Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Reading Time Data Collection Using Crowdsourcing

National Institute for Japanese Language and Linguistics, Japan
Tokyo University of Foreign Studies
Masayuki Asahara

How do vocabulary rating scores affect reading time in the Japanese language?

Word Familiarity Survey
https://github.com/masayu-a/WLSP-familiarity

Word Familiarity rating using crowdsourcing
Rating five perspectives
KNOW, WRITE, READ, SPEAK, LISTEN

Stimuli: 84,114 surface forms from the Word List by Semantic Principles (WLSP)

Modeling using the Generalized Linear Mixed Model

Vocabulary Rating
Word Familiarity

Survey date Number of Ratings
2021/09/12-14 2,396 385,380
2020/10/09-12 2,372 943,295
2019/11/14-22 2,421 288,000
2018/11/15-21 3,391 1,617,215

Reading Time Collection
https://github.com/masayu-a/BCCWJ-SPR2

Self-paced reading using crowdsourcing
Stimuli:
Whitepaper, Textbooks, Books
in the Balanced Corpus of Contemporary Written Japanese (BCCWJ)

Modeling using:
Generalized Linear Mixed Model
Bayesian Linear Mixed Model

Reading Time = Vocabulary Rating

Statistical Analysis Generalized Linear Mixed Model: Logarithms of Reading Time

Statistical Analysis Generalized Linear Mixed Model: Reading Time

Future directions
Comparison with participants' vocabulary rating results
Working memory estimation = reading time