Word size in spoken and written Mandarin Chinese

James Myers (National Chung Cheng University, Taiwan)

49th Annual Meeting of the Societas Linguistica Europaea, 31 August – 3 September 2016, Naples, Italy

Summary

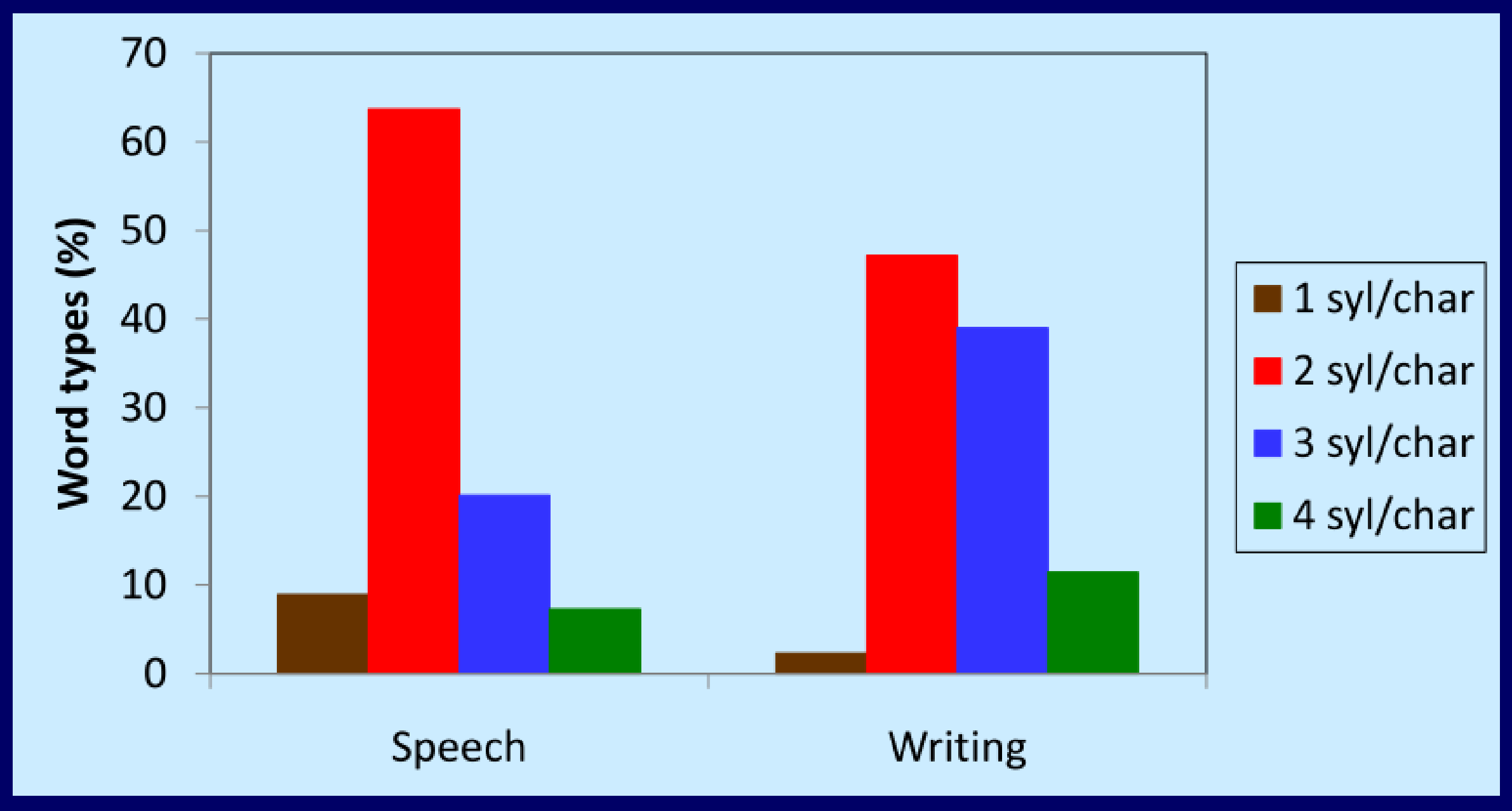
- Mandarin words tend to be disyllabic
- Do disyllables play a special role in spontaneous speech?
- To find out, we analyzed word sizes in spoken and written corpora
- **Productivity**
  - Disyllabic words most productive in speech
  - But trisyllables most productive in writing
- **Priming**
  - Disyllabic word size is chosen more in disyllabic contexts, in both modalities

Background

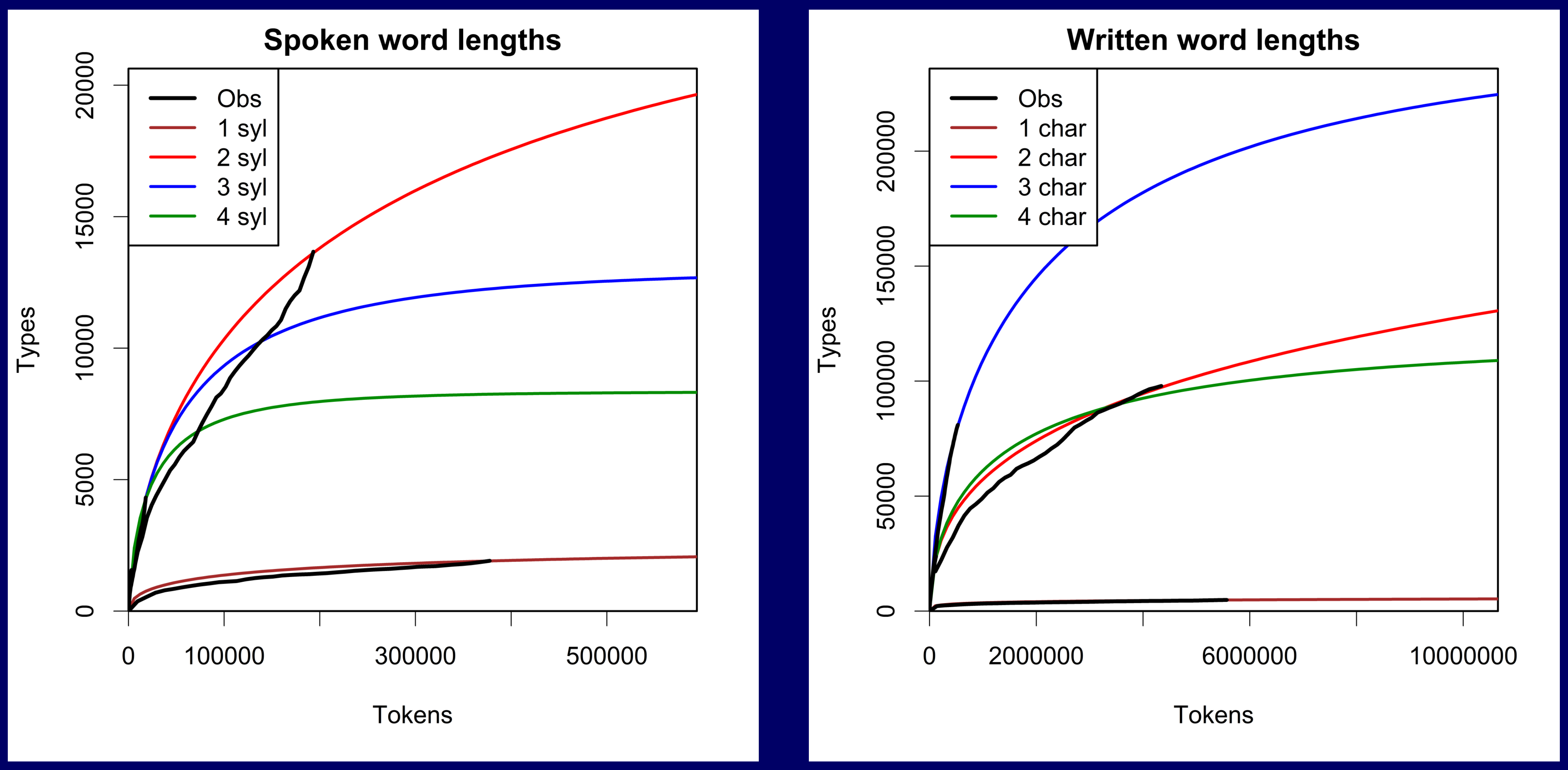
- Chinese characters represent **monosyllabic** morphemes, yet words in Sinitic languages, including Mandarin, tend to be **disyllabic**
- This is the size of a **metrical foot** in Chinese (Duanmu 2007, Myers & Tsay 2015)
  - Cf. also tone sandhi, poetry, stress, morphology
- Are disyllables special in **natural speech**?
  - Source: **Academia Sinica Balanced Corpus** (Chen et al. 1996)
  - Spoken portion: ca. 500,000 word tokens
  - Written portion: ca. 10,000,000 word tokens
- Are disyllables most productive? Is there prosodic priming? Does modality matter?

Productivity

- **Disyllabic/two-character words** predominate in both speech and writing



- Productivity was quantified as **growth curves** (Evert & Baroni 2007)
  - The number of distinct word types as a function of the number of sampled word tokens
- Word lengths compared by **extrapolating** to the same sample size via LNRE modeling
- The growth curves showed a dramatic **effect of modality** (see plots)
- **Speakers may plan word choice in terms of disyllabic feet**, but writers do not
  - Difference in productive word length across modalities may also relate to **information load**



Number of types per number of tokens, as observed and as extrapolated by LNRE models

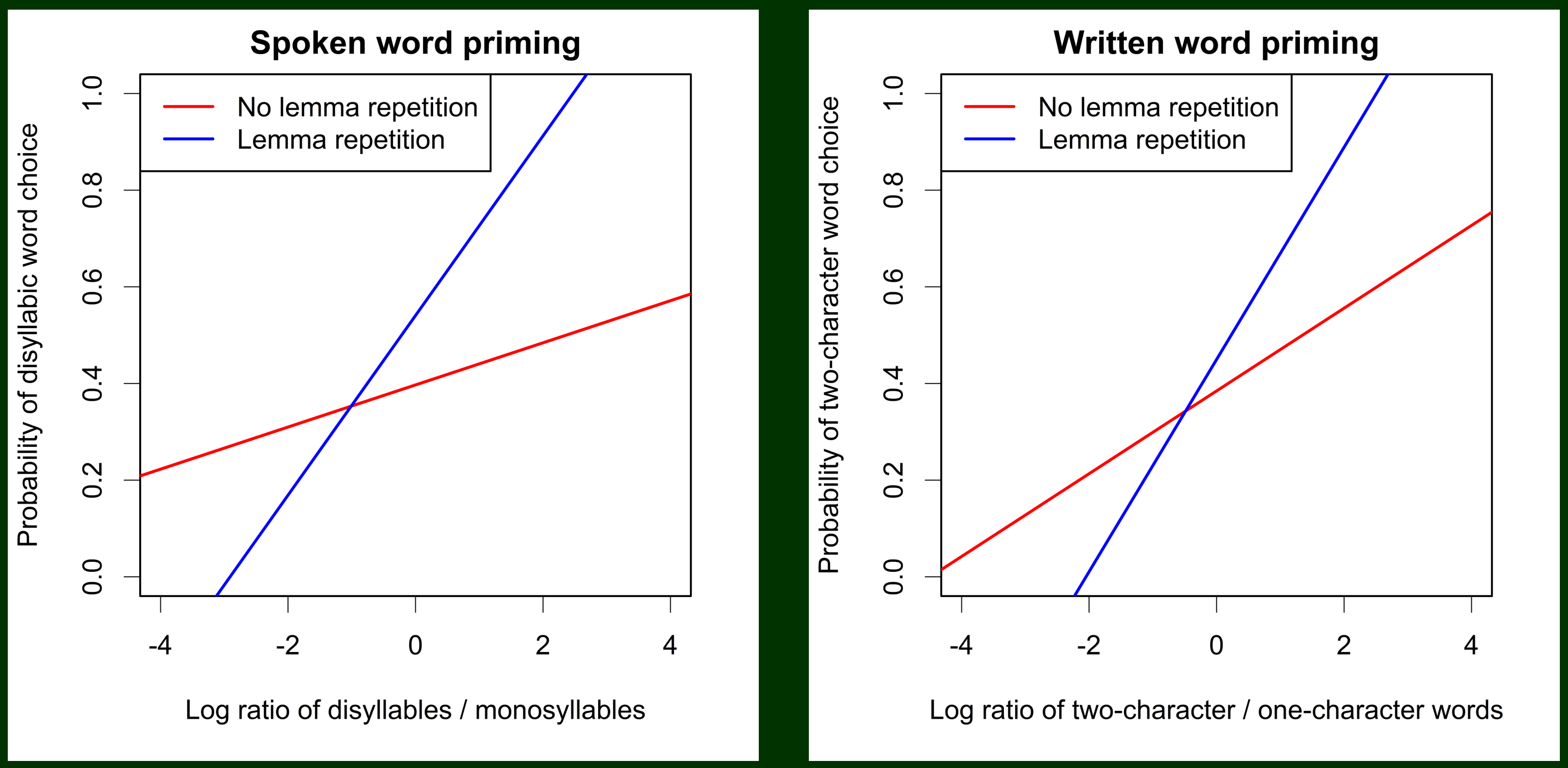
(Generalized Inverse Gauss-Poisson Large Numbers of Rare Events modeling with χ^2 goodness of fit)

Priming

- Thousands of Mandarin lemmas (syntactic/semantic lexical entries) are **elastic**
  - "Freely" vary in word length as monosyllabic or disyllabic (Duanmu & Dong 2016)

optional empty suffix	**zhuō** 'table'	**zhuōzi** 'table'
optional reduplication	**dì** 'younger brother'	**dìdì** 'younger brother'
redundant modifier	**gē** 'elder brother'	**dàgē** 'big elder brother'
redundant head	**dōng** 'east'	**dōngfāng** 'eastern direction'
superordinate category	**guā** 'melon'	**xīguā** 'watermelon'

- An experimental **priming** effect (Perry & Zhuang 2005)
  - In a picture-naming task, speakers were more likely to choose the **disyllabic variant** of elastic words when there also were pictures with **fixed disyllabic names in the test set**
- A corpus-based analysis
  - **Predict disyllabic variant** of elastic lemmas from log **ratio of disyllabic to monosyllabic** in adjacent ten words, with or without repetition of the target lemma
  - Elastic **word size was primed by context**, both preceding (see plots) or following (same pattern), even without lemma repetition, for **both speech and writing**
- Prosodic effects in writing? Or an indirect effect of shifting **degrees of formality**?



Effect of prosody and lemma repetition on probability of producing disyllabic elastic variant

(mixed-effects logistic regression on 146 spoken and 990 written elastic nouns)

What next?

- Productivity
  - Are affixed and compound words different?
- Priming
  - What about trisyllabic or longer words?
  - Is there cross-speaker priming?
- Other languages
  - Are there similar patterns in other languages with disyllabic feet?
  - What happens in languages with bimoraic or unbounded feet?
- Perception/recognition
  - Does only production show such patterns?
- **Wordlikeness**
  - Do Mandarin speakers judge disyllabic nonwords as particularly wordlike?
  - What about speakers of other languages?
  - Try our web app to collect and share wordlikeness judgments across languages (Chen & Myers 2016)

Worldlikeness

http://lngproc-4083.nitrouspro.com:3000/

Acknowledgments

MOST 103-2410-H-194-119-MY3 (Taiwan), Tsung-Ying Chen, Jane Tsay, three anonymous SLE reviewers

References

Chen, Huang, Chang, & 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.

Chen & Myers. 2016. Worldlikeness: A Web-based tool for typological psycholinguistic research. *Proceedings of the 40th Annual Penn Linguistics Conference*. University of Pennsylvania.

Duanmu. 2007. *The Phonology of Standard Chinese, 2nd ed*. Oxford University Press.

Duanmu & Dong. 2016. Elastic words in Chinese. In Sin-Wai Chan (ed.) *Routledge Encyclopedia of the Chinese Language*, pp. 452-468. London: Routledge.

Evert & 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.

Myers & Tsay. 2015. Trochaic feet in spontaneous spoken Southern Min. *Proceedings of the 27th North American Conference on Chinese Linguistics*, Vol. 2, 368-387. Los, Angeles: UCLA.

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