Measuring Textual Difficulty—Estimating Text Readability and a Person's Vocabulary Size

Sanae Fujita

Abstract

Whether it feels difficult or easy to read text depends on both the readability of that text and the amount of knowledge possessed by the reader. If text readability and the amount of knowledge of the reader (vocabulary size) can be easily and accurately estimated to automatically estimate the difficulty of text, it should be possible to recommend text having a suitable level of difficulty, thereby support the learning process. This article introduces methods we, NTT Communication Science Laboratories, devised for estimating text readability and vocabulary size.

Keywords: vocabulary size, word familiarity, readability

1. What does it mean to measure difficulty?

Taking, for example, a child who is just learning to read, one can imagine how a parent might select a picture book for the child to read on his or her own but end up reading the book to the child who simply finds it too difficult. Additionally, one can imagine how English text that may have been a struggle for a Japanese student to read in the first year of junior high school seems very simple when he/she is a college student. Given the same text, whether it feels difficult or easy depends on the amount of knowledge the reader has.

If recommendations can be made for picture books, novels, textbooks, or even such books in English that are perfectly readable or readable with a little effort, it should be possible to increase the reader's knowledge without difficulty. However, *determining perfectly readable* or *readable with a little effort* is not a trivial task. This is because both a person's amount of knowledge and text readability must be appropriately estimated.

2. Measuring a person's vocabulary size

One type of knowledge that a person needs is vocabulary. At NTT Communication Science Laboratories, researchers have been surveying and estimating the vocabulary size of people in various age groups for over 20 years.

In a survey targeting infants, their vocabulary is not that large, so it is not impossible to investigate all words that can be understood and spoken. We, NTT Communication Science Laboratories, constructed the Child Vocabulary Development Database (CVD) by collecting data on when infants learn and speak what types of words from over 1500 parent-infant pairs.

It is difficult to survey all the words a person knows from elementary school onward. In this case, the approach adopted for estimating vocabulary size is to ask a person whether he or she knows certain presented words. The more words that are presented, the more accurate the estimation becomes, but vocabulary size can still be estimated with only several dozen words.

A key point of this estimation method is hypothesizing how many words a person knows when answering that he or she knows a certain word. For example, given the words "salty" and "tidal gauge," there would probably be less people who know "tidal gauge." The assumption can therefore be made that a

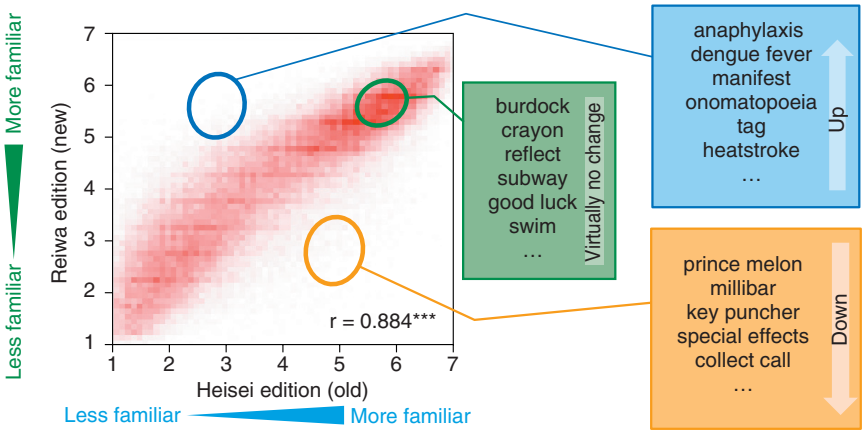


Fig. 1. Change in word familiarity: from Heisei to Reiwa.

person who knows the words "tidal gauge" has a larger vocabulary than a person who only knows "salty." Given that a person is knowledgeable of "tidal gauge," how many words can we assume that person knows? The basis for such an assumption is *word familiarity*.

3. Word-familiarity database

The quantification of the degree of word recognition through rating experiments is called *word familiarity*. A word assigned a large number corresponds to a word that many people are familiar with, and a word assigned a small number corresponds to a word that many people are unfamiliar with.

We have been constructing fundamental language resources, such as a word-familiarity database, for over 20 years. The Heisei edition of such a database (*Heisei* is the name of the previous Japanese era from Jan. 8, 1989 to Apr. 30, 2019) consisting of about 77,000 entries was released in 1999 as part of the NTT Database Series "Lexical Properties of Japanese." This series of databases has been widely used as a basic reference in such fields as psychology, language education, and speech and language therapy. However, over 20 years have passed since the initial surveys, which opened up the possibility that word familiarity has changed with the times. There is also the problem that new words (such as "Internet" and "convenience store") are not included.

In the face of these issues, we re-examined the relevance of words appearing in the Heisei edition, took up the inclusion of new words, and constructed the largest word-familiarity database consisting of about 163,000 entries as the Reiwa edition (*Reiwa* is the name of the current Japanese era that began on May 1, 2019) [1]. We also investigated changes in the word-familiarity database from the Heisei edition and found that a strong correlation exists between the two editions with no major change in familiarity for many words after more than 20 years. However, some words did in fact undergo significant change in familiarity, and we clarified what types of words underwent such change (**Fig. 1**).

4. Estimating vocabulary size using word familiarity

To estimate vocabulary size using word familiarity, we set up questions by sampling words from high to low degrees of familiarity (**Fig. 2**). Specifically, given a word for which a participant's answer is "Yes, I know this word," we assume that that participant knows all words with a degree of familiarity higher than that of that word and estimate vocabulary size accordingly. For example, if a participant's knowledge extends up to "salty," we assume the vocabulary size to be about 1500 words, but if extending up to "tidal gauge," we assume a vocabulary size of about 139,000 words. We consider, however, that variation likely exists as to what is known near the boundaries between known and unknown words. We therefore apply a logistic regression curve to the participant's answers and take the vocabulary size for which the probability of knowing is exactly 50% as the estimation result (**Fig. 3**).

With this method, the vocabulary size of a participant can be easily estimated by simply having the

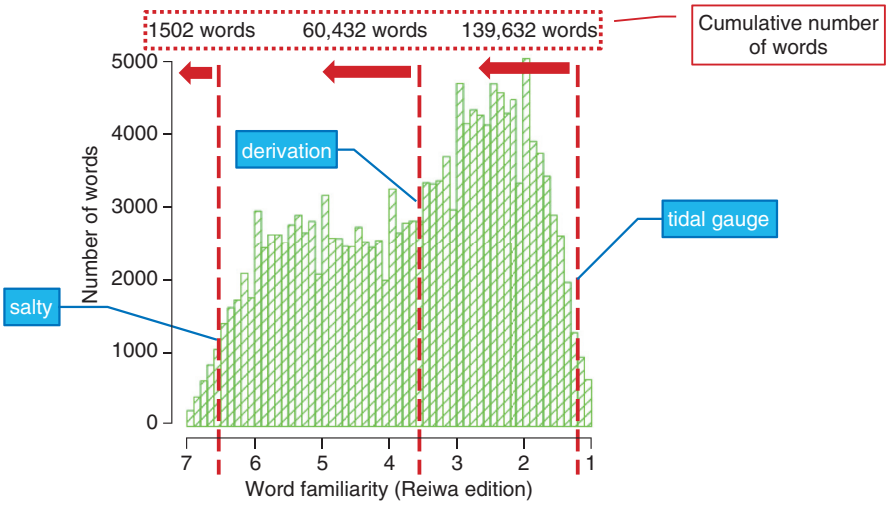


Fig. 2. Histogram of word familiarity: sampling words form high to low degrees of familiarity.

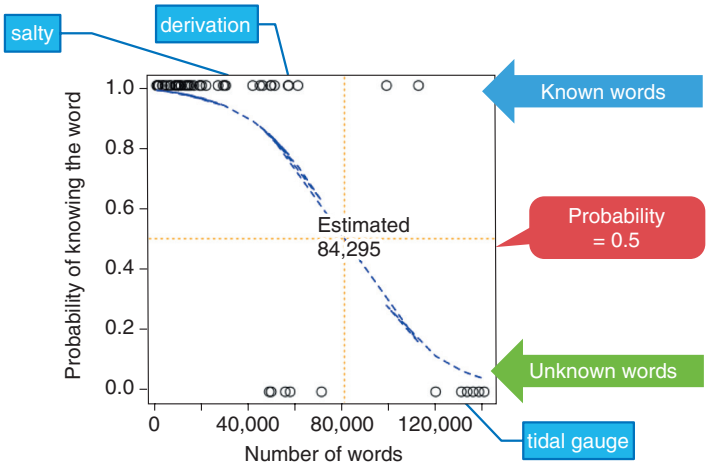


Fig. 3. Method of vocabulary-size estimation.

participant check a small number of sampled words. In addition, since there are words that theoretically have the same degree of familiarity, any of those words may be posed as a question, which means that changing presented words is easy and tests with a certain amount of variation can be prepared. Of course, the more words a participant is asked to check, the greater the estimation accuracy.

The upper limit of vocabulary size that can be estimated with this method depends on the size of the word-familiarity database, but construction of the Reiwa edition of the word-familiarity database significantly raised the upper limit of the vocabulary size that can be estimated, thereby enhancing the versatility of the test.

5. Large-scale survey of vocabulary size

There has essentially been no large-scale vocabulary-size surveys targeting children and students, but we used the method described above to carry out a vocabulary-size survey of 4600 individuals including more than 2800 public-school students from elementary school to high school.

The results revealed that vocabulary size increased rapidly for elementary-school students (6–7 to 11–12

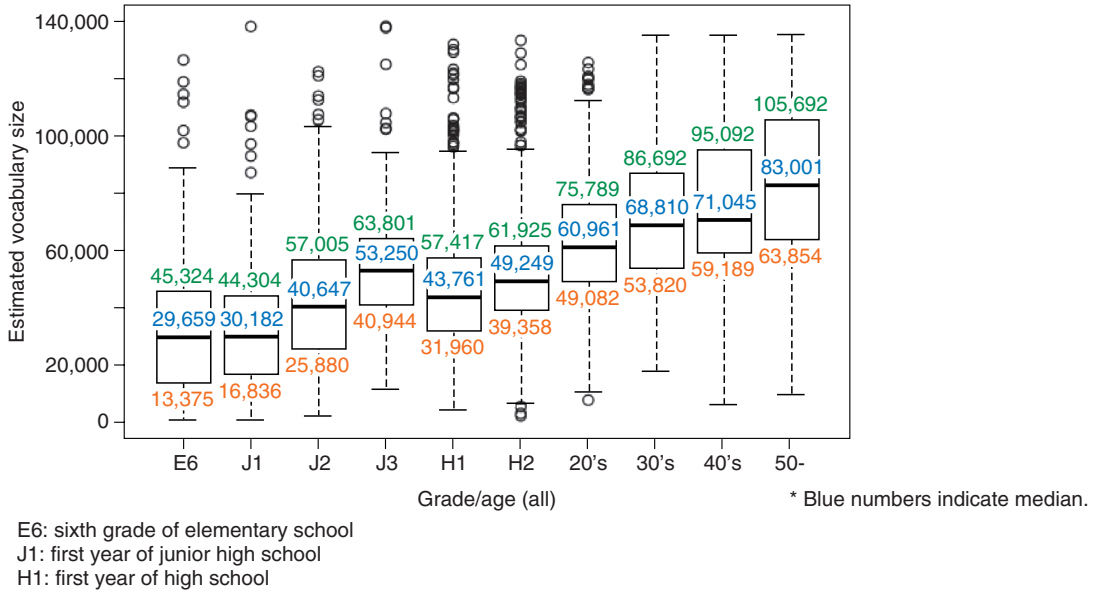


Fig. 4. Estimated results for each grade/age.

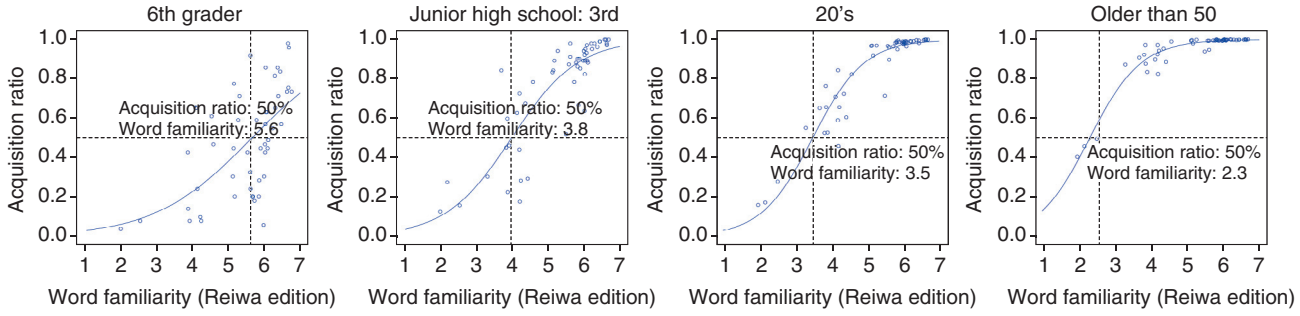


Fig. 5. Relationship between word familiarity and acquisition ratio for each grade/age.

years old) and junior-high students (12–13 to 14–15 years old) and increased for adults along with age (**Fig. 4**). These results also indicate large variation in vocabulary size even among same-year students, so such survey results should be useful in identifying students in need of assistance [2].

We also analyzed the relationship between word familiarity and vocabulary acquisition (percentage of individuals knowing that word) for various school years and age groups, as shown in **Fig. 5**. This figure shows the percentage of individuals answering "Yes, I know this word" for words of various degrees of familiarity (acquisition ratio) for various school years and age groups. These results indicate that the percentage of individuals answering in the affirmative tends to increase for higher degrees of familiarity for any school year or age group and that this trend becomes especially clear with age. However, compared with adults, there is much individual difference and variation regarding knowing or not knowing among elementary-school students and junior-high students, even for words that have a relatively high degree of familiarity. Based on this analysis, it should be possible to use word familiarity as a clue in identifying vocabulary that children and students should prioritize from then on or vocabulary that should be acquired.

We also released a vocabulary-size estimation test based on word familiarity (Reiwa edition) on the web for use by the general public in conjunction with NTT

Communication Science Laboratories Open House (released on June 4, 2020, http://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/). About ten days after its release, more than 30,000 individuals have taken the test. We encourage everyone to give it a try.

6. Measuring the difficulty of text

We now introduce a method for estimating text readability. We began this research by estimating the readability of a picture book. It was thought that combining this method with research on child vocabulary development conducted at NTT Communication Science Laboratories might contribute to clarifying and supporting vocabulary development.

Text readability is influenced by both the difficulty of the vocabulary used and the difficulty of sentence structure. In the case of picture books, this requires an accurate analysis of *hiragana* (characters that represent sounds in written Japanese). For example, the readability of the hiragana characters "とうさん" (pronounced *tousan*) varies greatly depending on whether they correspond to "父さん" (also pronounced *tousan* but meaning "father") or "倒産" (likewise pronounced *tousan* but meaning "bankruptcy"), where "父" and "倒産" are *kanji* (Chinese) characters. To improve the accuracy of such *hiragana* analysis, we used features that reflect the difficulty of the vocabulary used but referring, for example, to the CVD, and features that reflect the difficulty of sentence structure such as sentence length. We were able to estimate text readability with a level of accuracy as high as 87.8% in terms of classifying picture books into target readership, namely, the four age groups of 0–2, 3, 4, and 5 [3]. However, even if a certain picture book is described as "recommended for children 3 years of age," for example, it may in fact be appropriate for 3-year olds closer to 2 years of age or 3-year olds closer to 4 years of age. In short, estimating the target age at a year before or after the stated recommended age should not present much of a problem. If we therefore consider that such approximate estimations are acceptable, the level of accuracy that we achieved would in fact be 96.7%. It can therefore be said that the proposed method is robust and highly reliable in estimating text readability with practically no instances of greatly mistaking the target age. This method is currently being used to search for picture books suitable for children of various ages as part of NTT's *Pitarie* picture book search system [4].

Much research on estimating text readability is focused on evaluating whether the school year targeted by individual textbooks can be estimated. With this in mind, we also applied this method to estimating the target year of textbooks and found that it has a high level of accuracy above 98% for nine categories from the first year of elementary school to the third year of junior high school. This result indicates that our method for improving the accuracy of estimating the readability of picture books is also useful in estimating the readability of text for elementary and higher students.

7. NTT corpus of picture books and children's books

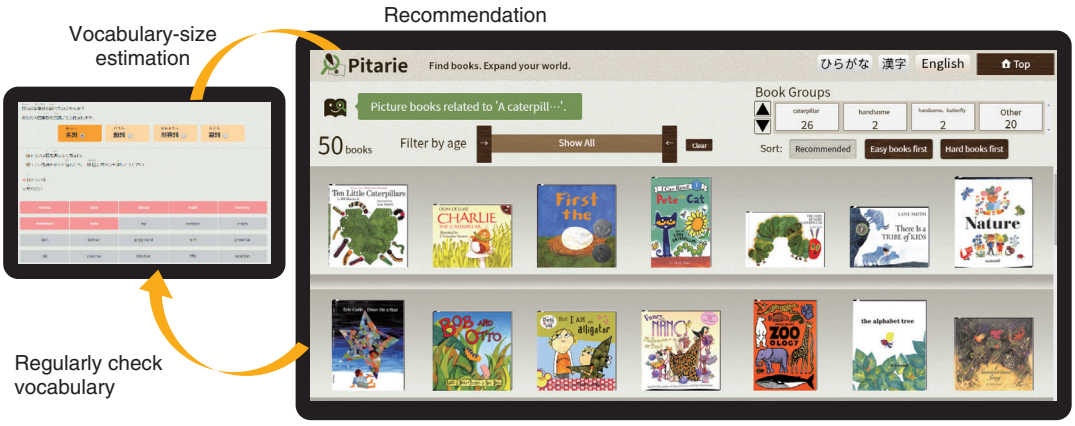
Estimating the text readability described above uses digitized picture-book-text data (corpus). However, no picture book corpus existed, so we had to begin our research by constructing such a corpus. This part of our research was the most arduous and time consuming. The pictures in picture books often include characters, which makes character recognition by optical character recognition infeasible. In the end, most text had to be entered manually. This painstaking work paid off in the form of an NTT picture book corpus consisting of more than 6000 volumes of picture books in Japanese and more than 2500 volumes of picture books in English, a corpus of unprecedented size. There are plans to expand this corpus even further.

We are also actively engaged in many new research projects using this picture-book corpus and children's books such as investigating the relationship between picture books and child vocabulary and emotional development.

8. Future developments

Estimating a person's vocabulary size and text readability have up to now been conducted independently. However, combining the two should make it possible to recommend text that is *perfectly readable* or *readable with a little effort* for any individual while periodically checking that person's vocabulary size. In fact, we are also researching the estimation of English vocabulary size and English text readability and beginning an initiative to recommend picture books in English appropriate to a person's vocabulary level for use in English language education at Japanese schools (**Fig. 6**) [5].

Going forward, we aim to provide child-rearing and educational support in both Japanese and English for every person from elementary through high school as



*Information on the picture books shown here is given after the reference section.

Fig. 6. Content recommendation suitable for vocabulary size.

well as adults while collecting evidence on its effectiveness.

References

[1] S. Fujita and T. Kobayashi, "Reexamination of Word Familiarity and Comparison with Past Examination," Proc. of the 26th Annual Meeting of the Association for Natural Language Processing, pp. 1037–1040, Mar. 2020.

[2] S. Fujita, T. Kobayashi, T. Yamada, S. Sugawara, T. Arai, and N. Arai, "Vocabulary Size of Elementary, Junior High and High School Students and Analysis of Relationship with Word Familiarity," Proc. of the 26th Annual Meeting of the Association for Natural Language Processing, pp. 355–358, Mar. 2020.

[3] S. Fujita, T. Kobayashi, Y. Minami, and H. Sugiyama, "Target Age Estimation of Texts for Children," Journal of Japanese Cognitive Science, Vol. 22, No. 4, pp. 604–620, 2015.

[4] S. Fujita, T. Hattori, T. Kobayashi, Y. Okumura, and I. Aoyama, "Picture-book Search System 'Pitarie'—Finding Appropriate Books for Each Child—," Journal of Natural Language Processing, Vol. 24, No. 1, pp. 49–73, 2017.

[5] S. Fujita, T. Hattori, T. Kobayashi, and F. Naya, "Estimation Method of Vocabulary Size of English Learners," The 34th Annual Conference of the Japanese Society for Artificial Intelligence, June 2020.

Information on the picture books shown in Fig. 6

Top row (from left):

B. Martin Jr (author), L. Ehlert (illustrator), "Ten Little Caterpillars," Beach Lane Books, 2011.

D. DeLuise (author), C. Santoro (illustrator), "Charlie the Caterpillar," Aladdin, 1990.

L. Vaccaro Seeger (author and illustrator), "First the Egg," Roaring Brook Press, 2007.

J. Dean, "Pete the Cat and the Cool Caterpillar," Series: I Can Read! 1, HarperCollins, 2018.

E. Carle (author and illustrator), "The Very Hungry Caterpillar," Philomel Books, 1969.

L. Smith (author and illustrator), "There Is a Tribe of Kids," Two Hoots, 2018.

A. Grée, (author and illustrator), "Nature," Button Books, 2012.

Bottom row (from left):

E. Carle (author and illustrator), "Draw Me a Star," Philomel Books, 1992.

R. O. Bruel (author), Nick Bruel (illustrator), "Bob and Otto," Roaring Brook Press, 2007.

L. Child (author), "But I Am an Alligator," Series: Charlie and Lola, Grosset & Dunlap, 2008.

J. O'Connor (author), R. Preiss Glasser (illustrator), C. Bracken (illustrator), "Fancy Nancy Halloween...or Bust!," Series: Fancy Nancy, Harper Festival, 2009.

J. Jolivet (author and illustrator), "Zoo Ology," Roaring Brook Press, 2002.

L. Lionni (author and illustrator), "The Alphabet Tree," Alfred A. Knopf, 1968.

I. Haas (author and illustrator), "A Summertime Song," Margaret K. McElderry Books, 1997.

Sanae Fujita
Senior Research Scientist, Linguistic Intelligence Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

She joined NTT Communication Science Laboratories in 1999 and received a Ph.D. in engineering from Nara Institute of Science and Technology in 2009. She is currently researching natural language processing, especially word disambiguation, text-readability estimation, and vocabulary-size estimation. She received the 2013 Best Paper Award of Journal of Natural Language Processing and 2018 NTT President Award. She is a member of the Association for Natural Language Processing and the Information Processing Society of Japan.