Experimenting with a Japanese automated essay scoring system in the L2 Japanese environment

Jun Imaki
Shunichi Ishihara
College of Asia and the Pacific, Australian National University, Australia

The purpose of this study is to provide an empirical analysis of the performance of an L1 Japanese automated essay scoring system which was on L2 Japanese compositions. In particular, this study concerns the use of such a system in formal L2 Japanese classes by the teachers (not in standardised tests). Thus experiments were designed accordingly. For this study, *Jess*, Japanese essay scoring system, was trialled using L2 Japanese compositions (n = 50). While *Jess* performed very well, being comparable with human raters in that the correlation between *Jess* and the average of the nine human raters is at least as high as the correlation between the 9 human raters, we also found: 1) that the performance of *Jess* is not as good as the reported performance of English automated essay scoring systems in the L2 environment and 2) that the very good compositions tend to be under-scored by *Jess*, indicating that *Jess* still has possible room for improvement.

Key words: L2 writing, automated essay scoring system, Japanese, rater reliability, essay evaluation, rater training

Introduction

This study investigates a Japanese Automated Essay Scoring (AES) system, assessing how well the system compares with human raters in scoring compositions written by second language (L2) Japanese learners. In particular, with the use of an AES system for regular L2 Japanese classes (not for standardised tests) in mind, we designed the experiments to be as realistic as

Address for correspondence: Jun Imaki, School of Culture, History and Language, College of Asia and the Pacific, Baldessin Precinct Building #110, The Australian National University, Canberra, ACT 0200, Australia; Email: jun.imaki@anu.edu.au.

possible for ordinary L2 Japanese classes. In other words, we aim to assess the validity of an AES system as a reliable tool for L2 Japanese teachers.

Lonsdale and Strong-Krause (2003, p. 1) note that AES are desirable for use in L2 classes for the following three reasons:

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; and
3. Feedback: providing feedback to a learner is important, and AES can provide prompt feedback with specific suggestions.

The first point, practicality, is obvious. Traditionally, assessment of written tasks involves various activities, such as design of prompts, creation of rubrics, training of assessors, scoring and giving feedback. This can be quite an expensive process in terms of time and human resources but AES systems can drastically reduce some of their burdens (Chapelle & Douglas, 2006; Weigle, 2013). Practicality is important considering the current situation surrounding L2 education of high staff-student ratios, the increase in staff teaching load and the replacement of permanent teaching staff with sessional teaching staff (White, Baldauf & Diller, 1997). AES systems are consistent, whereas essay grading is a very subjective task for people, being susceptible to various humanistic conditions such as impatience, fatigue or the *halo effect* (Enright & Quinlan, 2010; Hughes, 1989; Norbert & Williamson, 2013; Weigle, 2010). The advance of natural language processing (NLP) and automatic text analysis techniques has enabled AES systems to give diagnostic feedback on various aspects of writing, such as lexical complexity, syntactic variety, discourse structures, grammatical usage, word choice and content development. This is done promptly and also for different levels of language skills (e.g. beginner or advanced), or in a multilingual manner, facilitating students' ability to interpret the feedback (Attali & Burstein, 2006; Chen & Cheng, 2008; Deane, 2013; Vantage Learning, 2008; Warschauer & Ware, 2006).

The developers of AES systems report high correlations of their systems with human raters (Attali & Burstein, 2006; Burstein, Kukich, Wolff, Chi & Chodorow, 1998; Elliott, 2003; Vantage Learning, 2008). However, AES is not free from scepticism. Criticism includes the non-transparency of the marking process and criteria (Weigle, 2002), the validation procedure of AES (Weir, 2005), the difference in cognitive processes between human raters and AES (Drechsel, 1999), the large amount of training data for AES (Conian, 2009) and the possibility of fooling AES (Deane, 2013; Powers, Burstein, Chodorow, Fowles & Kukich, 2002). Regarding the use of AES systems in L2 classes in particular, Conian (2009, p. 263) suggests that major studies of AES systems "have been skewed…[and that] there has been a focus on native speaker writing". Similarly, Warschauer and Ware (2006) also note there has been little empirical research on how accurate and useful AES systems are in ESL classes (or more generally, L2 classes).

Shifting the focus to Japanese, there are currently two AES systems: *Jess* (Ishioka & Kameda, 2004) and *Moririn* (Nihongo sakubun shouronbun kenkyuu kai, 2008), which are readily available online. However, to the best of our knowledge, neither of these systems has been evaluated on their accuracy in the L2 Japanese environment (Deane, 2013). This was the primary motivation for this research.

Japanese is a popular L2 choice among students. The number of people who are studying Japanese overseas is about 4 million in over 100 countries (The Japan Foundation, 2013). However, the increase in students has not been met by that of teachers. According to the Japan Foundation (2013), over the period from 1979 through 2012, the number of students increased 31 times, while the number of teachers increased only 15 times. Considering the large potential of AES and the constant increase of L2 Japanese learners, L2 Japanese education could receive large benefits from AES. However, the validity of AES needs to be assessed before its implementation. Therefore, this will be part of the rationale for investigating the validity of Japanese AES in the L2 Japanese environment.

English AES systems have been used in various settings, such as large-scale language testing programs, placement tests and classrooms. "[W]ith an increasing demand for technology use in educational institutions, more and more teachers and students have used, are using, or are considering using AES in their classrooms" (Chen & Cheng, 2008, p. 95). Given this fact, Chen and Cheng (2008) claim that a more pressing question is not whether this technology should or should not be used, but how it can be used to achieve more desirable learning outcomes. Consequently, it appears that the focus of AES studies has started shifting to the effectiveness of AES systems as a pedagogical tool both in first language (L1) (Attali, 2004; Shermis, Mzumara, Olson & Harrington, 2001; Warschauer & Grimes, 2008) and L2 English writing classes (Cheung et al., 2007; Weigle, 2013).

The current study focuses on the reliability of a Japanese AES system as a tool for teachers in their classes, by comparing the correlations between computer-generated and human-rater scores with the correlations obtained from a number of human raters. *Jess* is used in this study.

Automated Essay Scoring: English

AES is a relatively young area of study with a history of almost 45 years. In 1966, Ellis Page developed a computer based grading system named Project Essay Grader (PEG™) which is regarded as the pioneer of AES (Page, 1996). Since then, many AES systems have been developed. There have been many studies reporting that AES systems are comparable to human raters in that the human-computer agreement (r=0.75~0.85) is at least as high as the human-human agreement (r=0.70~0.85). Refer to Ben-Simon and Bennett (2007) for a summary of these studies.

The main commercial use of AES systems has been in the grading of large scale standardised tests, such as the Graduate Record Examination writing assessments, and TOEFL writing assessments (Weigle, 2010). In actual classrooms, AES systems are not merely used as tools which return an overall, holistic score of an essay, but are more likely to be used as an essay critiquing system or a writing assistance tool (Cheung et al., 2007).

Critiques from Conian (2009) and Xi (2010) that major studies of AES focus on native speaker writing are true; however, there are a few studies including Burstein & Chodorow (1999b), Conian (2009) and Enright & Quinlan (2010) which report that AES systems perform well with compositions written by ESL students, being comparable to human raters.

AES Overview: Japanese

Following success in English AES, there have been several Japanese AES produced since the 1980s. However, in the early stages of their development, these systems were developed not as 'scoring' systems but as editing and proofreading tools for Japanese newspaper and publishing companies (Ishioka & Kameda, 2004). Some scholars (Gao, Odaka & Ogura, 2002; Hasegawa, 2006; Oyama, 2010) attempted to develop AES for L2 Japanese, but these researches did not go beyond tentative proposals and they never implemented their proposals as evaluation tools in actual L2 Japanese classes.

Jess (Japanese Essay Scoring System) was designed and developed to be used in entrance exams for Japanese universities. The developers of *Jess*, Ishioka and Kameda (2004, 2006), describe it as a Japanese version of E-rater®, accommodating the linguistic characteristics of Japanese. *Jess* evaluates essays according to the following three criteria (Ishioka & Kameda, 2004);

1. Rhetoric: which includes a) the ease of reading such as median and

maximum sentence length, kanji/kana ratio, b) diversity in vocabulary, c) percentage of big words, and d) 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.

There is one significant difference between *Jess* and E-rater® in terms of training materials. *Jess* is trained with editorials and columns of newspapers. The developers of *Jess* emphasise this point as the most significant aspect of *Jess*, not requiring any pre-marked training essays (Ishioka & Kameda, 2004).

Jess allows users to designate weights for the above criteria. However, the default weights are set as rhetoric (5), organisation (2), and content (3). The final score of the learner's essay is calculated by deducting from a perfect score of 100%.

Ishioka, Sagisaka and Nimura (2003) conducted a series of experiments to examine the reliability and comparability of *Jess* with human raters. They used essays written by new university graduates as part of a selection process for company entrance examinations. These essays were scored by *Jess*, as well as by a number of human raters. Prior to marking, the human raters received training of about an hour in order to eliminate subjective judgments as much as possible. Inter-class correlation (ICC) coefficients were also calculated to see the correlations between the human raters. They found that the correlation coefficients (r) between the average of the human raters and *Jess* were between 0.43 and 0.57 for the three sets of essays. They also reported that the ICC coefficients between the human raters were between 0.45 and 0.59. From the similarities between Pearson's correlation coefficients and the ICC coefficients, they concluded that *Jess* can substitute one of the human raters, although they acknowledge that these coefficients appear to be smaller than equivalent surveys of English (Powers et al., 2002).

Methodology

Database

In this study, 50 Japanese compositions written by Thai (T) and Korean (K) learners of L2 Japanese were randomly drawn from the Taiyaku Database which was compiled by the National Institute for Japanese Language (Usami, 2006). The compositions, totalling 50 (25 each from the Thai and Korean students), were written on the topic of smoking. These two countries were selected for practical reasons; they were the only countries in the database with sufficient numbers of compositions under the same topic. After reading an instruction sheet (see Appendix), students were asked to write a composition of about 800 characters. The majority of the students had studied Japanese for between 15 and 24 months. The students' level of language skill can be judged as being from the upper beginner to lower intermediate level. We decided to focus on students with this level of Japanese because 1) students at lower-beginner level, rather than an upper-beginner level, may have found it difficult to express their opinions in written Japanese and 2) a large number of students belong to this beginner-intermediate level so there would potentially be a high demand for AES due to the large number of compositions to be marked. The average length of these 50 compositions was 655 characters (standard deviation=140; min=323; max=998).

Raters

Nine teachers of L2 Japanese, including both native and non-native, male and female speakers, participated in this study as raters. The nine raters (R1~R9) had various levels of teaching experience, ranging from a single year to over ten years of teaching at Australian universities. Two of the raters were non-native speakers of Japanese. A set of 50 compositions was given to each rater in the same order together with an instruction sheet for the composition (Appendix) and an instruction sheet for marking given below. The raters were asked to mark the compositions holistically on a 100-point scale. The instructions given to the raters were minimal as shown below:

1. Grade this composition, written by students of Japanese, in a manner you find appropriate.
2. Preferably try to mark them so as to also rank the students within their level; however, the final judgement is left to your discretion.

As explained in the Introduction, we presupposed the use of an AES system by L2 Japanese teachers for their classes, and accordingly we tried to emulate the conditions arising from this presupposition in the experiments. Due to financial and time limitations that L2 education in Australian universities is currently facing, it is very difficult to provide extensive training for raters. Under these circumstances, the best practice that L2 teachers can undertake is perhaps to provide raters with specific marking criteria and exemplars, and then to hold a post-marking discussion for moderation between the raters. Some L2 teachers might say that even this is unreasonably optimistic. Therefore, we decided to simply ask the raters to mark the compositions with minimal instructions. Holistic scoring was chosen as "it saves time and money and at the same efficiency" (Bacha, 2001, p. 374). We also judged that a 100-point scale is more widely accepted in formal class settings than Likert-based 5 or 7 point scales (employed in many previous AES studies).

These 50 compositions were also evaluated by *Jess*. *Moririn* is not included in this study mainly because the system is completely black-boxed in terms of its algorithms for scoring. Also, since *Jess* was designed to evaluate entrance exams for Japanese universities, we decided *Jess* is more suitable as a valid instrument for academic Japanese.

Table 1 contains the configurations of the parameter weights examined in this study. We tried 8 different configurations in total.

Table 1. The 8 configurations of parameter weights used in this study*

Index	J523	J550	J4338	J6228	J8118	J0010	J1000	J0100
Rhetoric	5	5	4	6	8	0	10	0
Organisation	2	5	3	2	1	0	0	10
Content	3	0	3	2	1	10	0	0
Length	-	-	800	800	800	-	-	-

*For example, J523 stands for the weights of 5, 2 and 3 for Rhetoric, Organisation and Content, respectively.

Data Analysis

The correlations 1) between the nine human raters and 2) those between the average of the nine human raters and *Jess* were investigated using Pearson's correlation. Two different types of scores were used: the raw scores and the normalised z-scores.

Where not specified, raw scores were used for statistical analysis. The normalised scores were mainly used to examine how much the individual scores given by *Jess* deviated from the scores given by the human raters.

Results and Discussions

Between human raters

Inter-rater variability was investigated across the nine human raters. Table 2 contains Pearson's correlation coefficient (r) for each comparison between the nine raters. As can be seen from Table 2, all comparisons show positive r values ($0.293 \leq r \leq 0.707$). The correlation is statistically significant for all comparisons. However, different degrees of correlations were found between different comparisons of raters. For example, R3 shows a very strong positive correlation with R9 ($r = 0.707$) while R8 has only a weak positive correlation with R5 ($r = 0.293$). R9 shows a strong or very strong positive correlation with all of the other raters (r ≥ 0.479), indicating that R9 employed a typical marking criteria across the nine human raters. The correlation of R8 with the other raters is significant, but relatively weaker compared to the other raters.

Table 2. A matrix of the correlation coefficient values for the comparisons across the human raters

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**
R1	-	0.484	0.519	0.439	0.428	0.456	0.575	0.510	0.666
R2	0.484	-	0.494	0.313	0.515	0.401	0.529	0.324	0.531
R3	0.519	0.494	-	0.484	0.587	0.567	0.560	0.394	0.707
R4	0.439	0.313	0.484	-	0.418	0.537	0.402	0.413	0.584
R5	0.428	0.515	0.587	0.418	-	0.552	0.572	0.293	0.514
R6	0.456	0.401	0.567	0.537	0.552	-	0.484	0.436	0.593
R7	0.575	0.529	0.560	0.402	0.572	0.484	-	0.573	0.607
R8	0.510	0.324	0.394	0.413	0.293	0.436	0.573	-	0.479
R9	0.666	0.531	0.707	0.584	0.514	0.593	0.607	0.479	-

Table 2 displays the strengths/weaknesses of the correlations between raters. However, it is rather difficult to grasp the overall view of the inter-rater correlations. Therefore, the inter-class correlation (ICC) test was conducted. Two different sets of scores were used for this test: the raw scores and the normalised z-scores. The results of the ICC tests are given in Table 3.

Table 3. The results of the ICC tests for the human raters based on the raw and normalized scores

	Raw	**Normalised**
ICC	0.344	0.544

Table 3 indicates that there is a large difference between the ICC coefficient calculated from the raw scores (= 0.344) and that from the normalised scores (= 0.544). This difference indicates that there is a strong positive correlation between the raters in terms of the ranking order of the compositions while different raters use different levels of standard (i.e., 50% or 70% as the average) and different ranges of scores (i.e., scores are given in the range of 40%~90% or 0%~100%).

As explained above, the amount of marking instruction given to the nine raters was minimal. Despite this, judging from the ICC coefficient (= 0.544) based on the normalised scores, the correlation across the nine raters is high enough to say that the ranking order of the compositions is fairly consistent across the nine raters. This observation implies that there is a certain degree of shared judgement amongst Japanese teachers for marking compositions written by beginner and intermediate students. However, the actual raw scores assigned to any particular composition varied between the nine raters. This point can also be confirmed by Figure 1 which contains the boxplot of the raw scores for each human rater. It is evident from Figure 1 that different raters scored the compositions using different levels of standard and different ranges of scores. For example, R1's average score for compositions was very similar to R3's average score, but the range of the scores they used is very different. R5 used a wide range of scores to mark the compositions while R1 used a very narrow range of scores. R9 set the highest level of standard of all, resulting in R9's average score being the lowest out of the nine.

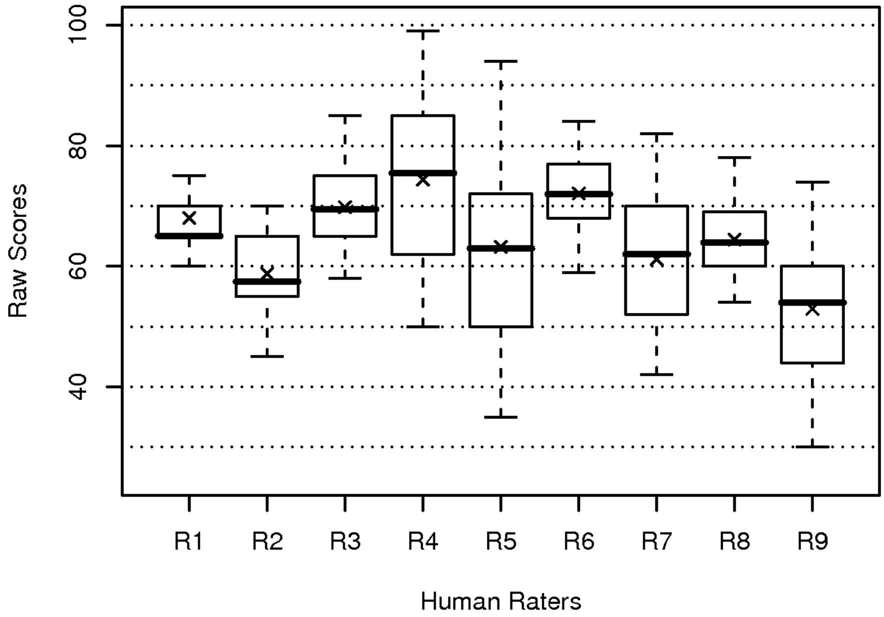

Figure 1. Boxplots showing the variabilities of the scores given by the nine human raters.

Ishioka, Sagisaka and Nimura (2003) reported that the ICC coefficient values across their human raters were between 0.45 and 0.59. Compared with these ICC coefficient values, the ICC coefficient of 0.344 based on the raw scores, which is comparable to Ishioka, Sagisaka & Nimura (2003), appears to be significantly lower. However, this may be explained by the following points: 1) their raters were trained for marking whereas our raters were not trained for this particular task; 2) their raters were asked to give scores holistically with the aid of an analytical process whereas our raters were asked to give only holistic scores; and 3) a five-point scale was used for their study while a 100-point scale was used for our study. As discussed in the Methodology section, the lack of the rater training is a possible scenario: the L2 climate in Australia is such that rater training may not be possible, and we would expect this is to contribute to low ICC coefficient values. However, the low ICC coefficient in the current context actually indicates that if one cannot provide appropriate training for raters, it is better to use an AES system. However, judging from the findings of Baird, Greatorex and Bell (2004, p. 346), that "exemplar work of students and discussion between examiners may not be so beneficial to inter-rater reliability if a community of practice already exists", our nine raters who had been working together in the same L2 Japanese program are considered to be already in an established community of practice, thus the effect arising from the lack of training could be minimised. The employment of holistic scoring can be also a reason for the low ICC coefficient. However, the relationship between analytical and holistic scoring is not clear; some researchers (e.g. Bacha, 2001; Chang, 2002) report significant correlations between them, and some even report that the holistic scale has higher inter-rater agreement than analytical ones (Barkaoui, 2007). Thus, it is difficult to conclude that the low ICC coefficient is also attributable to the holistic scale we employed in our study. Some studies see the value of holistic scoring for evaluating classroom essays (Charney, 1984; Cumming, 1990), and Bacha (2001, p. 374) says that "it saves time and money and at the same efficiency". Thus, the use of holistic scoring in our study is appropriate and realistic.

Another possible reason for lower ICC coefficient values is that the level of compositions used in our study was relatively similar, as the learners were assumed to be upper beginner to the lower-intermediate level. Under these circumstances the differences in the level of compositions are less clear. In a formal teaching setting, students belonging to the same class usually share a similar level of proficiency, unlike standardised tests or placement tests in which the level of proficiency varies.

Comparing the correlations between the nine human raters of the current study for Japanese and the correlations between human raters reported in previous studies for English L1 (Ben-Simon & Bennett, 2007) and L2 (Burstein & Chodorow, 1999a; Conian, 2009), the correlation for Japanese is lower than that for English. Estimates of relatively poor validity have been reported between human raters for L1 Japanese essays across various papers (Akutsu, Kikuchi, Suzuki & Watanabe, 2006; Ando, 1974; Ikeda, 1992), including Ishioka, Sagisaka and Nimura (2003). It is not clear at this stage what contributes to the large variability across human raters for evaluating Japanese essays. However, it has been pointed out that 'what is a good essay' is not well defined not only in L2 but also L1 Japanese education (Inoue & Okuma, 1985; The National Institute for Japanese Language, 1978). Therefore, there no well-accepted standard for assessing writing in Japanese (Yoshikawa & Kishi, 2006). Consequently, although the importance of so-called 'academic writing' has started to be recognised, particularly in tertiary education in Japan, and some fundamental research relating to academic writing can also be found in L2 Japanese teaching (Kinoshita, 1990; Sasaki, 2000), it is undeniable that education in academic writing has been less weighted in Japanese than in English.

Machine and human raters

The scores of *Jess* with different parameter weights were compared with the average scores of the nine human raters. The results are given in Table 4.

Table 4. The correlations between the average of the 9 human raters and *Jess* with different weights

Index	J523	J550	J4338	J6228	J8118	J0010	J0100	J1000
	0.514	0.552	0.526	0.562	0.556	0.355	0.465	0.521

As far as the r values are concerned, J6228 (Rhetoric: 6, Organisation: 2, Content: 2; Length: 800) performed best (r = 0.562), showing a strong positive correlation with the average of the human raters. J0010 (Rhetoric: 0, Organisation: 0, Content: 10), on the other hand, clearly under-performed (r = 0.355) compared to the rest of the weight configurations. This indicates that human raters put more weight on Rhetoric and Organisation than Content when they mark compositions, conforming to what is reported in Watanabe, Taira and Inoue (1988). It can also be seen from Table 4 that different weightings of Rhetoric and Organisation do not significantly influence the performance of *Jess* when a light weight is given to Content.

The r value of 0.562 achieved by J6228 is very similar to the r values (= 0.43~0.57) reported in Ishioka, Sagisaka and Nimura (2003). The r value of 0.562 is higher than the ICC coefficient values (= 0.344 and 0.544) given across the raters (refer to Table 3). As far as the variability between human raters and *Jess* is concerned, this result implies that *Jess* can be used as a rater, substituting one of the nine raters for L2 Japanese compositions.

We have demonstrated that *Jess* performs equally as well as human raters (or even better) for marking the compositions of L2 Japanese learners, in particular under the realistic conditions of L2 Japanese education in Australia. However, it is necessary to note that the performance of *Jess* in the L2 Japanese environment is not as good as the performance of English AES systems in the ESL environment, in that the correlation between English AES and human raters is higher than that between *Jess* and human raters. The average correlation coefficient between the scores of E-rater® and those of two human raters was 0.693 in Burstein and Chodorow (1999), and the correlation coefficients between the scores of BETSY™ and those of two human raters were very strong (r = 0.826~0.859) in Conian (2009).

Landauer, Laham and Foltz (2003) report that the correlations between human raters and AES may differ depending on the materials to be assessed, in that the human-machine agreement rate may be poorer for classroom essays than standardised test essays. The compositions used in our study have characteristics which are typical of classroom compositions, including its topic; the raters were classroom teachers and the compositions were marked using a 100-point scale. On the other hand, the majority of previous studies on the validity of English AES are based on standardised test essays both in L1 (Ben-Simon & Bennett, 2007) and L2 (Burstein & Chodorow, 1999a; Conian, 2009; Enright & Quinlan, 2010). Therefore, relatively poor agreement between *Jess* and the human raters may be attributed to the fact that the compositions used in this study have the characteristics of classroom essays.

We should not ignore the possibility that *Jess* may not be robust enough for L2 Japanese, being not as well developed as English AES systems for ESL essays. Ishioka, Sagisaka and Nimura (2003) point out a difficulty in accurately assessing the logical structure of compositions of only 600 to 800 characters. Particularly considering that the average length of *Jess's* training data is about 1600 characters, the lengths of the compositions (average length = 665 characters) in the current study may have adversely affected the results. As Cumming (1990) reports, essay raters clearly distinguish L2 proficiency and writing expertise (= writing skills) when evaluating ESL compositions, while different marking strategies are used for L1 and L2 writings. Needless to say, the compositions written by students whose level of Japanese is between upper-beginner and lower-intermediate contain more lexico-grammatical errors than those written by advanced-level students. Obviously *Jess* cannot detect these errors in the same way that human raters can; hence linguistic accuracy may be differently evaluated by *Jess* and the human raters. While it is possible to argue that *Jess* can be used as a rater, substituting one of the human raters due to its comparability with human raters in the L2 Japanese environment, the comparisons between *Jess* and its English counterparts also suggest that there may be room for *Jess* to improve as an AES system.

Deviations between human and machine scores

We investigated the overall performance of *Jess* compared with human raters under the realistic conditions in which *Jess* is used by L2 Japanese teachers in their classes. However, it is also useful to see what sort of scores *Jess* actually assigned to each composition in comparison to human raters. In other words, even if the overall performance of *Jess* is comparable to human raters in terms of correlation, if *Jess* assigns significantly different scores to individual compositions from human raters, it will raise a reliability question.

In

Figure , the overall average normalised scores (black circles) of the nine human raters are plotted in ascending order with 2 standard deviations above and below them (vertical bars). That is, assuming normally distributed scores, two standard deviations either side of the mean will include about 95% of all observations. The normalised scores of the compositions that J6228 evaluated are also plotted (crosses) in

Figure .

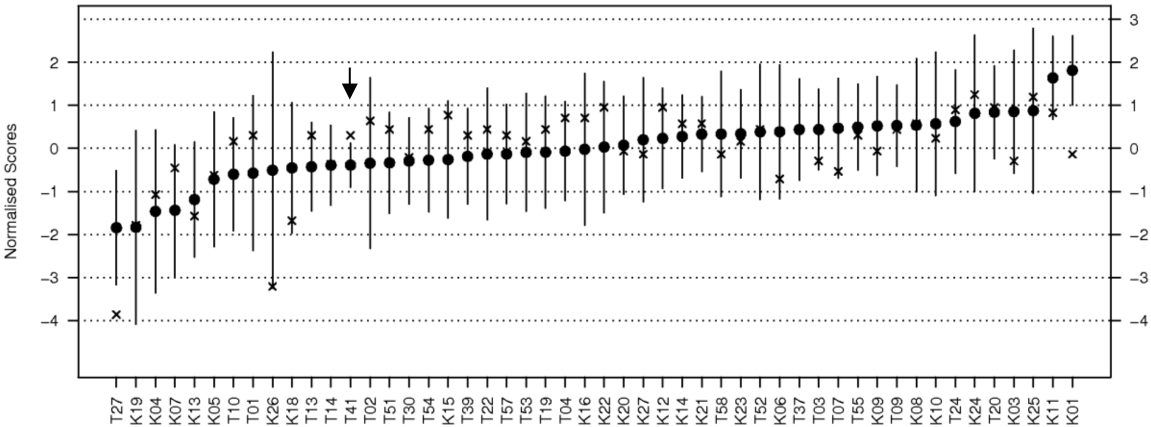

Figure 2. Human rater average vs. J6228. Arrows indicate those scores of J6228 which exist outside the standard deviation corridor.

As can be seen in Figure 2, some normalised scores of J6228 (T27, K26, T41 and K01) are located outside the standard deviation corridors (refer to the arrows in

Figure). Most of the normalised scores that are placed outside the standard deviation corridor (T27, K26 and K01) were underscored compared to the marks given by human raters. Those compositions which are ranked at both ends of the ranking order (T27 and K01) appear to be difficult to be accurately marked by *Jess*. It is also observable from

Figure that J6228 did not give very high scores (> 1.5) to any of the compositions, with many of the normalised scores being distributed between 0 and 1, despite human raters not hesitating to give high scores (i.e., the overall average normalised scores are evenly distributed between -2 and 2). The skewness of the normalised scores of J6228 is -1.49 while that of the average scores of the nine human raters is -0.351. The same trend was evident in other weight configurations of *Jess*. Since *Jess* is trained with editorials and columns of newspapers written by professional writers, it is understandable that L2 Japanese learners cannot compete with professional writers, resulting in no high scores. However, this could be a problem related to the use of *Jess* in the L2 Japanese environment as very good compositions may also be under-scored by *Jess*.

Conclusion

Having the use of an AES system for formal L2 Japanese classes in mind, we conducted some experiments in order to assess how well the AES system is comparable with human raters in scoring classroom compositions written by L2 Japanese leaners. First, we demonstrated that a Japanese AES system, *Jess*, performs well in scoring compositions written by L2 Japanese learners in that the correlation between *Jess* and the average of the nine human raters is as high as the correlation between the nine human raters. While the variability between human raters is high, this result suggests that *Jess* is comparable to human raters in scoring L2 Japanese compositions. Possible reasons for the high variability of human raters are the inherent characteristics and difficulties associated with current L2 education, including the lack of the rater selections based on strict criteria, rater training and workshops, and monitoring

Having said that, we also became aware that *Jess* underperforms compared to English counterparts in L2 essays. While discussing possible reasons for this, we also commented that there is still room for *Jess* to improve as an AES system. We also noticed that very good compositions tended to be under-scored by *Jess* compared to the human raters, possibly due to the fact that *Jess* is trained on materials written by professional writers.

Acknowledgements

This paper is a revised and extended version of the study presented at a joint conference for the Japanese Studies Association of Australia and the International Conference on Japanese Language Education in Sydney, Australia in 2009. This study is financially supported by the College of Asia and the Pacific, the Australian National University.

References

Akutsu, H., Kikuchi, K., Suzuki, H., & Watanabe, Y. (2006). A study of essay tests (1): Agreement between raters. *Iwate daigaku kyouiku gakubu fuzoku kyouiku jisshi sougou center kenkyuu kiyou, 5*, 115–122.

Ando, K. (1974). Shouronbun saitenhou no ichi kentou (A study of essay evaluation method). *Proceedings of Annual Convention of the Japanese Association of Educational Psychology, 16*, 492–493.

Attali, Y. (2004, April 12–16). *Exploring the feedback and revision features of* Criterion. Paper presented at the Meeting of the National Council on Measurement in Education, San Diego, California.

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring with E–rater v.2. *Journal of Technology, Learning and Assessment, 4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371–383.

Baird, J.–A., Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331–348.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed–method study. *Assessing Writing, 12*(86–107).

Ben–Simon, A. & Bennett, R. E. (2007). Toward more substantively meaningful Automated Essay Scoring. *The Journal of Technology, Learning, and Assessment, 6*(1). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/.

Burstein, J. & Chodorow, M. (1999, June). *Automated Essay Scoring for nonnative English speakers.* Paper presented at the ACL99 workshop on computer–mediated language assessment and evaluation of natural language processing, College Park, MD.

Burstein, J., Kukich, K., Wolff, S., Chi, L. & Chodorow, M. (1998, August). *Enriching Automated Essay Scoring using discourse marking.* Paper presented at the Workshop on Discourse Relations & Discourse marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada.

Chang, Y.–F. (2002). EFL teachers' responses to L2 writing. *Education Resources Information Center*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED465283.

Chapelle, C. & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65–81.

Chen, C.–F. E. & Cheng, W.–Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perspective learning effectiveness in EFL classes. *Language Learning & Technology, 12*(2), 94–112.

Cheung, W., Cheung, W. K., Mørch, A. I., Wong, K. C., Lee, C., Liu, J. et al. (2007). Grounding collaborative learning in semantics–based critiquing. *International Journal of Distance Education Technologies, 5*(2), 40–55.

Conian, D. (2009). Experimenting with a computer essay–scoring program based on ESL student writing scripts. *ReCALL, 21*(2), 259–279.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51.

Deane, P. (2013). On the relation between Automated Essay Scoring and modernv of the writing construct. *Assessing Writing, 18*(1), 7–24.

Drechsel, J. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two–Year College, 26*(4), 380–187.

Elliott, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein, (Eds.), *Automated Essay Scoring: A cross–disciplinary perspective.* (pp. 71–86). Mahwah: Lawrence Erlbaum.

Enright, M. K. & Quinlan, T. (2010). Complementing human judgement of essays written by English language learners with E–rater® scoring. *Language Testing, 27*(3), 317–334.

Gao, J., Odaka, T. & Ogura, H. (2002). A trial of composition evaluation by n–gram distributions for Japanese composition of foreign Japanese–learners. *The transactions of the Institute of Electronics, Information and Communication Engineers, J85*(1), 1083–1087.

Hasegawa, M. (2006). Sakubun hyouka to koubunteki tokuchou ni tsuite – ryuugakusei no placement test wo rei ni– (Essay Evaluation and Features of Sentence Structure – Examples from exchange student placement tests). *Nihongo kenkyuu (Research of Japanaese Language, 26*, 59–73.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge UP.

Ikeda, H. (1992). *Tesuto no kagaku (Science of tests)*. Tokyo: Nihon bunka kagaku sha.

Inoue, N. & Okuma, T. (1985). *Jyugyou ni yakudatsu bunshouron buntairon (Useful theories for sentence and language style for the classroom)*. Tokyo: Kyouiku Shuppan.

Ishioka, T. & Kameda, M. (2004, August 30 – September 3). *JESS: An automated Japanese essay soring system.* Paper presented at the the 15th International Workshop on Database and Expert Systems Applications, Spain.

Ishioka, T. & Kameda, M. (2006, July). *Automated Japanese Essay Scoring system based on articles written by experts.* Paper presented at the 21st International Conference on Computional Linguistics and the 44th Annual Meeting of the Association or Computional Linguistics, Sydney, Australia.

Ishioka, T., Sagisaka, Y. & Nimura, H. (2003). *Jess: Nihongo shouronbun jidousaiten shisutemu – nyuushashiken niyorusakubundeeta no hyouka (Jess: Japanese Essay Scoring System – Evaluation of Essay Data at Company Entrance Examinations).* Paper presented at the Conference of the Japanese federation of Statistical Science Association Nagoya, Japan.

Kinoshita, K. (1990). *Repooto no kumitatekata (How to compose essays).* Tokyo: Chikuma Shobo.

Landauer, T. K., Laham, D. & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross–disciplinary perspective* (pp. 87–112). NJ: Erlbaum.

Lonsdale, D. & Strong–Krause, D. (2003, May). *Automated rating of ESL essays.* Paper presented at the HLT–NAACL 03 Workshop on Building education applications using natural language processing, Edmonton, Canada.

Nihongo sakubun shouronbun kenkyuu kai. (2008). Nihongo no bunshou kaiseki soft moririn (*Japanese sentence analysis system – Moririn).* Retrieved December 5th, 2008, from http://www.mori7.info/moririn/index.php.

Norbert, E. & Williamson, D. M. (2013). Assessing Writing special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*(1), 1–6.

Oyama, H. (2010). Automatic error detection method for Japanese particles. *Ritsumeikan Asia Pacific University Polyglossia, 18*, 55-63.

Page, E. B. (1996, April 11). *Grading essays by computer: Why the controversy?* Paper presented at the NCME, Invited Symposium, New York.

Powers, D., Burstein, J., Chodorow, M., Fowles, M. & Kukich, K. (2002). Stumping E–rater: challenging the validity of Automated Essay Scoring *Computers and the Humanities, 18*, 103–134.

Sasaki, Y. (2000). Case studies in theme–oriented expressions of opinion: Native speakers of Japanese and non–native speakers whose first language is Chinese. In the National Institute for Japanese Language (Ed.), *Nihongo* *kyouiku no tame no asia shogengo no taiyaku sakubun data no shuushuu to corpus no kouchiku* (pp. 219–230). Tokyo: the National Institute for Japanese Language.

Shermis, M. D., Mzumara, H. R., Olson, J. & Harrington, S. (2001). On–line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26*(3), 247–259.

The Japan Foundation. (2013). *Japanese–language education overseas*. Tokyo: The Japan Foundation.

The National Institute for Japanese Language. (1978). *Ability of written expression and composition in elementary school children*. Tokyo: Tokyo Shoseki Publishers.

Usami, Y. (2006). *Sakubun taiyaku database sakusei no mokuteki to sono tayou na katsuyou ni tsuite (The purpose of creating an essay translation database and its various uses)*. Tokyo: National Institute for Japanese language.

Vantage Learning. (2008). *My Access! Efficacy report*. Newtown, PA: Vantage Learning.

Warschauer, M. & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*, 22–36.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching research, 10*(2), 1–24.

Watanabe, H., Taira, Y. & Inoue, S. (1988). An analysis of essay examination data. *Bulletin of the Faculty of Education, University of Tokyo, 28*, 143–164.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non–test indicators of writing ability. *Language Testing, 27*(3), 335–353.

Weigle, S. C. (2013). English language learners and automated Scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85–99.

Weir, C. J. (2005). *Language testing and validation: An evidence–based approach*. NY: Palgrave MacMillan.

White, P., Baldauf, R. & Diller, A. (1997). *Languages and universities: Under siege*. Canberra.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300.

Yoshikawa, A, & Kishi, M. (2006). An examination of items for evaluating compositions: What is the factor that affects evaluation of opinion texts. *Bulletin of Tokyo Gakugei University. Educational sciences, 57*, 93–102.

Appendix

Instructions for Composition

Read the following text and write your own opinion in around 800 Japanese characters.

Smoking is becoming a problem in Japan. Some people are saying "A law should be made to prohibit smoking in public places such as offices, restaurants, buses and trains. On top of that smoking commercials on TV have a bad influence on children so they should also be disallowed." However, others say "It's strange to create a rule forbidding smoking. Anyone should have the right to smoke."

What do you think? Write your own opinion on smoking.