



Automatic Essay Scoring Systems: Can We Use Them in a Japanese Language Learning Environment?

JUN IMAKI & SHUNICHI ISHIHARA

Faculty of Asian Studies, The Australian National University, ACT 0200 Australia

ABSTRACT

This study aims to empirically test whether the automated essay scoring (AES) systems—which were designed and developed for Japanese essays written by native speakers, can be used for Japanese compositions written by non-native speakers. If the current ready-online systems are reliable enough to evaluate Japanese language learners' compositions, and are comparable with human raters, these systems could be used in a Japanese language learning environment. Therefore AES may be able to rectify some time and labour constraints in many teaching institutions, as currently many teachers cannot assign writing assignments as often as they would like to. In order to find the possibility of using the AES systems, the Japanese AES systems *Jess* and *Moririn*, as well as nine human raters, evaluated 50 Japanese compositions written by non-native speakers and the scores from the human raters were statistically compared with those from the AES. All scores were normalised by means of the z-score normalisation technique and the correlations between human raters and AES systems were analysed using Spearman's rank correlation test.

Keywords: teaching Japanese, automated essay scoring, evaluation, learner's composition, rater differences

1 Introduction

Writing assessments in language classes is said to be one of the most effective ways to improve learners' language skill as writing employs both linguistic skill and communicative meaning-making act [1]. Furthermore, writing is a great measure of learners' writing ability. However, language teachers share the dilemma that due to the continuous growth of class sizes and expectations to provide more meticulous evaluations rather than multiple choice quizzes. Therefore, AES could be an economically feasible alternative to using human raters for marking learners' written assignments.

AES systems are desirable for use in language classes for the following three reasons [2]:

1. **Practicality:** human essay grading is very time-consuming for teachers
2. **Consistency:** human essay grading is subjective in nature and consistency may sometimes suffer
3. **Feedback:** providing feedback to a learner is important, and AES can provide prompt feedback with specific suggestions.

2 Previous Literature

AES has only existed for about 40 years, and is relatively young area of study.

2.1 AES for first language (L1) writing

2.1.1 AES for English essays written by native speakers

In 1966, Ellis Page developed a computer based grading system named Project Essay Grader (PEG) and it is regarded as a pioneer of AES. During the 1990s, due to the rapid improvement of computing technology and the progress of the field of natural language processing, several major AES systems were designed.

Numerous studies have been conducted to compare correlations between AES systems and human raters. The result of selected studies are summarised in Table 1.

System	Sample size	Human-Human correlation	Human-AES correlation
PEG	300	0.65	0.74
IEA	285	0.88	0.85
E-rater	270	0.69	0.75
IntelliMetric	102	0.84	0.78-0.85

Table 1: Selected researches comparing a correlation between human raters and that between human raters and AES systems.

Results from Table 1 indicate AES systems for L1 English generally have very high and statistically significant correlation rates with expert human raters.

2.1.2 AES for Japanese essays written by native speakers

After successful reliability experiments in the US, several AES were designed in Japan from the 1980s for the Japanese language. Two main Japanese AES used in this research are *Moririn* and *Jess*.

In *Moririn*, essays are seen as having four aspects [3]. These four aspects are:

1. **Surface form:** which includes the structure, expression, topic and theme
2. **Interior form:** which is the aesthetic aspect of expressions
3. **Past content:** which is vocabulary as a raw material, and
4. **Future content:** which includes originality and impression.

Moririn can evaluate the second and the third aspects of learners' writing. In *Moririn*, raw material vocabulary refers to vocabulary which has a higher correlation with content depth. *Moririn* has a correlation of 0.86 with expert raters' marking as shown in Table 2, which contains the results of reliability experiments of *Moririn* and *Jess*.

System	Sample size	Correlation of AES and expert raters	Correlation amongst human raters
Moririn	35	0.86	N/A
Jess	143	0.83	0.70
Jess	500	0.57	0.48

Table 2: Summary of AES performance.

In *Jess*, essays are evaluated from the following three aspects [4].

1. **Rhetoric:** which includes the ease of reading (such as median and maximum sentence length, diversity in vocabulary, percentage of long words, and percentage of passive sentences),
2. **Organisation:** which includes the usage of conjugation relationships such as forward connection words and reverse connection words, and
3. **Content:** which includes relevant information and precise or specialised vocabulary.

Unlike all other AES, including both English and Japanese system, *Jess* is the only AES which uses essays written by professional writers for system training [5]. In comparison, other AES systems use essays that have been evaluated by expert human raters.

It can be stated that *Jess* has a high enough correlation with human raters to be used to evaluate essays. In one of the reliability experiments, the correlation between *Jess* and expert raters was 0.83 when evaluating 143 essays written by Japanese university students. In another experiment the average correlation was 0.57 between human raters and *Jess* when evaluating 500 essays whereas the correlation among five human raters was 0.48 [5].

2.2 AES for the second language (L2) writing

2.2.1 AES for English compositions written by non-native speakers

In the field of second language learning, even though none of the major AES were originally designed for evaluating L2 compositions, many of the AES were used for English as a second language (ESL) [6]. In ESL classes, Jacobs' five criteria [7] are generally used as a base of evaluation measurement.

In Jacobs' criteria, ESL compositions are evaluated according to:

1. **Content:** which can be evaluated in terms of knowledge, substantiveness, relevance to an assigned topic and the development of the topic
2. **Organisation:** which can be evaluated in terms of fluency of expression, clear statement of the composition, cohesiveness and the logic of sequence
3. **Vocabulary:** which can be evaluated in terms of the effectiveness in the choice of words and idioms, sophistication, mastery of word form and appropriateness of register
4. **Language use:** which can be evaluated in terms of the effectiveness of complex constructions, agreement of tense and number, articles and pronouns, and
5. **Mechanics:** which can be evaluated in terms of conventions, errors, punctuation and phrasing.

The currently used AES systems have close evaluation items to Jacobs' criteria, therefore this can be one of the rationales to why AES could be used in L2 classes.

In Burstein & Chodorow's experiment [8], three expert human raters and E-rater, evaluated 255 English compositions written by non-native speakers on six-point scale, with the mean agreement of assigning the point within a single point between human raters and E-rater was 92%.

2.2.2 AES for Japanese compositions written by non-native speakers

For Japanese, there is no one particular theory to what consists a good composition [9, 10, 11, 12].

Presently in Japan, there are several AES that have been developed in order to evaluate compositions written by non-native speakers. However, these AES are only able to detect one or two of the features of so called good compositions. For example, in the experiment of Gao et al [13], only vocabulary was used as an index to evaluate learner's compositions. Therefore, even though *Jess* is designed for L1 Japanese essays, it is meaningful to empirically analyse whether it can also be used for L2 Japanese compositions as the criteria of *Jess* seem to similar enough to the suggested criteria from the previous literature.

3 Methodology

In this study, 50 Japanese compositions were drawn from the Taiyaku Database by National Institute for the Japanese Language [14]. Nine teachers of the Japanese language were asked to evaluate these 50 compositions on a 100-point scale. These 50 compositions were also evaluated by *Moririn* and *Jess*. In order to seek the closest weights for evaluating compositions for AES in comparison to the human raters, nine different weights of *Jess* were used. All raw scores were z-score normalised so that the normalised scores can be directly compared between AES and human raters.

4 Results and discussions

The correlations between raters are analysed using Spearman's rank correlation tests.

	Ave.	r
J523	0.001826	0.430
J5238	0.001041	0.450
J712	0.001087	0.449
J811	0.000700	0.463
J622	0.000661	0.465
J550	0.000704	0.463
J433	0.002692	0.416
J0010	0.062887	0.265
J1000	0.000805	0.459
M	0.026200	0.314
T1	0.000000	0.737
T2	0.000001	0.635
T3	0.000000	0.672
T4	0.000000	0.703
T5	0.000001	0.629
T6	0.000001	0.638
T7	0.000000	0.753
T8	0.000061	0.536
T9	0.000000	0.861

Table 3: Average agreement and r-correlation of human average and machine scores.

Firstly, the average normalised scores given by the human raters were compared with *Jess*, *Moririn* and each of the human raters. The results of these comparisons are given in Table 3. Table 3 shows that both AES and each of the human raters have significant correlation ($p < 0.05$) with the average scores of the human raters. Within the nine different weights of *Jess*, the parameter weights of rhetoric 6, organisation 2, and content 2 (J622) resulted with the highest correlation to the average of the human raters (shown in blue in Table 3), and the parameter weights of rhetoric 0, organisation 0, and content 10 (J0010) resulted with the lowest agreement with the average of the human raters (shown in red in Table 3).

However as a whole, the r-correlation of the average of human raters and each human rater ranges from 0.536 to 0.861 (Table 3) and it is higher than the highest r-correlation from *Jess* with J622 (blue number in Table 3).

Within the human raters, T9 has the highest r-correlation with the average of human raters as shown in green in Table 3. Figures 1 and 2 show the comparison of J622 and the average of the human scores. As can be seen from Figure 2, T9 gave every student the score which is within the four standard deviation corridor, whereas J622 gave the score which is outside the four standard deviation corridor in a few occasions as shown in Figure 1.

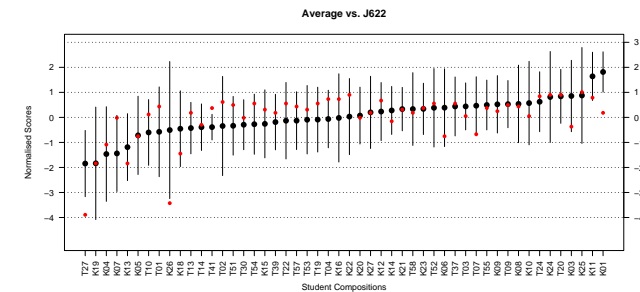


Fig. 1: The average scores of the nine raters (black) plotted in ascending order with two standard deviation above and the below the mean scores and the scores of J622 (red)

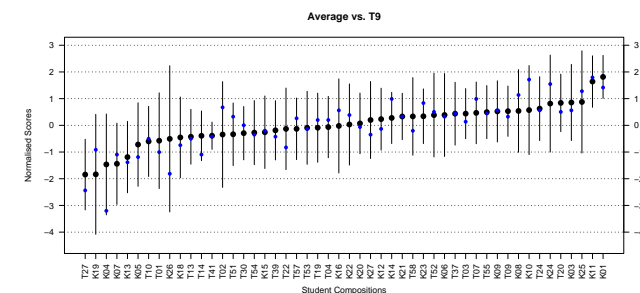


Fig. 2: The average scores of the nine raters (black) vs the scores of T9 (blue)

When each human rater is compared to each other, it can be seen from Table 4 that some human raters are very similar in ranking learners' compositions and that some human raters are not very similar with other human raters.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
T1	-	0.0015	0.0002	0.0019	0.0067	0.0126	0.0000	0.0041	0.0000
T2	0.0015	-	0.0070	0.0504	0.0005	0.0175	0.0002	0.1294	0.0003
T3	0.0002	0.0070	-	0.0020	0.0000	0.0269	0.0003	0.0782	0.0000
T4	0.0019	0.0504	0.0020	-	0.0056	0.0001	0.0068	0.0054	0.0001
T5	0.0067	0.0005	0.0000	0.0056	-	0.0002	0.0000	0.2415	0.0006
T6	0.0126	0.0175	0.0269	0.0001	0.0002	-	0.0080	0.0160	0.0002
T7	0.0000	0.0002	0.0003	0.0068	0.0000	0.0080	-	0.0001	0.0000
T8	0.0041	0.1294	0.0782	0.0054	0.2415	0.0160	0.0001	-	0.0045
T9	0.0000	0.0003	0.0000	0.0001	0.0006	0.0002	0.0000	0.0045	-
Ave.	0.0034	0.0258	0.0143	0.0090	0.0319	0.0102	0.0019	0.0599	0.0007

Table 4: Spearman's matrix of human raters (p values).

When the scores of each human rater and each AES parameter are compared as shown in Table 5, it can be observed that some human raters have similar tendencies in ranking learners' compositions with AES rather than other human raters. T4 and T9 have very high correlations with some parameters of *Jess* as shown in blue in Table 5. For example, as illustrated in Figure 3, T9 and *Jess* J1000 has a very similar ranking of learners' compositions.

	J523	J5238	J712	J811	J622	J550	J433	J0010	J1000	M	Ave.
T1	0.004	0.002	0.003	0.003	0.002	0.001	0.002	0.100	0.005	0.840	0.096
T2	0.182	0.166	0.152	0.139	0.150	0.298	0.363	0.949	0.087	0.132	0.262
T3	0.251	0.151	0.062	0.029	0.085	0.065	0.219	0.358	0.022	0.161	0.140
T4	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.001	0.011	0.002
T5	0.042	0.043	0.069	0.077	0.033	0.069	0.076	0.496	0.151	0.139	0.120
T6	0.026	0.020	0.009	0.009	0.013	0.105	0.060	0.033	0.010	0.291	0.058
T7	0.465	0.387	0.537	0.407	0.407	0.142	0.272	0.835	0.386	0.050	0.389
T8	0.184	0.098	0.148	0.096	0.137	0.125	0.079	0.148	0.119	0.021	0.115
T9	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.161	0.000	0.016	0.018
Ave.	0.128	0.096	0.109	0.084	0.092	0.089	0.119	0.343	0.087	0.184	-

Table 5: Spearman's matrix of 9 human raters and 10 different weights of *Jess* and *Moririn* (p values).

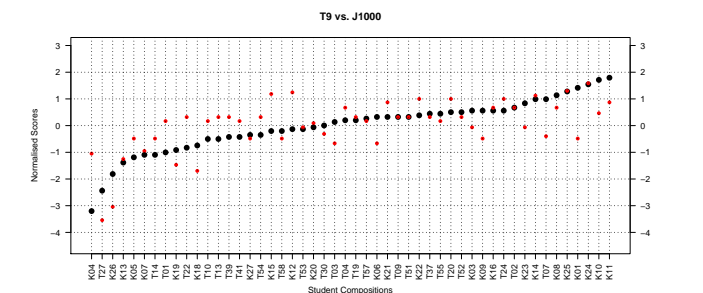


Fig. 3: The scores of T9 (black) vs the scores of J1000 (red).

5 Conclusions

In conclusion statistically speaking, AES should be able to be used in a Japanese language learning environment. However it is also true that human raters perform better in marking compositions in comparison to AES.

Each human rater has very different tendencies in marking compositions. Depending of parameter sets, some human raters show more similar tendencies with AES than with other human raters.

Some findings from this experiment suggest that future studies are needed to improve AES for L2 Japanese. For example, what Japanese teachers regard when they evaluate learners' compositions are different from what currently ready on-line essay scoring systems. Besides, there is not a standard criterion for L2 Japanese compositions unlike Jacobs' in English, and this might be one of the reasons why there is no AES for Japanese L2 composition yet.

6 Acknowledgements

The authors extend their sincere thanks to the nine teachers who responded to our request by evaluating compositions. This study was financially supported by the College of Asia and the Pacific, the Australian National University.

References

- [1] D. Castro-Castro, R. Lannes-Losada, M. Maritxalar, I. Niebla, C. Pérez-Marqués, N. Álamo-Suárez, and A. Pons-Porrata. 2008. A Multilingual Application for Automated Essay Scoring. *Advances in Artificial Intelligence*. IBERAMIA, 5290.
- [2] D. Lonsdale and D. Strong-Krause. 2003. Automated Rating of ESL Essays. *Technology Conference, Proceedings of the HLT-NAACL 03 Workshop on Building education applications using natural language processing*, 2:61-67.
- [3] Nihongo sakubun shouronbun kenkyukai. 2008. Nihongo no bunsyou kaiseiki sofuto moiririn. Retrieved 2008 from <http://www.mori7.info/moririn/index.php>.
- [4] T. Ishioka and M. Kameda. 2002. Kompyuutaa ni yoru nihongo shouronbun no jidou saiten sisutemu. *Technical report of IEICE*, 102(491):43-48.
- [5] T. Ishioka and M. Kameda. 2004. Automated Japanese Essay Scoring System: *Jess*. *Proceedings of the 15th International workshop on Database and Expert Systems Applications*.
- [6] M. Warschauer and P. Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):1-24.
- [7] H. Jacobs, S. Zingraf, D. Wormuth, V. Hartfiel and J. Hughey. 1981. *Testing ESL composition: A Practical Approach*, Rowley, MA: Newbury House.
- [8] J. Burstein and M. Chodorow. 1999. Automated Essay Scoring for Non-native English Speakers. Retrieved 2009 from <http://www.ets.org/Media/Research/pdf/erater.ac1999rev.pdf>.
- [9] F. Morita. 1981. Sakubun no hyouka. *Nihongo kyouiku*, 43:17-33.
- [10] Y. Kikuchi. 1987. Sakubun no hyouka houhou ni tsuite no ichi shian. *Nihongo kyouiku*, 63:87-104.
- [11] M. Tanaka, Y. Tsubone and A. Hajikano. 1998. Daini gengo toshite no nihongo ni okeru sakubun hyouka kijun—ii sakubun no kettei youin. *Nihongo kyouiku*, 99:60-71.
- [12] K. Murakami. 2001. Sakubun no hyouka no shinraisei to hyoukasha no tokusei—Nihonjin bogo washa to hi bogo washa no hyouka no hikaku. *Nagoya daigaku nihongo, nihon bunka ronshuu*, 9:47-69.
- [13] B. Gao, T. Kodaka and H. Ogura. 2002. N-gram bumpu ni yoru gaikokujin nihongo gakushuuya nihongo sakubun hyouka no kokoromi. *Journal of the Institute of Electronics, Information and Communication Engineers*, 85:1083-1087.
- [14] Y. Usami. 2006. Sakubun taiyaku deetaa sakusei no mokuteki to sono tayou ni kansuru kenkyuu chousa houkokusho, 9-42.