Learner Corpus Studies in Asia and the World, vol. 5 (Papers from LCSAW2020) ISSN: 2435-2632
Published by School of Language & Communication, Kobe University, Japan

Quantitative Analysis of JFL Learners' Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency

Jae-Ho LEE
Waseda University

Yoichiro HASEBE
Doshisha University

Abstract

We developed jWriter (https://jreadability.net/jwriter), a computer-based writing-evaluation system for learners and teachers of Japanese as a foreign language (JFL). The system has two primary functions: 1) to analyze the input essay in Japanese and output an estimated proficiency level and 2) to provide advisory comments, or diagnostic analysis, about the variety and usage of the expressions in the essay. With this system, learners can work to improve their writing skills while obtaining feedback, albeit informal, in real-time. jWriter is built upon a regression model constructed using data from the I-JAS (International Corpus of Japanese as a Second Language), and the accuracy of this discriminant formula is 76%.

Keywords

Writing Proficiency, Quantitative Analysis, Automated Evaluation, Multiple Regression, Learner Corpora

1. Introduction

Improving skills in writing, one of the four aspects to language, is essential for learners to go to the next level of language learning. However, improving one's writing is a rather "lonely" practice: unlike when practicing conversation, the learner sits still while working in a notebook or on a computer screen without a partner. They just "write." In fact, this is the case for instructors, to some extent. Evaluation of text submitted by learners is a lonely practice for teachers, as they do not have anything around them that can assist with the work and can only rely on their own linguistic knowledge and intuition. Thus, it has been difficult to standardize the evaluation of learners' writing. This is not ideal for students or teachers.

With the hope of making a breakthrough in this situation—even a tiny one—we developed jWriter, a computer-based writing-evaluation system for learners and teachers of Japanese as a foreign language (JFL). The system has two primary functions: 1) to analyze the input essay in Japanese and output an estimated proficiency level and 2) to provide advisory comments, or diagnostic analysis, about the variety and usage of the expressions in the essay. With this system, learners can work to improve their writing skills while obtaining feedback, albeit informal, in real-time. Also, teachers can use the system to obtain guidelines and/or a rough sketch of the students' work, with which they can fine-tune their evaluation and prepare detailed advisory comments.

jWriter is built upon a regression model constructed using data from the International Corpus of Japanese as a Second Language (I-JAS Corpus, see Sakoda et al. (Eds.), 2020). This paper presents details of the quantitative linguistic analysis we conducted to construct the regression model and then explicates upon the interfaces and workings of the jWriter system.

2. Background

In this section, the theoretical/conceptual background of our project is presented. It consists of three parts that respectively expound on ideas of educational text mining, quantitative linguistic analysis, and performance evaluation.

Nowadays, teaching practices are being significantly improved by utilizing a large volume of data collected from various types of actual education settings, such as class activities, take-home assignments, and communicative interactions between students and teachers outside class. Romero & Ventura (2012) use the term "educational data mining" to refer to a framework for resolving various challenges that educators confront in their teaching. "Mining" is carried out in a number of areas to extract regularities and/or associations embedded within a large volume of data (Adriaans & Zantinge, 1996). Educational data mining, a specialized type of data mining, is conducted to benefit both educators and learners by unearthing facts and ideas that are applicable to the practices of educational activities. jWriter, which utilizes the results of analyzing a large volume of data from the I-JAS corpus, is one such application based on this trend of educational text mining.

The concepts and methods of quantitative linguistic analysis are also central to our project. In quantitative linguistics, text is transformed into numbers, and the numbers are evaluated for investigating various aspects of language (Lee, 2017). Data, in a sense, must be quantitatively evaluable, and since strings of text printed on paper or computer screens are not quantitatively evaluable, they need to be first transformed into a set of numbers for them to be treated as data (Toyoda, 2008). Such data, then, can be treated as if they were "fingerprints." People's fingerprints are all different, and a set of various text features also differ. Properties such as the average length of sentences, variety of vocabulary items, and overall frequency of particular types of expressions are all different. Hence, they can be collectively used to characterize a given set of text. jWriter identifies such fingerprints from input text and evaluates them on the basis of knowledge extracted from the I-JAS corpus.

Also worth mentioning here is that we developed jWriter with a perspective of performance evaluation. Performance evaluations are various attempts to grade learners' performances in activities such as oral interviews or essay writings as objectively as possible (Bachman, 1990; Green, 2014). Traditionally, many educational institutions have used what are often called objective tests, which typically include multiple-choice questions, to evaluate learner's linguistic abilities. As Kondo-Brown (2012) and Lee (2015) point out, deeper understandings of learners' abilities require detailed analysis of spoken and written text produced by learners themselves. Spoken and written texts, however, are difficult to evaluate objectively, and they have been difficult to deal with using computational systems. In fact, this was one of the reasons why institutions have long relied on objective tests instead.

We developed jWriter to change this tradition in Japanese writing education. In the next two sections, we present how we built a regression formula to evaluate Japanese text and produce output that can be used as a reliable index to estimate the level of the learners who authored the text.

3. Data and Methods

The I-JAS corpus, which we used to construct the regression model, is composed of learners' essays authored under the following conditions:

- Topic of essay: food preferences—fast food versus home-cooked food
- Quantity of essay: approximately 600 words
- Authoring environment: consulting dictionaries and online references was allowed; there was no time limit; asking for help from teachers or other people was prohibited.

The nationalities and native languages of the learners who authored the essays in the I-JAS corpus are diverse, making this corpus well balanced in terms of the participants' linguistic backgrounds.

One of the most important characteristics of the essays in the I-JAS corpus is that they contain Simple Performance-Oriented Test (SPOT) scores for the essays' authors, so that researchers can investigate associations between text features of the essay texts and certain levels of proficiency in learning Japanese (Kobayashi, 2015). In the development of jWriter, we used 373 essays and grouped them into three levels (elementary, intermediate, and advanced) according to the SPOT scores associated with them. The results are shown in Table 1.

Table 1

Distribution of I-JAS essays in terms of learners' proficiency and native language

	Number of Essays			Average Number of Characters		
	Elementary	Intermediate	Advanced	Elementary	Intermediate	Advanced
Chinese	3	23	56	622.3	638.9	647.9
Korean	0	11	53	—	561.9	602.6
Spanish	23	9	1	569.7	635.3	533.0
Indonesian	19	14	0	522.5	614.6	—
Vietnamese	14	16	1	558.8	642.4	989.0
English	16	9	3	542.9	631.3	687.3
German	4	17	3	544.0	614.8	604.7
Russian	8	9	5	479.9	674.2	701.4
Hungarian	2	12	7	408.0	626.3	641.7
Thai	3	12	1	606.0	618.8	722.0
French	7	3	0	473.7	553.0	—
Turkish	4	4	1	488.0	617.0	644.0
Sum	103	139	131	—	—	—

Table 1 also shows that the native language distribution of the learners contributed to the I-JAS corpus. Of the total 373 learners, speakers of Chinese and Korean accounted for the largest portion, at 82 and 64 learners, respectively.[1] These are followed by speakers of Spanish (33), Indonesian (33), Vietnamese (31), English (28), German (24), Russian (22), Hungarian (21), Thai (16), French (10), and Turkish (9). It is worth noting here that learners at more advanced levels wrote more characters than those at less advanced levels, as observed in the figures in the columns "elementary," "intermediate," and "advanced."

We then identified various text features using computational text-processing tools to do statistical analysis for determining associations between the text features and the SPOT scores. This procedure was conducted as follows. First we input the text of each of the essays into the morphological analyzer program MeCab to segmentize the text into words.[2] Next, we obtained the following text features from each of the essays: 1) the total number of word tokens, 2) the total number of word types, 3) the frequency distribution of words of different origins (*wago* [Japanese], *kango* [Chinese], or *gairaigo* [Western]), and the frequency distribution of parts of speech (12 parts in total). Then, we also obtained the frequency distribution of different levels of words according to the Japanese Educational Vocabulary (JEV) word list for each of the essays.[3] The JEV consists of 17,920 entry words grouped into six levels ("lower elementary," "upper elementary," "lower intermediate," "upper intermediate," "lower advanced," and "upper advanced") (Sunakawa et al., 2012).

As a result, we had a learner corpus of 373 essays grouped into three levels ("elementary," "intermediate," and "advanced") along with data on various text features, such as the number of words and the variety of the vocabulary items used. We then applied statistical analyses to this dataset using both the descriptive method and multiple linear regression. The results of the statistical analyses are presented in the next section.

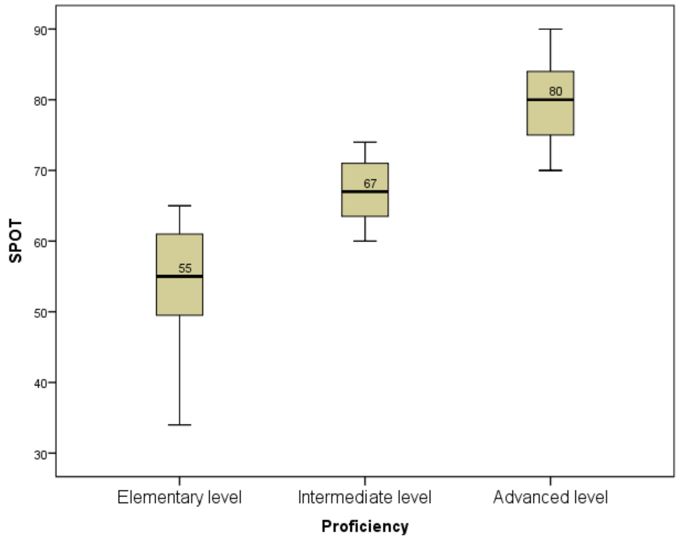
4. Results and Analysis

4.1 Descriptive Analysis

First, we present the distribution of the SPOT scores among the three groups of proficiency levels in Figure 1. Although the elementary-level essays show a wider range on the SPOT score scale than the other two groups, the three groups are placed with fairly balanced distances, with mean scores of 55 (elementary), 67, (intermediate), and 80 (advanced).

Figure 1

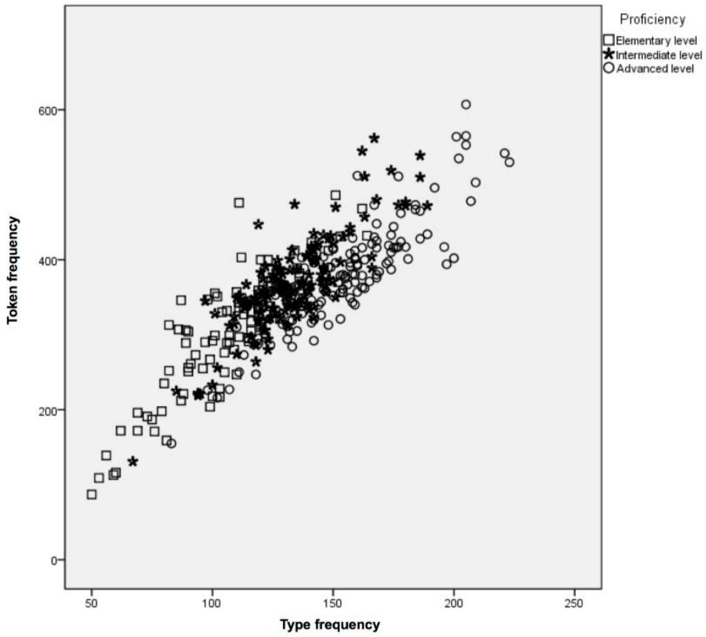
SPOT scores and proficiency levels



Next, the scatter plot in Figure 2 shows the distribution of the essays from the three groups in terms of their token frequencies and type frequencies. The token frequency and type frequency are positively correlated in the corpus, and the essays from more advanced learners tend to show larger token frequencies and type frequencies.

Figure 2

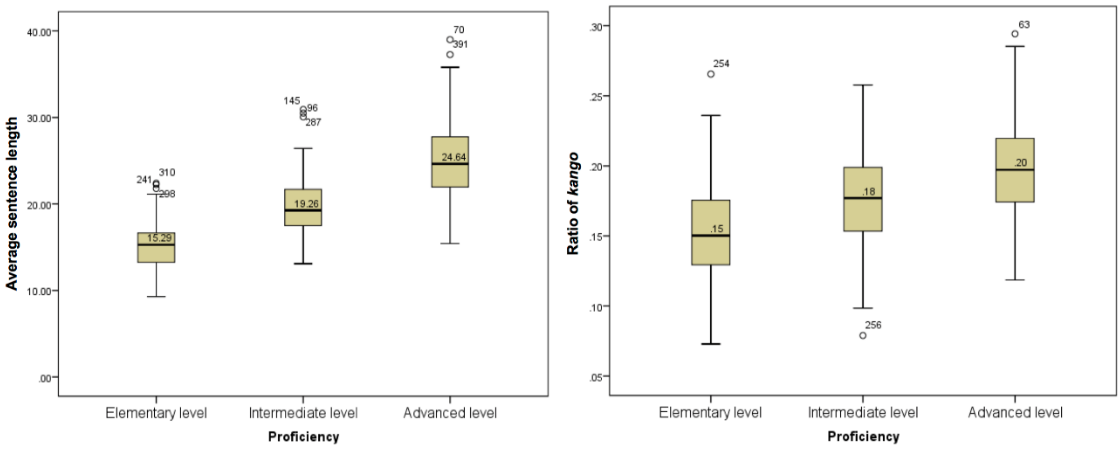
Token frequencies and type frequencies of the essays from the different level groups



However, some essays belong to one level group but show type-token frequencies that are comparable to the essays from other level groups. Thus, the type-token frequency alone is not a reliable index of the proficiency of the learners' essays. We need to turn to other text features to build a robust model of proficiency-level estimation by essay text. Then, we investigated the correlation between proficiency (based on the SPOT scores) and the average length of sentences (i.e., the mean number of words/morphemes per sentence) as well as the correlation between proficiency and the ratio of *kango* (i.e., words of Chinese origin) to the total number of words in the essay. These are presented in Figures 3(a) and 3(b), respectively.

Figure 3 (a): Correlation between proficiency and average sentence length

(b): Correlation between proficiency and ratio of *kango*



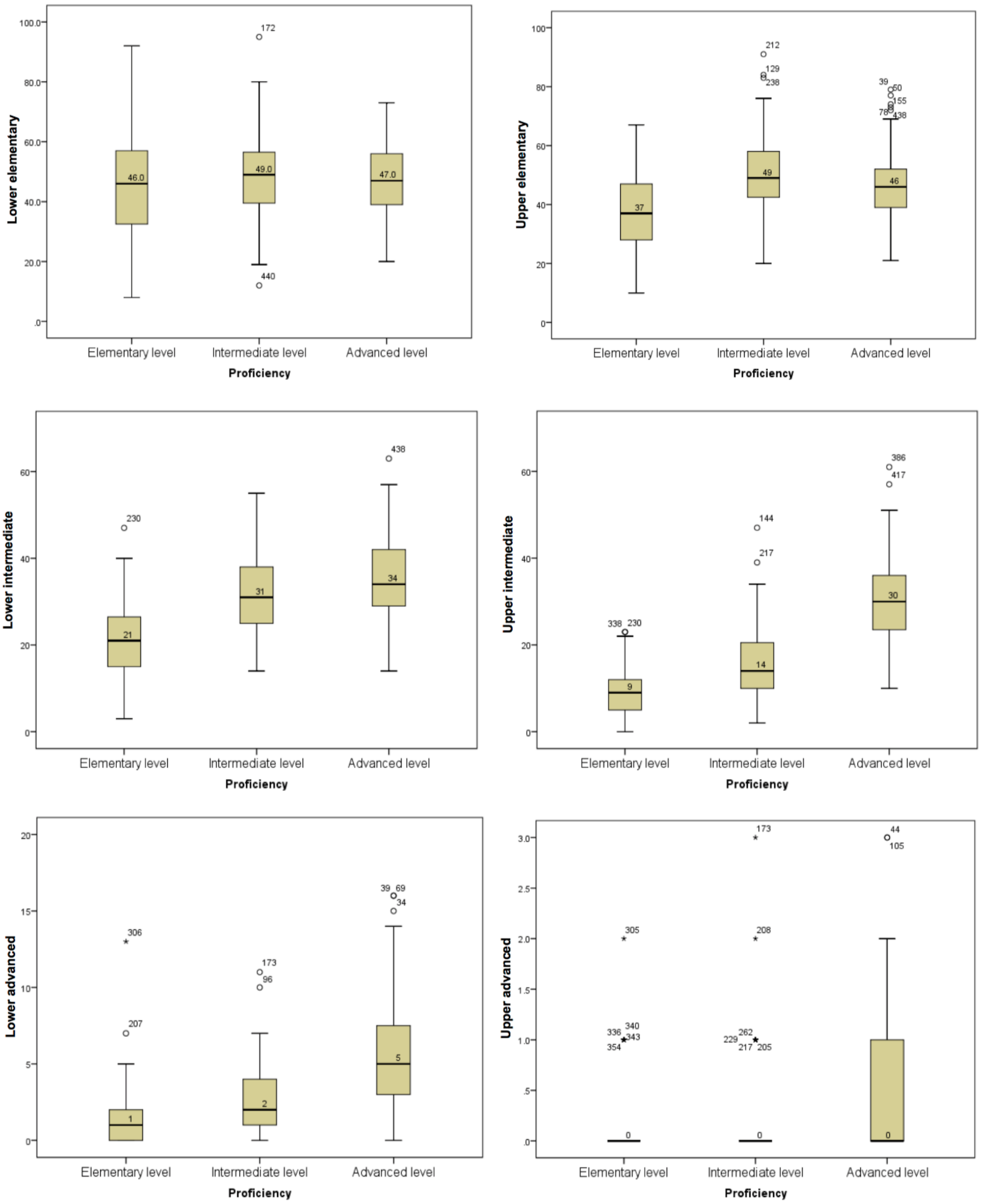
(a) (b)

From the results shown in Figure 3, it can be said that elementary-level learners used 15 words per sentence, on average, and 15% of their words were *kango*. Intermediate-level learners write sentences with 19 words, on average, and 18% of them were *kango*; advanced-level learners wrote sentences with 24 words, on average, and 20% of them were *kango*. It is almost obvious that the more advanced the learners are, the longer their sentences become and the higher the rate of *kango* in their text grows.

To further explore the possibility of finding good indices with which to distinguish different levels of proficiency in learners' essay texts, we also turned to the relations between the SPOT scores and use of different levels of vocabulary items. As mentioned earlier, we analyzed the text of the essays and calculated the distribution of different levels of vocabulary items in them using the JEV word list. Figure 4 presents the differences in rates of words corresponding to the six levels (from "lower elementary" to "upper advanced") used by learners of three proficiency levels ("elementary," "intermediate," and "advanced").

Figure 4

Relation between vocabulary levels and proficiency levels



Though we should not say anything overly definite based on Figure 4 alone, it can be observed that 1) learners of all levels widely used lower-intermediate-level words (more than 40 times in an essay, on average); 2) learners of all levels rarely used lower-/upper-advanced words, if at all; and 3) elementary learners used lower-intermediate-level words less frequently than other groups of learners, and advanced learners used upper-intermediate-level words more frequently than other groups of learners.

These facts presumably show that weighing the use of lower-elementary-level words does not highly contribute to constructing a good estimation model because those basic words are included in abundance in texts at all levels. In an opposite fashion, learners at all three levels rarely used lower-/upper-advanced-level words, making the frequency of these advanced words not very useful either. What is more promising is looking at the frequency of lower-/upper-intermediate-level words. With this in mind, we conducted a multiple regression analysis on the I-JAS dataset to verify our expectation and constructed a robust regression model to estimate the author's proficiency level for a given essay.

4.2 Multiple Regression Analysis and Its Results

We conducted a step-wise multiple regression analysis on the I-JAS learner corpus data using IBM SPSS (Ver. 24) and obtained eight models with different choices of variables and weights. Among the eight models, the two models presented below shewed especially high coefficients of determination, which are generally used as the index of a model's estimation accuracy.

Model A:
Estimated proficiency level = 1.592 + average sentence length × 0.046 + number of upper-intermediate words × 0.026 + type-token ratio × −0.416 + number of verbs × 0.014 + number of lower-intermediate words × 0.015 + total number of characters × −-0.004 + total number of *wago* × 0.006 ($R^2 = 0.755$)

Model B:
Estimated proficiency level = 1.637 + average sentence length × 0.045 + number of upper-intermediate words × 0.021 + type-token ratio ×−0.430 + number of verbs× 0.015 + number of lower-intermediate words × 0.011 + total number of characters × −0.004 + number of *wago* ×0.007 + number of *kango* × 0.007 ($R^2 = 0.760$)

Model A and Model B share many variables (the average sentence length, the number of upper-intermediate words, the type-token ratio, the number of verbs, the number of upper-intermediate words, the total number of characters, and the number of *wago*); the variable for the number of *kango* is the only variable used in just Model B. The coefficients of determination (R^2) of the two models are 0.755 and 0.760 respectively. The numbers are both high, but it is not very clear how well these models estimate text at particular levels (elementary, intermediate, and advanced). To examine this, we applied both models to text from the three levels. The results are presented in Table 2 (Model A) and Table 3 (Model B).

Table 2

Results of estimation test with Model A

		Estimated levels			
		NA(L)*	Elementary	Intermediate	Advanced
Actual levels	Elementary	9	88	6	
	Intermediate	1	56	82	
	Advanced			85	46
	Sum	10	144	173	46

Note: * NA(L) refers to cases with values that exceed the lower end of the model's expected value range

Table 3

Results of estimation test with Model B

		Estimated levels				
		NA(L)*	Elementary	Intermediate	Advanced	NA(H)**
Actual levels	Elementary	1	68	34		
	Intermediate		6	124	9	
	Advanced			18	103	10
	Sum	1	74	176	112	10

Note: * NA(L) refers to cases with values that exceed the lower end of the model's expected value range
** NA(H) refers to cases with values that exceed the higher end of the model's expected value range

The numbers in the tables' gray cells represent the essays correctly estimated by our models. Looking at the estimations of the elementary level, Models A shows a better score than Model B, with 88 essays correctly estimated by the former (85 percent recall) and 68 by the latter (66 percent recall). Looking at the figures for the other levels, however, a different picture emerges: for the intermediate level, Model A only attains 59 percent recall, while Model B presents 89 percent recall; for the advanced level, Model A presents 35 percent recall, while Model B reaches 85 percent recall.

In sum, Model A is a good estimator at evaluating elementary-level essays, whereas

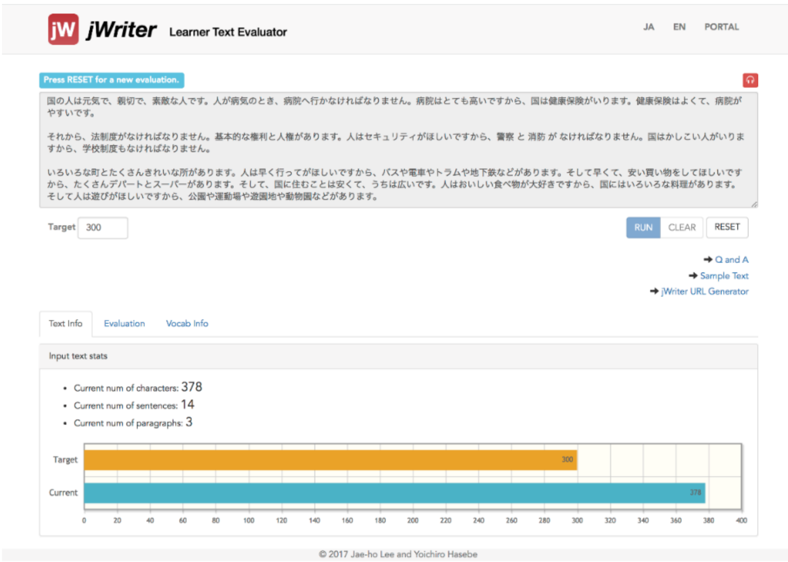
Model B is a better estimator for intermediate and advanced essays. It is speculated that the difference between the two models arises partially because Model B includes the number of *kango* in the model, while Model A does not. Which of the two is a better estimator then? We concluded that Model B is more appropriate for our purpose of developing a computational system to evaluate learners' essays in Japanese. This is because essay writing is more important in the JFL classroom for intermediate and advanced learners. Elementary-level learners do practice writing, but it tends to be sentence-level or single-paragraph writing. Then, practically, it makes more sense to choose a model that is better at estimating intermediate- and advanced-level text than elementary-level text.

5. Developing a Computational Evaluation System

Using Model B, we developed jWriter, an online computational system that estimates the Japanese proficiency level of the author of an input essay text.[4] Figure 5 is a screenshot of the system's input panel.

Figure 5

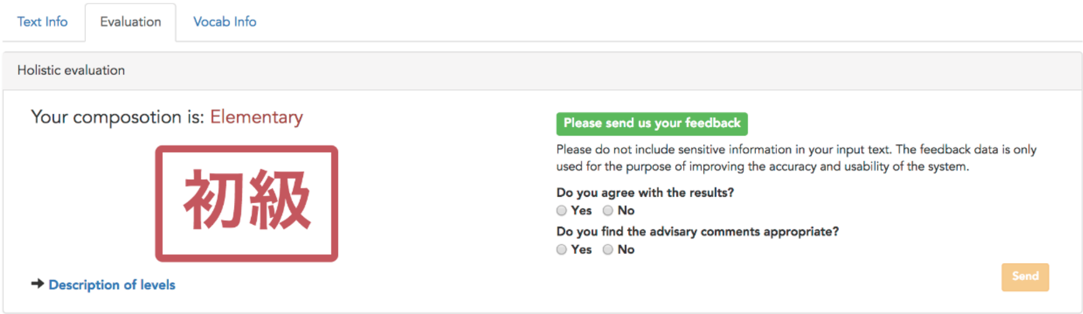
Input panel of jWriter



The jWriter system does two things: it 1) analyzes and presents the user with the estimated proficiency level of the author of the essay submitted via the online interface, and 2) provides the user with advice generated according to the essay's proficiency level and text features. In Section 3, we described how the data from the I-JAS corpus was processed using the morphological analyzer program. The essay text submitted to jWriter is first processed in the very same fashion and using exactly the same toolset as in the process described in Section 3. Then, the regression model (Model B) is applied to the resulting text features, such as the number of words of a particular level, the number of verbs, and the type/token ratio, producing a single numeral value, which jWriter in turn translates into a corresponding label, i.e., "elementary," "intermediate," or "advanced." Figure 6 is a screenshot of jWriter presenting the estimated proficiency level.

Figure 6

jWriter presenting the estimated proficiency level



This three-level estimation is useful to quickly check the learner's writing proficiency. It does not give much practical information, though, for learners to improve their skills. Thus, we implemented an advice-generation functionality for jWriter that allows the user to see particular points with which the text's author presumably has trouble or, rather, excels. Figure 7 shows an example set of advice, or "diagnostic analysis," generated by jWriter.

Figure 7

jWriter's diagnostic analysis

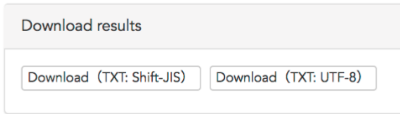
As shown in Figure 7, jWriter uses four types of text features from the input essay and presents their relative distance from the average values among essays at the same level contained in the corpus. The four features are: 1) the type/token ratio, referred to as "word variation"; 2) the number of *kango*, or "*kango* frequency"; 3) the average sentence length, or "sentence complexity"; and 4) the number of intermediate-level words, referred to as "high-level words frequency."

Diagnostic analysis is helpful to show the reasoning behind jWriter's evaluation of the learner's essay. As mentioned in Sections 3 and 4, we analyzed 373 essay texts from the I-JAS corpus and constructed our regression model to estimate the proficiency level of a given input essay. While the estimation presented by jWriter has only three levels ("elementary," "intermediate," "and advanced"), the essays' text features can take diverse values. It is therefore helpful to be able to see whether the values of certain text features exceed or lag behind the average of the essays at the same level.

In addition to these two main functionalities, jWriter offers a downloadable text file containing features extracted from the input essay (obtained rather contingently via morphological analysis to conduct evaluation): included are the values for 37 text features, the estimated level (both as a numeral value and as a text label), and the original essay text. This functionality, illustrated in Figure 8, will be especially helpful for researchers who intend to investigate learners' data more closely and/or to use the text features as inputs for further statistical analyses, for instance.

Figure 8

Downloaded file of text features



感動詞	0	普通名詞	53	上級後半語	0	総文字数	386
形状詞	5	連体詞	0	その他の品詞	33	内容語	73
形容詞	17	総文数	14	ひらがな	241	名詞	53
助詞	64	平均語数	16. 29	カタカナ	16	述べ語数	228
助動詞	33	総形態素数	228	漢字	89	異なり語数	86
接続詞	3	初級前半語	34	和語	158	多様性	0. 377193
代名詞	1	初級後半語	32	漢語	34	漢語力	0. 1708543
動詞	3	中級前半語	12	外来語	5	レベル付語数	88
副詞	3	中級後半語	10	混種語	2	難解語使用率	0. 1136364
固有名詞	0	上級前半語	0	定型句	0	リーダビリティ・レベル	4. 6
						リーダビリティ・ガイドライン	初級後半
						作文評価・レベル	1. 143256
						作文評価・ガイドライン	初級

6. Conclusion

This paper presented a study using the I-JAS corpus to develop a computational system for innovative writing education for JFL learners. First, we identified text features that are characteristic of the essay texts of learners with different proficiency levels by analyzing the relations between various text features and the learners' SPOT scores. Then, we developed jWriter, a computational evaluator of essay text written by learners, using the regression model constructed as a result of analysis of the I-JAS data.

As mentioned in the first section of this paper, JFL educators have had difficulties with evaluating essays with objectivity, especially when the number of texts is large. We believe that our attempt to develop jWriter shows that the situation has started changing and that future efforts in this line of research and development are very promising.

Acknowledgments

The present paper is a modified and extended version of our earlier report presented in Japanese (Lee at al., 2019). This work was supported by JSPS KAKENHI Grant Number 16K02794, 19H01273, 19K21637.

Notes

[1] The learner essays in the I-JAS corpus were all written by different individuals.

[2] MeCab can be used with different dictionary datasets that suit the user's purposes. We

used MeCab (https://github.com/jordwest/mecab-docs-en) with the Japanese dictionary dataset UniDic Ver. 2.1.2 (http://unidic.ninjal.ac.jp).

[3] Here, the term "level" refers to that of the JEV word list. In the present paper, "level" is used in two ways: the level specified in the JEV word list and the learner's level of proficiency.

[4] jWriter is publicly available online at https://jreadability.net/jwriter/en .

References

Adriaans, P. & Zantinge, D. (1996). *Data mining*. Addison-Wesley.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Green, A. (2014). *Exploring language assessment and testing*. Routledge.

Kobayashi, N. (2015). SPOT. In Lee, J.-H. (Ed.), *Nihongo kyooiku no tame no gengo tesuto gaidobukku* [Guidebook of language testing for Japanese language education] (pp. 110-126). Kurosio.

Kondo-Brown, K. (2012). *Nihongo kyooshi no tame no hyooka nyuumon* [Introduction to assessment for Japanese language teachers]. Kurosio.

Lee, J.-H. (Ed.) (2015). *Nihongo kyooiku no tame no gengo tesuto gaidobukku* [Guidebook of language testing for Japanese language education]. Kurosio.

Lee, J.-H. (Ed.) (2017). *Bunshoo o kagaku suru* [The science of measurement and evaluation of text]. Hituzi.

Lee, J.-H., Hasebe, Y., & Murata, Y. (2019). Automatic determination of learner's writing proficiency and development of a web system. In Lee, J.-H. (Ed.), *ICT×nihongo kyooiku* [ICT×Japanese Language Education] (pp. 38-53). Hituzi.

Romero, C. & Ventura, S. (2012). Data mining in education. *WIREs: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Sakoda, K., Ishikawa, S., & Lee, J-H. (Eds.) (2020). *Nihongo-gakushuusha koopasu I-JAS nyuumon* [Introduction to the Japanese Learner's Corpus I-JAS]. Kurosio.

Sunakawa, Y., Lee, J.-H., & Takahara, M. (2012). The construction of a database to support the compilation of Japanese learners' dictionaries. *Acta Linguistica Asiatica*, *2*(2), 97-115.

Toyoda, H. (2008). *Deeta mainingu nyuumon* [Introduction to data mining]. Tokyo Tosho.