more complex concepts and polysyllabic borrowings in a more formal register. The similar degree of priming of disyllabic/two-character words across modalities suggests either that word selection (as opposed to word coinage) remains subject to prosodic priming even in writing, or that word lengths "clump" in natural language for non-prosodic reasons.

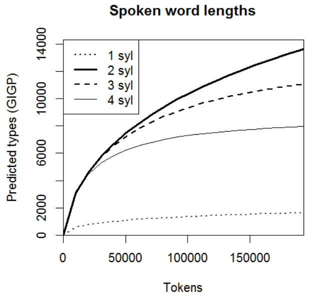

Figure 1

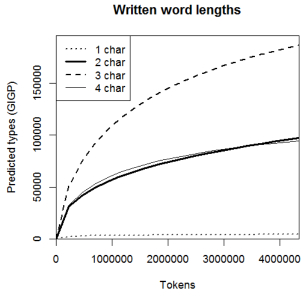

Figure 2

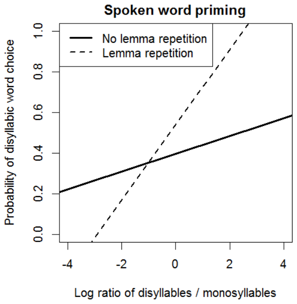

Figure 3

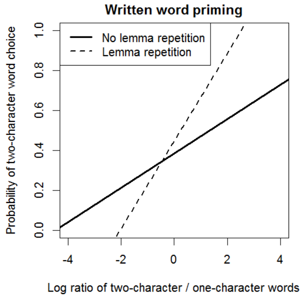

Figure 4

References

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu. 1996. Sinica Corpus: Design methodology for balanced corpora. *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation*, Seoul, Korea, pp. 167-176.

Duanmu, San. 2007. *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

Duanmu, San, and Yan Dong. Forthcoming. Elastic words in Chinese. In S.-W. Chan (ed.) *Routledge Encyclopedia of the Chinese Language*. London: Routledge.

Evert, Stefan, and Marco Baroni. 2007. zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29-32, Prague, Czech Republic.

Perry, Conrad and Jie Zhuang. 2005. Prosody and lemma selection. *Memory and Cognition* 33: 862-870.