Inferring CEFR Reading Comprehension Index Based on Japanese Document Classification Method Including Pre-A1 Level

My Nguyen Tra HUYNH[*1], Yoshinori MIYAZAKI[*1], Seiji TANI[*2]
*1 Shizuoka University
*2 Tokoha University

The CEFR (Common European Framework of Reference for Languages) has drawn great interests with can-do statements (CDSs), designed to provide assessments of foreign language proficiency, but only limited studies of CEFR for learners of Japanese have been conducted. Some of them worked on the classification of Japanese sample sentences into the corresponding CEFR reading comprehension indices (CDSs), using 3 features: length, document type and technicality. In this presentation, we will add new seven CDSs of Pre-A1 level released in CEFR Companion Volume with New Descriptors in 2017, and carry out experiments for the classification problem. In line with the incorporation of new CDSs, we have extended to 4 features for inference, and used Kanji rate as one of the vital elements in reading comprehension. Further, instead of conventional 7 document types, we divided sentences into 8 types. The results of experiments will be shown in the presentation.

Keywords: CEFR, Classification, Machine learning, Corpus, Can-Do Statement

1. INTRODUCTION

The Common European Framework of Reference for Languages (CEFR) (1) is an international standard for describing second language proficiency developed by the European Council. The framework includes four language skills with six stages of A1 to C2 level, and series of description statements (Can-Do Statements, hereafter called CDS) are described for each level indicating what can be done. Moreover, each CDS can be used together with concrete example sentences for utilization. For instance, a CDS includes “Can understand everyday signs and notices in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings.”, whose example sentence can be “Caution: Do not attempt to leave the train when doors are closing”. Hence, CEFR is an important linguistic assessment to evaluate the language proficiency of learners in testing.

Regarding the utilization of CEFR for Japanese language education, Japanese CEFR-compliant text corpus has not been created, and only limited studies of CEFR for learners of Japanese have been conducted. Takada et al., (2) studied the semi-automatic classification of Japanese example sentences corresponding to reading comprehension indices (CDS) for the creation of Japanese CEFR-compliant text corpus using machine learning. In the study of (2), “technicality”, “length”, and “document type” were chosen as features in CDS classification, but a feature of “document type” was conducted manually. Therefore, Hirakawa et al., (3) conducted a research to automatically estimate “document type” using the method of (2), word2vec and fastText. In addition, (3) worked on the improvement of the accuracy of “technicality”, and carried out the experiment using these estimation results. (2) and (3) covered 27 CDSs except C1, C2 levels at proficient level and a CDS at B2 level which

focuses on vocabulary ability rather than reading comprehension skill.

In 2017, The CEFR Companion Volume with New Descriptors (4) was published to be intended as a complement to the CEFR. The focus in the project was to update the CEFR illustrative descriptors by highlighting certain innovative areas of some CEFR descriptors on the previous version; further development of the CEFR with fulling defining 'plus level' and a new 'Pre-A1' level; responding to demands for description of listening and reading in detail; and enriching the description at A1, and at the C levels. Among these updates, this study will focus on Pre-A1 level, which supports for the beginners of Japanese. We will add new seven CDSs of Pre-A1 level in the Companion Volume with corresponding example sentences. As a result, to cope with the accuracy of classification of new CDSs, we have extended to 4 features for inference, and used Kanji rate as one of the vital elements in reading comprehension. Further, instead of conventional 7 document types, we divided sentences into 8 types.

2. PRE-A1 LEVEL

A scale is defined as the overall proficiency of a foreign language describing in detail about language use and language ability. The scale of CEFR is based on two levels (A1, A2) of "Basic language users", two levels (B1, B2) of "Independent language users", and two levels (C1, C2) of "Proficient language users". However, even at the most fundamental A1 level, proficiency is too high for the beginners of foreign language; as a result, Pre-A1 level before reaching A1 level has been complemented in the CEFR Companion Volume.

Pre-A1 level is a band of proficiency at which the learner has not yet acquired a generative capacity, but relies upon a combination of words and formulaic expressions (4). At this level, learners are the beginners, who do not have a vocabulary structure yet and know the simple words they learned in class. As is appropriate for learners of Pre-A1 level, reading comprehension tasks are supported by pictures. Longer tasks are mainly based on simple stories, so learners should be provided as much opportunity as possible to read and enjoy stories at their level. In addition, Pre-A1 level produces simple utterances, and generally responds at word or phrase but may also produce some longer utterances (The 2018 Pre A1 Starters, A1 Movers and A2 Flyers revisions (5)). To support for these requirements, the CEFR Companion Volume has provided seven descriptors (CDSs) in terms of reading comprehension for learners of Pre-A1 level.

According to the descriptors, learners of Pre-A1 level can recognize familiar words that there are visual supports such as photos and illustrations (for example, menus of fast food restaurants, picture books). On the other hand, learners can understand familiar names and words on the message if necessary. Among the information in cards and e-mails, Pre-A1 level emphasizes the understanding of very simple information such as "date" and "time". Moreover, because this level supports for the language in daily use, it requires learners to understand the very short, simple instructions, signs or notices in everyday contexts, in which simple words and sentences are used. The detail of seven new descriptors is shown in Figure 1.

Fig. 1 Lists of CDSs of Pre-A1 Level

CDS No.	Can-Do Statement
32	Can understand simple everyday signs suchas 'Parking', 'Station', 'Dining room', 'No smoking', etc.
33	Can find information about places, times and prices on posters, flyers and notices.
34	Can understand the simplest informational

	material that consists of familiar words and pictures, such as a fast-food restaurant menu illustrated with photos or an illustrated story formulated in very simple, everyday words.
35	Can understand from a letter, card or email the event to which he/she is being invited and the information given about day, time and location.
36	Can recognize times and places in very simple notes and text messages from friends or colleagues.
37	Can understand very short, simple instructions used in familiar, everyday contexts such as 'No parking', 'No food or drink', etc., especially if there are illustrations.
38	Can recognize familiar words accompanied by pictures, such as a fast-food restaurant menu illustrated with photos or a picture book using familiar vocabulary.

3. CLASSIFICATION FEATURES

In this study, we continue to use the feature of "length" (number of characters, number of words, number of sentences, number of line feeds) which is used in (2). In terms of the feature of "document type", we extend to 8 types instead of conventional 7 types in (2), (3). Regarding "technicality", we still use the estimation method of fastText in (3), which was considered to have the high improvement in accuracy. Finally, a new feature added in the classification is Kanji Rate, which is suggested as one of the elements affecting reading ability at low levels (Uno et al., (7)).

3.1 Document Type

In document type estimation, from ten types used in Takada (2), Hirakawa (3) selected seven document types from CDS because some of the ten types had similar types of documents, and the difference was not clear. Especially, "letters" and "mails" were referred to "communication texts", "signs" and "posters" were integrated as "signs + posters". As a result, seven document types were "article + news", "newspaper articles", "public documents", "signs + posters", "communication statements", "instructions" and "others".

In regard to the descriptors of Pre-A1 level, most of them emphasize the language in daily use with very short, simple words and sentences. From the list of CDSs in Figure 1, it can be known that, example sentences of Pre-A1 level can be seen in every day signs; posters, flyers and notices; materials illustrated by pictures; letters, cards or email, notes or text messages; and instructions. However, these descriptors do not focus on the first three document types used in (3): "article + news", "newspaper articles", "public documents", whose example sentences are longer and required the knowledge in certain fields. Therefore, we will keep these three document types in this study.

On the other hand, as in Figure 1, posters, flyers and notices are one of the main document types used in everyday context, while signs mainly emphasize the recognition of simple and familiar words. Moreover, when collecting the example sentences of Pre-A1 level, most of them belong to type of notices that provides much information about places, times and prices, rather than in posters and flyers. As a result, to improve the accuracy in estimation of document types of example sentences of Pre-A1 level, we divide the last four types used in (3) into five ones. "sign + posters" is divided into "signs" and "notices", in which "notices" includes the example sentences can be seen in posters, flyers, notices, etc.

In addition, example sentences belong to menus, picture books, stories, etc., with visual illustration are listed as "others".

Consequently, there are eight document types used in this study: "article + news", "newspaper articles", "public documents", "signs", "notices", "communication statements", "instructions", and "others".

3.2 Kanji Rate

The most common Japanese writing system is based on the mixture of Kanji and Kana (Hiragana and Katakana). Kana has clear and direct character-to-sound correspondences where each Kana represents Japanese mora. In contrast, Kanji (originally derived from Chinese characters) is commonly used for writing content words – most of nouns, verbs and adjectives are written in Kanji – and Kanji characters, alone or combination with other characters, represent whole words. (Alexandra S. Dylman et al., (8)). Hence, it implies that reading Kanji ability may require different reading strategies or different cognitive skills to acquire reading comprehension of Japanese.

In regard to the applications of learning of foreign language, "readability" is used to define how difficult users evaluate a reading text. Functions which define readability are called readability formula. In the learning of Japanese, there has been many of readability formulas for Japanese. Morioka et al., (9) and Yasumoto et al., (10) proposed formulas using the sentence length measured in letters, and words and the percentage of Kanji characters for estimating the difficulty of the vocabulary. It is said that a text with longer sentences is estimated as difficult, and a text with more Kanji characters is also estimated as difficult. Morioka (9), whose study focused on school textbooks, suggested that upper grade text-books contain longer sentences on the average and more Kanji characters. Yasumoto (10) showed that documents with more Kanji characters are less readable even for adults. Therefore, documents using more Kanji characters are likely to include more different words and demand more reading skills.

On the other hand, the number of Kanji characters which students have to master is specified in each grade affecting the reading comprehension of language learners. According to the curriculum for the school subject "Japanese language" by the Ministry of Education, there are 6 Grades 1-6 of Kanji acquisition. However, the number of Kanji taught in Grade 1 is limited to 76 characters, students in Grade 1 can only achieve the accuracy of around 80% in reading. The accuracy is suggested to improve when learners are taught more Kanji characters for higher grades (Tamaoka et al. (11)). Acknowledged the important role of Kanji characters in reading comprehension, especially at low level, we decided to use Kanji rate as a feature of CDS classification to improve the accuracy of classification in Pre-A1 level.

Moreover, besides the number of Kanji characters, the degree of difficulty of each Kanji characters affects reading ability of learners. The more Kanji characters at low level appear in a reading text, the more readable the text is (12). In this study, we use four lists of Kanji characters of Japan-Language Proficiency Test (JLPT) (13). The four lists are named JLPT level 4, JLPT level 3, JLPT level 2 and JLPT level 1 (14); in which the difficulty rises from level 4 to level 1. In addition, in case a Kanji characters appearing in a reading text does not belong to any of four lists, it will be added to the list named "Other". Consequently, there are five Kanji lists to be estimated in this study.

4. CDS CLASSIFICATION

Regarding the CDS classification experiment, we continue to use " length", "document type" and "technicality" proposed by Takada (2), and a new feature "Kanji rate".

4.1 Technicality

In terms of technicality, we use the estimation result (1 as technical, -1 as non-technical) by fastText. The classifier used to estimate the technicality of example sentences is constructed as proposed in (3). Using this classifier, we assign technicality information to example sentences.

4.2 Length

Estimation of length is conducted the same as the one used by Takada (2). Especially, we use MeCab (15) to extract "number of words", "number of characters", "number of sentences", and "number of line feeds".

4.3 Document Type

The estimation result in 3.1 is used in the CDS classification experiment. When using the estimation result as the input values, it is converted to applicable: 1 or non-applicable: 0, corresponding to a document type with an example sentence as shown in Figure 2.

4.4 Kanji Rate

Regarding the Kanji Rate, we use the result estimated in 3.2 for the classification of corresponding CDS. The input values include the percentage of the number of Kanji characters in the total number characters in an example sentence, the number of Kanji characters of each level 1, 2, 3 ,4, other in an example sentence.

5. EXPERIMENT

5.1 Experiment on estimating document type

The data set used for the experiment on document type estimation includes 1,423 example sentences collected in (3), and 149 new example sentences of level Pre-A1. These example sentences were collected by ten collaborators who are currently teachers of Japanese foreign language (eight of them are Vietnamese, two of them are Japanese). In total, there are 1,572 example sentences in this experiment.

fastText (16) is one of the methods of text classification library for word representations and sentence classification. From (3), fastText was considered as the most effective methods to estimate document type. Therefore, we use fastText with the same parameters proposed in (3) to carry the experiment. The result is shown in Figure 2.

Fig. 2 Results of Document Type Estimation

Document Type	No. of example sentences	No. of truly predicted	Rate
article + news	239	178	74.48%
newspaper articles	214	178	83.18%
public documents	200	181	90.50%
signs	245	165	67.35%
notices	55	34	61.82%
communication statements	231	192	83.12%
instructions	211	173	81.99%
others	177	117	66.10%
TOTAL	1,572	1,218	77.48%

From Figure 2, there are five types of document types that can be estimated with the correct answer rate of 70% or more. Three other document types can also be estimated with the result more than 60%. The document type with the lowest percentage is “notices”, which is considered to be difficult to estimate because the small number of example sentences was used. In average, the experiment got the result of 77.48%.

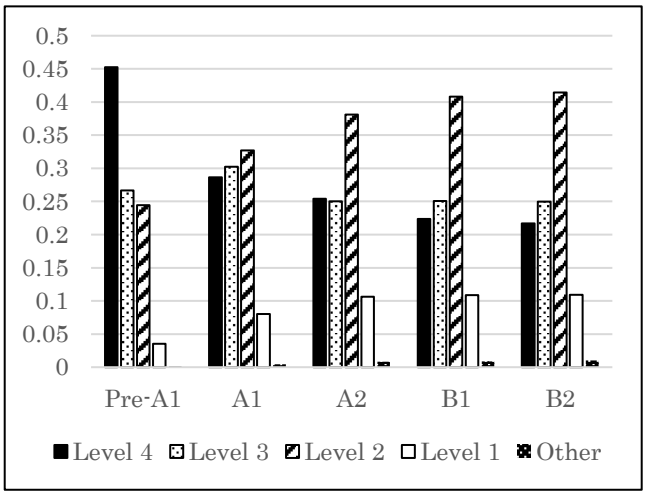
5.2 Experiment on estimating Kanji rate

In this experiment, we use 370 example sentences used in (3) and 149 example sentences of Pre-A1 level collecting this time as mentioned in 5.1; in total there are 519 example sentences. Machine learning is used to carry the experiment with the Kanji rates calculated for each level.

Figure 3 shows the percentage of Kanji rate of example sentences in five levels of CEFR. It can be seen that the percentage of Kanji rate in level 4 belonging to Pre-A1 level is the highest. It implies that example sentences of Pre-A1 level comprises of the easiest Kanji characters, in comparison to four

other levels. Moreover, the number of Kanji characters in level 4 is decreasing from Pre-A1 level to B2 level. It is a good result to be considered as the input value for CDS classification, especially in the example sentences of Pre-A1 level.

Fig. 3 Results of Kanji Rate in each CEFR Level



5.3 Experiment on CDS classification

For classification experiments, as in (2), we assumed multi-label classification corresponding to multiple CDSs in an example sentence. For the data set of the experiment, we used 519 example sentences with multi-label information collected from 10 experienced Japanese educators with knowledge in CEFR. The average number of CDSs corresponding to one example sentence is about 2.82. 34 binary classifiers (SVM) are used for the multi-label classification method, and the cross validation with three divisions is performed for the evaluation. As a result, the overall average accuracy was about 70.27%.

Fig. 4 Overall Results of CDS classification

	Recall	Precision	F-value
Positive	70.29%	38.39%	49.66%
Negative	89.82%	97.10%	93.32%

As in Figure 4, the classification results for all example sentences were about 70.29% positive recall, about 38.39% positive precision, about 49.66% positive F-value, about 89.82% negative recall, about 97.10% negative precision, and the negative F-value was about 93.32%. Although it was confirmed that the positive recall, negative recall and precision ratio were maintained at a relatively high level, the accuracy of the positive precision rate was quite low. The positive average predicted number for one example sentence was about 5.15.

ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid (KAKENHI) for "Scientific Research (C) (18K00722)."

REFERENCES

(1) Council of Europe: "Common European Framework of Reference for Languages: Learning , Teaching Assessment", Cambridge University Press (2001)

(2) 高田宏輝, 宮崎佳典, 谷誠司: "韓国人日本語学習者のための CEFR 読解指標に基づく例文分類", 韓國日本學會第 94 回國際學術大會, pp. 299-303 (2017)

(3) 平川遼汰, 宮崎佳典, 谷誠司 : "日本語例文自動分類による CEFR 読解指標推定支援 Web アプリケーションの開発", 情報処理学会全国大会, pp. (4)-635-636 (2018)

(4) Council of Europe: "Common European Framework of Reference for Languages: Learning , Teaching. Companion Volume with New Descriptors" (2017)

(5) Cambridge English Assessment: "The 2018 Pre A1 Starters, A1 Movers and A2 Flyers revisions" (2018)

(6) Global Test of English Communication: "GTEC スコアと CEFR レベル関連付け調査報造－Pre-A1/A1 レベル追加調査" (2018)

(7) Akira Uno, Taeko N. Wydell, Noriko Haruhara, Masato Kaneko, Naoko Shinya: "Relationship between reading/writing skills and cognitive abilities among Japanese primary-school children: normal readers versus poor readers (dyslexics)", Reading and Writing, Vol.22, No.7, pp. 755-789 (2009)

(8) Alexandra S. Dylman, Mariko Kikutani: "The role of semantic processing in reading Japanese orthographies: an investigation using a script-switch paradigm", Reading and Writing, Vol.31, No.7, pp. 503-531 (2018)

(9) 森岡健二: "ことばの教育" ,明治書院 (1988)

(10) 安本美典 : "説得の文章技術", 講談社現代新書 (1983)

(11) Katsuo Tamaoka: "A Japanese Perspective on

Literacy and Biliteracy: A National Paper of Japan", The Reading Research Symposium (1996)

(12) Komori, Saeko: "A Study of Kanji Word Recognition Process for Japanese as a Second Language", Tokyo: Kazama Shobo (2009)

(13) "JLPT Japanese-Language Proficiency Test", http:///www.jlpt.jp/ (Reference date: 2018.6.20)

(14) http://kanjicards.org/kanji-lists.html (Reference date: 2018.6.20)

(15) "MeCab", http://taku910.github.io/mecab/ (Reference date: 2018.6.19)

(16) "fastText", https://fasttext.cc/docs/en/supervised-tutorial.html#content (Reference date: 2018.6.19)