IMPROVEMENT OF AN ENGLISH READABILITY SCALING TOOL WITH MACHINE LEARNING: A STUDY ON THE DEVELOPMENT OF A NEW DICTIONARY DATABASE IN THE READABILITY TOOL

Mariko Sakamoto
(sakamoto@kagoshima-ct.ac.jp)
National Institute of Technology, Kagoshima College, Japan

Akira Nakano
(nakano@kurume-nct.ac.jp)
National Institute of Technology, Kurume College, Japan

Abstract

The purpose of the current study is to improve the English readability scaling tool for Japanese English teachers and learners, named "CheckRead (Fukui, 2022)" developed by Fukui and Ozasa, with using machine learning. As the first step of the study, the present paper dealt with the improvement of the dictionary database used in Ozasa-Fukui Year Level Index in "CheckRead". It was based on the former version of the educational guidelines of foreign language provided by Ministry of Education, Culture, Sports, Science and Technology, Japan, with using the English co-occurrence dictionary and machine learning so that the readability index could precisely measure the readability level of the new English textbooks authorized by the new educational guidelines of foreign language. The authors extracted common verbs as "have", "ask", and be-verbs to see their behavior in the co-occurrence dictionary, and found that it would be possible to collect the information on the structure of the sentences in the dictionary with machine learning.

1 Purpose of the Study

In this paper we reported the first trial of the whole study on the improvement of a readability index called "Ozasa-Fukui Year Level Index (OFYL; Fukui and Ozasa, 2017, Sakamoto et. al., 2019, Fukui, 2022). We require several processes, experiments, and researches in order to achieve our purpose of the study, and this paper was presented as one of the processes on building databases. The purpose of the present paper was to improve the dictionary information which consists OFYL, producing databases through machine learning.

2 Motivation of the Study

The Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan had started the revision of the educational guidelines from 2020, from an elementary school to a high school in turn year by year. Textbooks of all the subjects were written based on the educational guidelines (= course of study) to be authorized by the government, meaning that

the publishers were needed to change all the contents of the textbooks along with the new course of study.
The present English readability measuring index, OFYL, has a high reliability value reflecting the judgment of the experienced experts of English teaching methodology and pedagogy. However, it also has to be renewed, since it also has to follow the modification of the foreign language education guidelines with keeping its reliability and validity.

In the present study, we hoped to introduce machine learning as a new method for the improvement of OFYL index instead of adopting the manual procedure by experts of English teaching methodology and pedagogy. One purpose of introducing this new method was to integrate the experts' knowledge and skills into the automatic system so that anyone would be able to make the best use of the heritage in future. The other purpose was to reduce the burden of the developing process of creating a completely new database for the new index. On the process of making a database in OFYL, the researchers were required to discuss, judge, and tag the acquiring-year/month-level to every single sentence all by manual. This method could be possible this time, too, since it had been done before, but it would not be realistic to carry out the same procedure every time when the course of study was modified, because of the amount of work and the shortage of experts. These circumstances brought us to the research method of the use of machine learning instead of human resources.

The present paper shows the process of producing the databases, especially of the "dictionary information" and an English textbook database, which are parts of the components of OFYL. The "dictionary information" refers to not the ordinary dictionary which we use in language learning but an EDR digitalized dictionary, which we will mention in the section of the Method.

3 Background

3.1 Ozasa-Fukui Year Level Index (OFYL)

Ozasa-Fukui Year Level Index (OFYL) had been developed as an objective English readability measuring index for Japanese English learners. It judges the readability level of an English sentence in terms of acquired year and month in Japanese schools as elementary school, junior high school, and high school. The calculation formula of OFYL is as follows:

4.6579*exp(-17.7116*0.3716^Diff)
where
Diff = 0.995*Words/S + 0.4302*Syllables/W + 0.9800*WordDiff/W + 0.0633*IdiomDiff/S + 0.2815

(Sakamoto et. al., 2019, p.24)

The explanatory variables in this formula refer to the number of words in a sentence, the number of syllables in a word, the value of difficulty level of a word, and the value of difficulty level of idioms in a sentence.

Also, OFYL consists of 916 sentences with the information of the difficulty level, gathered from three sets of textbooks and judged by three experts of English teaching methodology and pedagogy. Fukui and Ozasa conducted nonlinear analysis with the variables and achieved a high reliability value as r2=.8217 (Fukui and Ozasa, 2017, p.8). A strong point of this index is that the meaning of the value is set for Japanese EFL learners so that a user can interpret the level instantly. For example, the readability levels of an English graded reading material "The Adventure of Tom Sawyer" are 81.97 in Flesch Reading Ease Score (FRE), 5.87 in Flesch-Kincaid Grade Level (FKG), and 3.8 in OFYL. It means that this English reading material is "Easy" level in FRE, corresponding to a 6-year-grade level of US readers according to FKG. In OFYL, on the other hand, it shows that it is suitable for a 3.8-year-experienced English learner in Japan, meaning a student who is in January of the first-year-grade in junior high school. This enables Japanese English learners to know the readability level of materials easily, and for English teachers in Japan, it helps to choose or develop authentic materials which are suitable for their students (Sakamoto, 2018).

3.2 Machine Learning

As for the development of the readability measuring tool with machine learning, there are several researches not only on English as a native language (L1), but also on English as the second language (ESL) and as a foreign language (EFL). For example, Xia et. al. (2016) reported their process of development of English readability measuring tool for European ESL learners. In their research, they experimented with domain adaptation and self-training approaches to use the plentiful amount of L1 data in order to compensate for the large amount of corpus.

Machine learning requires plenty of data for their self-learning and also for assessment dataset of the results of the learning. In the case of OFYL, the database for the assessment corresponds to the one developed by the experts of English teaching methodology. It also be needed to renew the information of dictionaries and idioms. Hence, for this research, it was mandatory to collect sufficient amounts of data for machine learning and to find a tool which would be possible to apply on the information of dictionaries and idioms.

4 Method

The present research required tools such as the co-occurrence digitalized dictionary, junior high school English text data and SQLite3. We explained the process of building up the databases in the following sections.

4.1 Co-occurrence Digitalized Dictionary

As the first issue towards the new dictionary information, we needed to select a dictionary data which would be able to compensate for the amount of English textbook dictionary information with similar, or at least not too different, characteristics. The dictionary items and sentences in the textbooks in Japan were controlled based on the course of study provided by MEXT, meaning they contained the grammatical items which were thought to be suitable for each school level in simple structures, unlike natural corpus (Sakamoto, et. al., 2018). Therefore, we needed to look for the corpus which contained a certain quantity of items enough to compensate for the English textbook data without L1 or natural English corpus. Under this circumstance, we focused on a digitalized dictionary. The digitalized dictionary we chose was a co-occurrence dictionary (ECC.DIC), which was one of EDR digitalized dictionaries created by a Japanese project which aimed to be useful in R&D of natural language processing and the next generation of knowledge processing systems.

The EDR was not an ordinary dictionary which language learners generally use, but it was a database which was compiled in order to use for programming. Also, the EDR Electronic Dictionary was developed for advanced processing of natural language by computers, and it was composed of eleven sub-dictionaries, for instance a concept dictionary, word dictionaries, bilingual dictionaries, and so on.

ECC.DIC, which was one of EDRs, dealt with the collocation and/or the relationship between two words in a sentence. It made us expected that it was suitable for obtaining the information on idioms which also consisted of the index. ECC.DIC consisted of 615,997 text data items with 227MB. A data at the English co-occurrence dictionary was composed of the record number, headword information, co-occurrence constituent information, syntactic information, semantic information, co-occurrence situation information, and management information. By using this dictionary, a user could learn actual examples of how words were combined in sentences, and their frequencies of the combinations.

4.2 Database of English Textbooks

The textbook database was created by digitizing junior high school English textbooks published in Japan. All of the sentences were separated into each line, and inputted with the information of sentence structures and grammatical items as relative-clause and to-infinitive. This database was made of four kinds of textbooks called “New Crown”, “New Horizon (2018 edition and 2020 edition)”, and “Prominence”, each of them had three volumes, and the database contained 5,316 sentences in total. For the present study, we used 1,317 sentences from “New Crown" vol.1-3 in the database.

4.3 Conversion of ECC.DIC and the Database of English Textbooks through SQLite3

ECC.DIC and the file of English textbooks were text datum. In ECC.DIC, each data was a complex description by using the semi-colons and nest constructions. Therefore, as the next procedure, we created a program to fix these symbols, and extracted headword information and co-occurrence situation information. The headword information had three kinds of information: Word 1 Notation, Co-occurrence Relation Label and Word 2 Notation. The co-occurrence situation information had four kinds of information: Surface Level Co-occurrence Frequency, Co-occurrence Entry Frequency, Governing Morpheme Frequency and Dependent Morpheme Frequency. In this procedure, converted datum were registered to the databases of SQLite3. From the converted ECC.DIC database, we obtained the seven kinds of information mentioned above. Some of the extracted examples were shown in Table 1.

Table 1. Seven kinds of information extracted from ECC.DIC.

No.	Word1	Connector	Word2	FS	FCC	FW1	FW2
1	ask	about	it	5	1	413	12441
2	ask	@d-object	you	28	1	413	8861
3	lend	without	ask	1	1	51	413
4	question	@passive-subj	ask	3	1	251	290

Co-occurrence Relation Label (=Connector) showed the relationship between Word 1 Notation (=Word1) and Word 2 Notation (=Word2) with semantic/syntactic tags, prepositions, or grammatical information as was shown in the examples in Table 1. For example, the connector showed the semantic/syntactic/structure relationship with @-tags, or prepositions.

Surface Level Co-occurrence Frequency (=FS) showed the frequency of the combination of Word1 and Word2. According to the FS of no.1 and 2 in Table 1, the combination of "ask + you" showed higher frequency than "ask + about + it".

Co-occurrence Entry Frequency (=FCC) had to do with the co-occurrent relationship of the abstract levels in the concept dictionary. Governing Morpheme Frequency (=FW1) meant a frequency of Word1 with the usage in a sentence, while Dependent Morpheme Frequency (=FW2) meant a frequency of Word2 with the usage in a sentence. It meant that the frequency of the usage of "ask" in no.3 line in Table 1, appearing with "lend" and "without" had higher frequency than the one in no.4, passive with its subject "question".

4.4 Processing English Textbook Database with TreeTagger

On the English textbook database, we applied a tagger program called the "TreeTagger (Schmid, 1994, 1995, 2002)".

The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart (Schmid, 2022).

After the words in the database were all tagged with the TreeTagger, we extracted ten kinds of verbs as "ask", "play", "read", "have", "has", "had", "is", "was", "are", and "were" from the database. We chose these verbs because we expected that the frequencies of these verbs were high in the junior high school English textbooks, and we would have liked to see if there were any differences in the conditions of the transitivity, person, plurality, or tense. We also expected to be able to observe the present perfect aspect from "have/has", and passive structure and progressive aspect from be-verbs.

4.5 Processing the Selected Verbs with ECC.DIC

We examined the verbs mentioned above, with applying the database of ECC.DIC onto the English textbook database. We listed up the verbs as Word1 and Word2, and extracted the combinations by the connectors. The idea was that it would be expected the items with high-frequency combinations would appear in sub-materials, supplement materials, drills, or examinations yet they did not show up in the textbooks. Therefore, we considered that it would be reasonable to add those combinations to the database in the future research as the expressions which the learners would acquire in certain year-grades.

Table 2. Variety of Connectors of "ask"

Word1		Word2	
Connector (@-tags)	Connector (prep.)	Connector (@-tags)	Connector (prep.)

@d-object	about	over	@d-object	-
@i-object	after	through	@i-object	
@o-complement(to_be_done)	at	to	@o-complement(to)	
@o-complement(to)	by	under	@passive-subj	
@passive-by	for	via	@pre-modifier	
@passive-complement(to)	from	with		
@passive-object	in			
@passive-subj	into			
@post-modifier	of			
@pre-modifier	on			

Using the database which we developed with the procedure in the Method, first we extracted the information of a word "ask" as was shown in Table 2. In the "Connector" row, we could observe the semantic/syntactic/structure relationships between Word1 and Word2. When we set "ask" as Word2, most of the relationships between Word1 were @subject, @passive-subj, or @s-complement(to), as "ask" mostly appears as a verb in a sentence. When we set the word as Word1, there were more variable connectors as seen in Table 2.

5 Analysis

We made an assessment on the produced database. Table 3 showed the examples of what we obtained from the processed database.

The total number of the co-occurred cases of "ask" was 1,804 cases, 1,134 cases as Word1 and 670 cases as Word2. The varieties and frequencies of the combinations which appeared in the database on "ask" were shown in Table 2.

Table 3. Sample of the Output: "ask" as Word1

Word1	Connector	Word2	FS	FCC	FW1	FW2
ask	@d-object	you	28	1	413	8861
ask	@d-object	time	4	1	413	241
ask	@d-object	question	55	1	413	191
ask	@d-object	people	9	1	413	55
ask	@d-object	me	22	1	413	4
ask	@i-object	you	18	1	413	8861
ask	@o-complement(to)	giv	3	1	251	187
ask	@o-complement(to)	liv	1	1	251	215
ask	@o-complement(to)	join	6	1	251	200
ask	@passive-complement(to)	give	3	1	251	1
ask	@passive-complement(to)	investigat	2	1	251	1
ask	@post-modifier	much	3	1	51	4

ask	@post-modifier	politely	2	1	51	8
ask	about	it	5	3	413	12441
ask	for	it	1	1	413	12441
ask	for	fund	2	1	251	197
ask	in	English	1	1	413	229
ask	of	monitor	1	1	251	4
ask	to	be	2	1	251	191

In Table 3, @d-object stood for the direct object and @i-object for the indirect object, which suggested that the structure of the sentence tagged @d-object was interpreted as a SVOO structure, and @i-object was interpreted as a SVO structure. Also, the cases which tagged with @post-modifier or prepositions seemed to show that they had SV structures or phrasal verbs as "ask for something". @o-complement(to) and @passive-complement(to) were the connectors which involved verbs in Word2. These connectors were interpreted as the to-infinitives which appeared in an active or a passive sentences. The value of FS in "ask + @d-object + question" was far higher than others. It suggested that this combination was able to be interpreted as a collocation.

The English teaching method in Japan, based on the course of study provided by the government, has adopted the concept of the "the five sentence structures (Onions, 1903, Routledge and Paul, 1971)". Therefore, to assign the information about each sentence structure as SV, SVC, SVO, SVOO, and SVOC would become a clue of presenting the acquiring year level or difficulty level of the sentence in the database.

Obtaining the viewpoints above, we applied the same analysis on the verbs as "play", "read", "be-verbs (= is/was/are/were)", and "have/has/had". The frequency of each verb in Word1 and Word2 was shown in Table 4. We considered that the reason for the high frequency of "have/has/had" was that they appeared not only as the general verbs but also as the perfect aspect and modality. As for the tense, we could observe from the frequencies in Word2 that the verbs with past tense were lower than the one in the present tense. We considered the reason was that the past tense was introduced later than the present tense in English class in a junior high school.

Table 4. Frequency of Selected Verbs in Word1 and Word2

verb	Word1	Word2
play	975	764
read	281	219
is	12	224
was	3	103
are	3	137
were	5	42
have	3050	1742
has	1896	1455

had	1761	1154

In Table 5 and 6, common examples of other verbs from the database were shown. The output as Word1 was in Table5, and as Word2 was in Table 6.

Table 5. Sample of the Output: selected verbs as Word1

Word1	Connector	Word2	FS	FCC	FW1	FW2
had	@d-object	address	1	1	531	23
had	@o-complement	salt	1	1	279	1
had	@post-modifier	lately	1	1	6	33
has	@d-object	chance	9	4	997	198
have	@d-object	way	13	2	558	652
have	@d-object(ing)	try	1	1	52	6
have	@d-object(to_be_done)	confirm	1	1	37	95
have	@d-object(to)	act	1	1	558	42
have	@o-complement(pp)	cancel	1	1	121	67
play	@d-object	game	30	2	176	77
play	@d-object	role	96	1	176	77
play	@post-modifier	together	2	1	176	42
play	@post-modifier	yesterday	1	1	176	185
play	by	rule	4	1	176	3
read	@d-object	book	23	6	114	407
read	@d-object	me	1	1	114	1547
read	@post-modifier	immediately	1	1	85	137
read	at	random	1	1	85	1
read	in	order	1	1	18	86
read	to	class	1	1	18	14
is	among	best	1	1	240	90
is	for	more	1	1	240	395
are	among	long	1	1	157	314
was	in	city	1	1	98	11
were	in	red	1	1	43	1

Setting the selected verbs in Word1, the semantic/syntactic/structure and preposition tags which were detected in the Connector row were shown in Table 5. In the database, the most observed connector of the verb “have” was @d-object, which we considered as reasonable, since the verb “have” generally needs an Object. There were 2,290 cases out of 3,050. 85 cases in the connector were classified as @d-object(to) and 5 cases were as @d-object(to-be-done), which suggested that these verbs functioned as the modality or idioms in each sentence. For the case that “have” was used as the present perfect, the tag @o-complement(pp) was attached, followed by another verb in Word2. We observed 35 cases of them. The tag @passive-subj, which seemed to be related to the passive construction, was observed when the verb “play” and “have” were set as Word2. Also, we observed @d-object, @post-modifier, and prepositions in the Connector row of “read”.

As for the be-verbs in Word1, the frequency in the database was very low (see, Table 4), and most of them appeared with prepositions. There seemed to be no differences in the conditions of tense or person.

Table 6. Sample of the Output: selected verbs as Word2

Word1	Connector	Word2	FS	FCC	FW1	FW2
undoubtedly	@pre-modifier	had	1	1	531	28
typhoon	@subject	had	1	1	127	37
also	@pre-modifier	has	54	1	997	8
Chicago	@subject	has	1	1	997	8
yes	@pre-modifier	have	11	1	1338	45
we	@subject	have	331	6	1338	5053
also	@pre-modifier	play	16	3	176	3729
sit	@s-complement(ing)	play	1	1	23	176
sign	@s-complement(to)	play	1	1	7	176
he	@subject	play	46	5	89	12340
want	@o-complement(to)	read	1	1	152	114
really	@pre-modifier	read	1	1	85	403
we	@pred-subj	read	4	1	1	5053
child	@subject	read	1	1	7	135
also	@pre-modifier	is	6	2	240	3729
defin	as	is	1	1	12	43
first	@pre-modifier	are	3	1	157	58
There	@pre-modifier	are	5	2	157	6
so	@pre-modifier	was	4	1	98	190
unfortunately	@pre-modifier	was	1	1	98	37
fortunately	@pre-modifier	were	1	1	43	34
There	@pre-modifier	were	4	1	43	6

From the output which we set the verbs as Word2, we could observe the following results. The sample of the output was shown in Table 6. For all of the general verbs as “have/has/had”, “play”, and “read”, the semantic/syntactic/structure tags in the Connector row were mostly @subject, then followed by @pre-modifier. This would be because of the basic word order of English S (+ modifier) + V. There were a few outputs meaning the complements: @s-complement(ing)/(to) and @o-complement(to), too. It suggested that the @s-complement(ing) represented the particle structure and @s/o-complement(to) represented the to-infinitive. As for the be-verbs, almost all of the outputs had the connector @pre-mofdifier. No particular differences among tenses or persons were observed.

6 Discussion

Because the developed database was based on a series of the English textbooks for Japanese junior high school learners, the variety of grammatical forms and verbs appeared in the database were controlled based on the course of study. Hence, we expected it to be possible to add grammatical information to the database automatically by selecting the grammatical items and tagging the year, grade, or month level to the database. For example, the information on the five sentence structures would show a high versatility because it was related to all kinds of verbs and sentences. The sentence structures tended to be introduced in a certain learning order in the junior high school English class in Japan, so it seemed to be possible to set the acquiring level or timing according to the sentence structures. Also, the tags as @d-object(to), @passive-subj, @o-complement(pp), @s-complement(ing), and @s/o-complement(to) were expected to become the sufficient clue for labeling items in the database with judging the acquiring level or timing of the grammatical forms.

As we have discussed above, it was considered to be a meaningful method to assign the difficulty level automatically with using the semantic/syntactic/structure tags and grammatical forms in the co-occurrence dictionary for the purpose of expanding the size of dictionary database, although it was still not on the level of discussing its accuracy. When considering that OFYL was created with the database which several experts of English pedagogy chose or produced each sentence and assigned the level to them all by manual, the method of the development of the database and the assessment in this paper were suitable for our purpose of the study, improving the readability measuring index without taking a manual procedure.

As for the dictionary information, we concluded that the English co-occurrence digitalized dictionary would be helpful as the material of developing the new readability measuring index by means of containing general word items which have similar characteristics to the words in the textbooks. It means that they would be useful because it gave us not only frequency information but also grammatical and sentence-structure information. Again, it would help to expand the dictionary information because it would be possible to generate the sentences which were not in the textbooks but plausible to be used in the sub-materials, supplement materials, or examinations, since they have the same grammatical characteristics and frequencies.

7 Conclusion

In this paper we reported the first stage of the process of improving the English readability measuring index through machine learning. We produced a database from the English co-occurrence digitalized dictionary and text data from a series of English textbooks, "New Crown", through machine learning. We could observe some of the sentence structures and grammatical items, for example the five sentence structures, present perfect, to-infinitive, and passive form from the semantic/syntactic/structure tags between Word1 and Word2 in the database. This suggested that it was possible to produce a database with the information of difficulty levels depending on the grammatical information through machine learning. In the future research, we would examine other verbs to find other tags to show the grammatical forms and collocations of the sentences in order to increase the size and accuracy of the assessed database. These findings brought us to conclude the possibility of expanding the number of sentences in dictionary information for developing the new readability measuring index with automatic means.

Acknowledgement

We would like to show our gratitude to JSPS for supporting this study by KAKENHI Grant Number JP19K00902.

References

Fukui, T. (2022). *CheckRead Ver.3.0, CheckRead Light Ver.3.0*. Retrieved from https://www.heisei-u.ac.jp/ba/fukui/readability.html

Fukui, T., Ozasa, T. (2017). Ozasa-Fukui Year Level, a Readability Measuring Tool: Its System and Measuring Program. *Journal of ICTATLL Japan, 4*, 1–12.

Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. (2014). *Educational guidelines (foreign language)*. Retrieved from https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_010.pdf

Negishi, M. et. al. (2021). *New Crown*. Vol. 1–3. Sanseido.

Onions, C.T. (1903). *An Advanced English Syntax*. OUP.

Onions, C.T. (1971). *Modern English Syntax*. In Routledge and Paul, K. Transferred to digital printing in 2002. https://books.google.co.jp/books?id=saxVEAAAQBAJ&printsec=frontcover &hl=ja&source=gbs_ge_summary_r&cad=0

Ozasa, T. and Fukui, T. (2009). Development of a Readability Index for EFL Learners. *Report of JSPS KAKENHI Grant Number JP19520535*. Retrieved from https://kaken.nii.ac.jp/grant/KAKENHI-PROJECT-19520535/

Sakamoto, M. (2018). An analysis of the readability of English graded readers: for the authentic use in classroom. A presentation handout of *the 6th conference of ICTATLL Japan*, Tokyo, Japan.

Sakamoto, M., Zhang, S., Akase, M. (2018). A study of perfect forms in English textbooks of China and Japan. *Journal of ICTATLL Japan, 5*, 23–35.

Sakamoto, M., Watanabe, K., Uenishi, K., Ozasa, T. (2019). An Analysis on the Readability of Thai and Japanese English Textbook. *Journal of ICTATLL Japan, 6*, 21–29.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of *International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid, H. (2022). *TreeTagger - a part-of-speech tagger for many languages*. Retrieved from https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Twain, M. (2008). *Penguin Readers: Level 1, The Adventure of Tom Sawyer*. Pearson Japan.

Xia, M., Kochmar, E., Briscoe, T. (2016). Text readability Assessment for Second Language Learners, In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Education Applications*, 12–22.