12

How many words do you know?

- Vocabulary size test, Reiwa edition -

Abstract

More than 20 years ago, NTT conducted psychological experiments to investigate word familiarity and thus construct a Japanese lexicon of about 77,000 words. This lexicon is still used in many fields.

Now, we reinvestigated and reconstructed an unparalleled-scale Lexicon of 163,000 words through crowdsourcing. By applying careful screening, we succeeded in obtaining highly reliable results. This makes it possible to make comparisons with results of 20 years ago. Based on this, we created the *Reiwa* edition of a vocabulary-size estimation test. Furthermore, this lexicon allows us to estimate the vocabulary size appropriate for the present day. We are working on vocabulary-size investigations for a wide range of ages, including elementary school to high school students. In the future, we will investigate and analyze the relationship between vocabulary size, reading comprehension, and academic ability, aiming to achieve effective educational support.

Word Familiarity Database, Reiwa edition Word Familiarity Examination via Cloud Ask good workers Comparison of Histograms with previous data to evaluate other sets Pre-screening Post-screening Cloud workers Rating word familiarity Cloud workers ■ 1995 (77,000 words) 2002 (added 32,000 words) Holder of official Kanji Evaluate 1000-2000 Correlation Reiwa ed.(163,000 words) test, level 2 or above, or words/set at a time coefficient of 3000 top performers in original (5% of words are duplicated evaluation Kanji reading test duplicated) is more than 0.5 2000 Point (1) Earnest and Point ② About 163,000 target eager evaluators who words, including new words, have knowledge of Kanji e.g., smartphone, internet, Word Familiarity heterodimer, laba... More Less familiar familiar Estimated vocabulary size, **Vocabulary-size estimation** Sample user A Estimated results for each grade Probability of knowing the word 80000 8.0 Please select words that Estimated Vocabulary Size 54176 you know. 9.0 □ bank Estimated 37366 □ charge 9.4 □ intimacy □ recast 000 20000 40000 60000 J3 No. of Words Point 3 Checklist of Point 4 About 4,600 persons, only about 50 words including elementary/junior high/high school students Vocabulary-size test: http://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/

References

- [1] S. Fujita, T. Kobayashi, "Reexamination of word familiarity and comparison with past examination," *The 26th Annual Meeting of the Association for Natural Language Processing*, pp. 1037-1040, 2020.
- [2] S. Fujita, T. Kobayashi, T. Yamada, S. Sugawara, T. Arai, N. Arai, "Vocabulary size of elementary, junior high and high school students and analysis of relationship with word familiarity," The 26th Annual Meeting of the Association for Natural Language Processing, pp. 355-358, 2020.

Contact

Sanae Fujita, email: cs-openhouse-ml@hco.ntt.co.jp Linguistic Intelligence Research Group, Innovative Communication Laboratory

