Item Development and Scoring for Japanese Oral Proficiency Testing

Hitokazu Matsushita, Deryle Lonsdale

Department of Computer Science, Department of Linguistics
Brigham Young University
h.matsushita@byu.edu, lonz@byu.edu

Abstract

This study introduces and evaluates a computerized approach to measuring Japanese L2 oral proficiency. We present a testing and scoring method that uses a type of structured speech called elicited imitation (EI) to evaluate accuracy of speech productions. Several types of language resources and toolkits are required to develop, administer, and score responses to this test. First, we present a corpus-based test item creation method to produce EI items with targeted linguistic features in a principled and efficient manner. Second, we sketch how we are able to bootstrap a small learner speech corpus to generate a significantly large corpus of training data for language model construction. Lastly, we show how newly created test items effectively classify learners according to their L2 speaking capability and illustrate how our scoring method computes a metric for language proficiency that correlates well with more traditional human scoring methods.

Keywords: elicited imitation, Japanese oral proficiency, item development, speech recognition

1. Background

Reliable and timely second language (L2) oral proficiency tests are usually costly and complicated to administer. The oral proficiency interview (OPI)[1] is a widely administered test that proceeds as follows: (1) A test taker produces prompted speech samples based on particular chosen topics in an interview with a certified evaluator; these speech samples are recorded for later evaluation. (2) Human evaluators rate the speech samples based on a rubric to produce an overall score. Raters need to be highly trained and turnaround time for evaluation can be substantial. Because of its price and duration, it is typically used as a high-stakes, milestone evaluation. It tests several widely accepted features essential for measuring L2 oral proficiency including pronunciation, vocabulary choice, morphosyntactic formation, and discourse structure (ACTFL, 1999).

Automating oral proficiency scoring is highly desirable, but current automatic speech recognition (ASR) technology is not yet completely viable for handling disfluent speech samples, such as those produced by language learners at varying levels of proficiency. However, current ASR is capable of dealing with certain types of constrained language. In elicited imitation (EI) testing, the test taker hears and repeats back a set of carefully engineered sentences of varying complexity one-by-one. The responses are recorded for subsequent scoring. The EI test has several advantages:

- Unlike in conversational oral interview tests, test takers benefit from multiple fresh starts, an oft-cited desideratum (Hughes, 2003); each of the EI test sentences (also called 'items') offers a chance to regroup and restart.
- Test administration is very time-efficient, taking approximately ten minutes to complete a sixty-item test (Graham et al., 2008).
- The grading process is virtually automatic with ASR, reducing substantially the time to complete the entire grading process.
- The EI scoring results are strictly numeric due to the objective and analytic nature of the test. Test results are thus applicable to both longitudinal and cross-sectional studies to examine learners' progress and compare their performance in a quantitative manner.
- The speech samples collected during the test can be later used for qualitative studies: when educators or researchers need to investigate learners' particular characteristics (e.g., speech error patterns) in the EI tasks. It is also possible to conduct similar qualitative studies, such as examining the relationship between learners' test performance and their learning experience. In fact, the findings obtained through such qualitative analyses are highly important and necessary for further testing and grading procedure refinement.

Past work in EI testing has identified key features that EI test items need to address: (1) The length (whether controlled by morae, syllables, or words) must exceed participants' short-term memory capacity, thus precluding rote repetition. (Jessop et al., 2007). (2) Lexical and morphological features must be carefully chosen or else items may be too easy or too difficult for learners to repeat (Tomita et al., 2009). (3) Target features should ideally be placed in the middle position of the sentence (Erlam, 2006). Satisfying these constraints is crucial for producing optimal EI items (Christensen et al., 2010). EI items must also target salient grammatical features that reflect the goals of the learners at various levels, curriculum designers, instructors, and evaluators. Taking all these requirements into consideration when developing EI items is daunting.

Ideal features vary across languages: effective Japanese EI items are considerably longer than English ones, and morphological complexity even in short Japanese sentences can be very high. Several theoretical and empirical linguistic studies have led to three classes of linguistic phenom-

[1] See `http://www.languagetesting.com`.

ena that we have found are necessary for high-precision EI testing of Japanese (Matsushita et al., 2010; Matsushita & LeGare, 2010):

- Japanese relative clauses are highly complex, requiring high memory load (since the clause precedes the head) and often leading to garden paths (Sawa, 2005; Carroll, 2008; Sawasaki, 2009). They are also more flexible in terms of NP extraction possibilities (Nakayama, 2002; Comrie, 2010). Creating short sentences with this feature for Japanese is relatively straightforward (Tsujimura, 2007).
- Processing multiple embeddings is challenging (e.g., Bader & Bayer (2006) for English). Since Japanese allows zero pronouns, embedded clauses with *pro*-drop are even more complicated to process, especially for language learners.
- Japanese evidentiality marking in a sentence implies sources of information the speaker relies on (McCready & Ogata, 2007), and its usage is often very subtle. Whether assertions are based on inference, previous direct experience, or hearsay often has very little or nothing to do with the referents in the sentence. Yet native Japanese speakers are able to judge where the knowledge comes from.

Since EI items must satisfy a number of linguistic constraints, they are often contrived and even nonsensical when developed wholly through manual means. On the other hand, they are more effective when they are more natural-sounding. This situation is ideal for application of the use of language resources: to the extent that "real-world" sentences from a corpus can meet the criteria, they can be used as EI items. By annotating corpus sentences, particular target features can be identified and useful sentences can be selected for EI testing purposes.

This paper describes our development of a set of Japanese EI items through the use of various language corpus and annotation resources. It also explains how we administered the resulting EI test to several students and scored their speech samples using various levels of current ASR technology. We demonstrate how the EI items created through the corpus-based approach effectively classify learners according to their oral proficiency levels and how the test scores correlate well with more traditional human scoring methods.

2. Using available resources

Scoring EI items can be done by hand: annotating each syllable, mora, word, or sentence with a binary score indicating whether it was successfully reproduced or not. We had human graders hand-score Japanese EI test items using our specially developed tool. Inter-rater reliability was high, as it was for other languages in prior research efforts. More interesting is the possibility of automatically scoring of EI items using automatic speech recognition (ASR). To achieve optimal EI scoring via the binary mora-based method, we conducted a series of studies to develop optimal language models.

2.1. System I

Our baseline system (System I) uses the Julius speech recognition engine (Lee & Kawahara, 2009). The acoustic model was trained with 20,000 sentences from the *Mainichi Shimbun* newspaper read aloud by 120 native speakers of Japanese. For each EI item we built its own custom language model consisting of just that sentence and its case marking variations, which often occur in Japanese learner speech. Recognition via these language models is in essence equivalent to a finite-state grammar approach to ASR (Shikano et al., 2007) We used more alignment via dynamic programming to produce binary scores.

To evaluate the performance of System I, we selected 60 Japanse EI items from a previous SPOT test[2] and administered them to 98 learners of Japanese of various proficiency. We then scored the collected speech samples with System I. For comparison with ASR-generated scores, seven human raters hand-scored the same speech data. Both ASR and human grading processes employed the mora-level binary scoring method.

Figure 1(a) shows that System I correlates strongly (at 0.9840) with human grading scores, though the lower-half ASR scores are overly generous compared to human ones. The grammar-based language models lacked wider coverage, and hence coerced wrongly pronounced morphemes to correct ones in the recognition process. Furthermore, several high-scoring achievers in this EI test caused a ceiling effect, because of the relative easiness of the chosen SPOT EI items.

2.2. System II

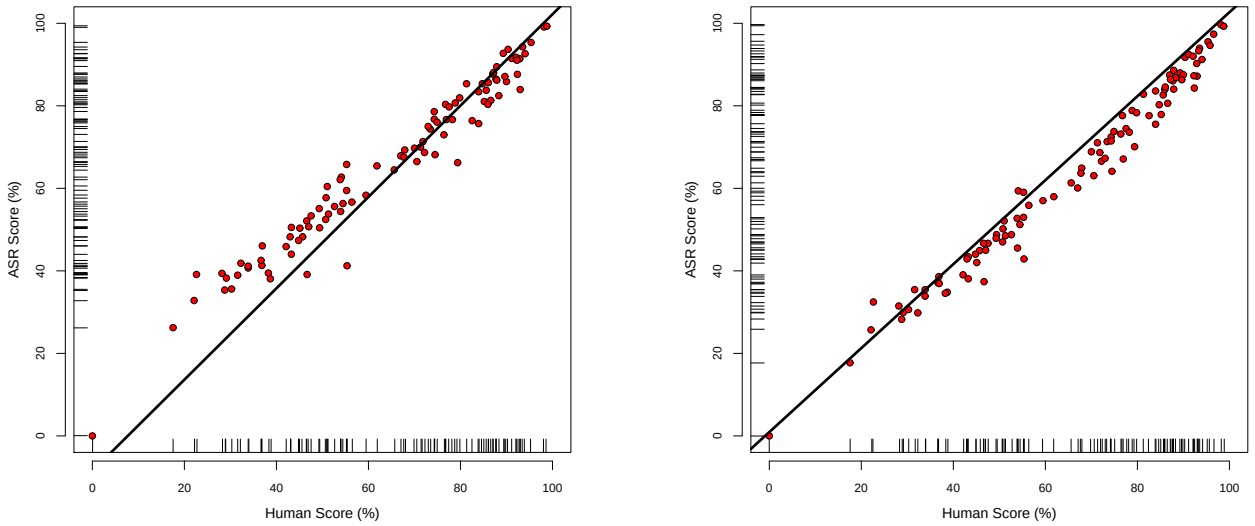
To solve the baseline system's overly-favorable grading issue, we developed a new grading system (System II) with a new language model. It is built with the whole Corpus of Spoken Japanese (CSJ), a large-scale transcribed Japanese speech corpus (Maekawa, 2003; NINJAL, 2006) with over 1,400 speakers (2/3 of whom are male), about 660 hours of speech, and over 7.5 million tokens. CSJ has several advantages for EI item development: it is speech-based and covers several genres, it precisely annotates a wide variety of speech-specific phenomena including disfluencies (e.g., fillers and fragments, repairs, word coalescence, etc.), and the data is stored in an XML database, permitting easy manipulation.

With the full set of CSJ language data in its language model, System II assures wider coverage in order to handle error-filled EI responses (Matsushita et al., 2010). We incorporated all 3,286 CSJ corpus files into a single language model via Palmkit[3], a tool for building Julius-compatible language models. Transcriptions of the correct EI sentences were also added to the language models and boosted to artificially exaggerate their influence in the dataset, thus predisposing the scorer to recognize them.

The System II evaluation used the same 60-item SPOT-based EI test mentioned previously for System I. Figure 1(b) shows that the System II scores correlate about as well to the human scores (at 0.9886) as System I did, even with

[2]Simple Performance-Oriented Test (Kobayashi et al., 1996).

[3]Available at `http://palmkit.sourceforge.net`.



(a) System I

(b) System II

Figure 1: Correlation analyses of human vs. ASR scoring

the substantially increased coverage with the CSJ data and its attendant increase in perplexity. Note that System I's overly generous grading for low scores is now lessened due to the additional language information provided by CSJ. However, the majority of ASR scores over 60% are now lower than human scores, presumably because System II was influenced by the perplexity increase. Interlanguage-influenced disfluency phenomena and unlikely collocations in EI responses are lacking in the language model; we suspect that CSJ's L1 speech alone cannot resolve the significant discrepancy between L1 and L2 production. We determined to incorporate L2 language data to System II to overcome the problem.

3. Resources for item creation and scoring

System III overcomes the shortcomings of its predecessors in two ways: (1) by using a set of EI items that we carefully engineered for this purpose, and (2) by incorporating more Japanese learner (JL2) language into the language models (Matsushita & Tsuchiya, 2011). In this section, we describe the corpus-based approach facilitating the item creation process in an efficient manner, and the bootstrapping method which integrates EI transcription data with CSJ in order to effectively specify L2 speech for scoring.

3.1. Item creation

In a previous study (Matsushita et al., 2010), we investigated the effectiveness of sentences containing unique morphosyntactic features occuring in Japanese as EI items. Along with the SPOT-based EI items described above, we generated eight additional test items containing the three Japanese linguistic features mentioned earlier. We created these items manually based on native speakers' linguistic intuitions, where native speakers selected target features, created several candidate items, judged them for naturalness to select the most plausible one, and then adjusted sentence length and lexical items as needed. These newly created test items were incorporated in the previous 60-item EI test, and we administered the test to a separate subject group ($n = 157$) and examined whether these new items would curb the ceiling effect we observed in the previous study. We evaluated the subjects' performance with System II and showed that there is a significant difference between the new eight items and SPOT-based items in terms of subjects' EI performance and their proficiency levels ($p < 0.0001$ with the factorial ANOVA).

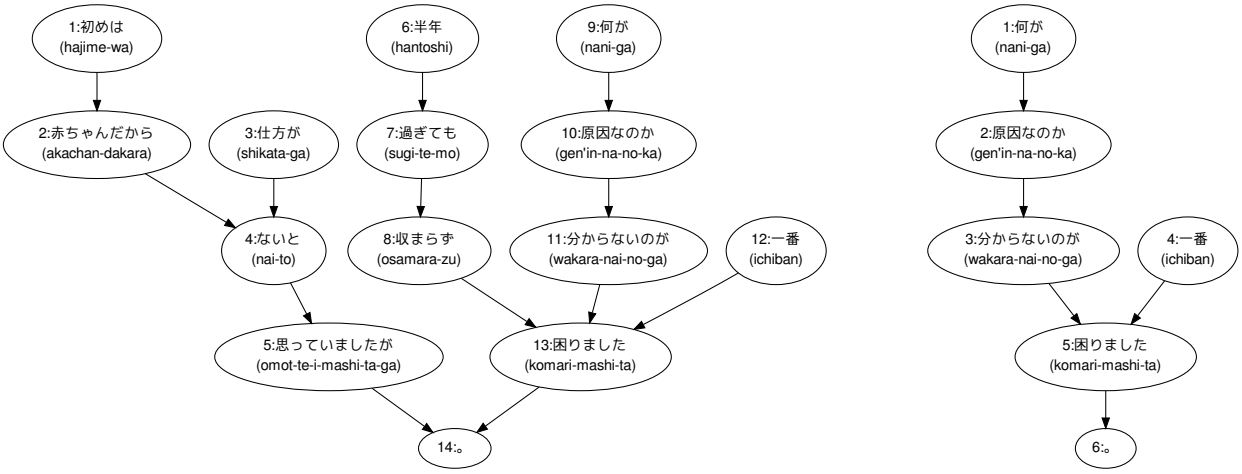
However, the problem with this item creation approach is that it is difficult to develop multiple high-quality items with desirable linguistic features in a short time, because generating item candidates with complex target features is quite labor-intensive. It is also rare that all item creators agree when judging artificially generated item candidates as well-formed and natural-sounding. Therefore, it is important to establish an systematic item creation method to lessen the burden on item creators and to generate a substantial number of desirable item candidates in an efficient manner.

To address this issue, we utilized the CSJ data and several corpus toolkits to facilitate the item creation process. We first processed the corpus with the Japanese lexical/morphological analyzer Mecab 0.98[4] which annotated morphemes, POS tags, pronunciation, and lexical classes. Next we used the Japanese dependency structure analyzer CaboCha 0.53[5] to create dependency relations for MeCab-generated morphemes. We then performed searches through the re-annotated CSJ transcripts using ChaKi.NET[6], a corpus search tool that allowed us to

[4]Available at `http://mecab.sourceforge.net`.

[5]Available at `http://chasen.org/taku/software/cabocha`.

[6]Available at `http://sourceforge.jp/projects/chaki/`.



(a) Original sentence

(b) Simplified sentence

Figure 2: Refining an overly complex sentence for EI use: (a) CSJ sentence with center embedding and (b) repurposed for EI item use (arrows indicate the dependency relationships).

retrieve draft sentences for item development. ChaKi.NET utilizes a dependency grammar framework to manage and access the constituency relations. Our queries executed most of the search functions utilized previously for English (Christensen et al., 2010): regular expressions, dependencies, word lists, and so forth. The tool retrieves multiple sentences that satisfy the queries, so humans must hand-select eventual EI items. Figure 2(a) shows an example sentence directly retrieved from CSJ, consisting of 67 morae and 13 bunsetsu[7]. This sentence is not ideal for EI use because of its length and complexity. However, trimming two branches from the sentence results in an appropriate sentence, shown in Figure 2(b) consisting of 27 morae and 5 bunsetsu. Final sentences are stored in an SQL database for future use.

3.2. Bootstrapping a learner corpus

Now we describe our use of learner corpus resources to further improve System III performance. Our previous EI test administrations have resulted in a small collection of JL2 speech errors, often with the same item across several speakers. By transcribing and annotating the EI responses we have collected, we have a characterization of some errors JL2 speakers make. These transcriptions can then be incorporated into new language models and thus taken into consideration when scoring EI items.

The main obstacle with this approach was quantity: the amount of transcribed JL2 speech samples is quite small compared to the CSJ L1 data. Our solution, often adopted in other natural language modeling contexts, is to bootstrap existing data to create an artificially induced corpus of analogous errors, in this case positing possible learner errors. Adding them to the previous system should increase its capability to more accurately deal with interlanguage-influenced EI responses. For this process we employed analogical modeling (AM), an exemplar-based learning and modeling system[8] that uses analogy and which has been shown to perform well in linguistic tasks, including modeling naturally occurring errors (Skousen et al., 2002). We thus extended the CSJ EI transcriptions by using AM to generate possible errors based on observed errors from these prior tests (Matsushita & Tsuchiya, 2011).

We transcribed a randomly selected 20% of the EI responses, and from it created some 500 AM exemplars to form a training dataset. Each exemplar contained as features a sliding window with surrounding morphemes, along with codes for each reflecting its status: correct (C), insertion (I), deletion (D), and substitution (S); the outcome was the morpheme itself. The test set ran the vectors with various output patterns to generate ample possible outcomes for each vector. AM thus behaved as a virtual learner, performing EI tasks one morpheme at a time based on knowledge about previous errors provided by the training set. We interpreted all possible AM outcomes as transitions, re-encoding them as finite-state grammars in Backus Naur Form (BNF). These grammars were used to generate 5,000 sentences by permuting AM-supplied morpheme patterns. The resulting artificial responses were then integrated into the statistical language models for Julius.

3.3. Combining item creation and System III

Figure 3 sketches the overall development of System III as just described. To summarize, a number of candidate items are extracted using target features via consulting the annotated CSJ corpus through a corpus management tool. Among these candidates, experimental EI items are selected according to the specified criteria and modified as necessary (e.g., Figure 2). The effectiveness of the experimental items is then examined in the test administration along with actual test items, and the results are returned as

[7] a unit of Japanese syntactic constituency

[8] Available at `http://humanities.byu.edu/am/`.

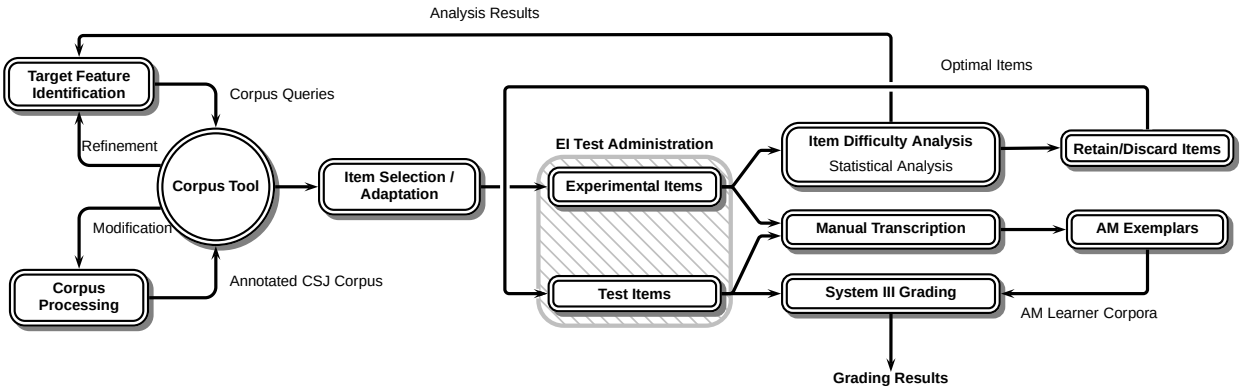


Figure 3: Combination of corpus-based item creation and System III

feedback to the next item creation. The collected EI speech samples are manually transcribed and utilized to train the AM system which produces artificial L2 corpora for the augmentation of System III's language models. Through this cyclic process, a substantial number of optimal test items are quickly generated and stored in the item database, and at the same time the System III is refined through AM-produced language data.

4. Results

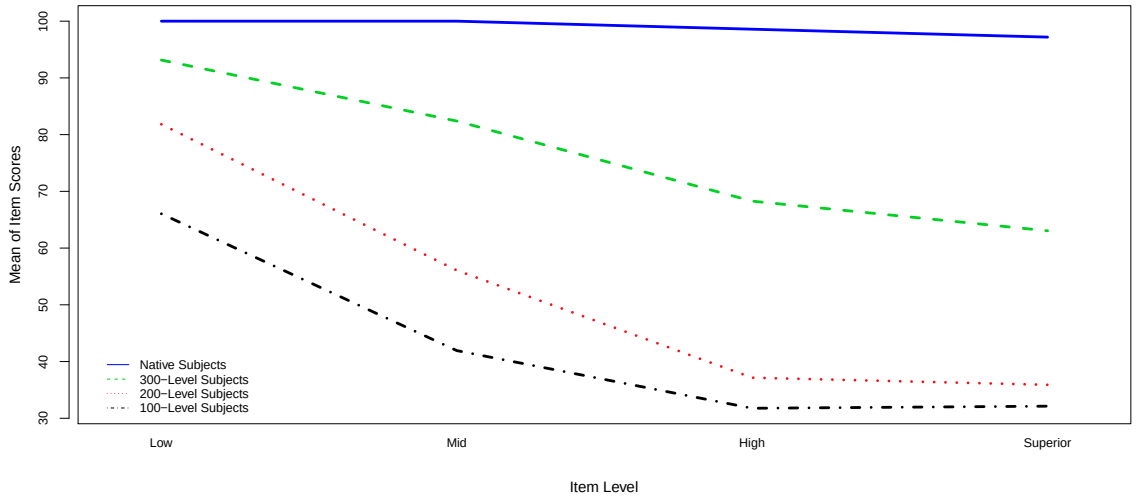
By following the item creation process described above, we created a new sixty-item Japanese EI test: thirty were generated from CSJ sentences, with four also incorporating the complicated Japanese morphosyntactic features mentioned earlier; the other thirty items came from textbooks used in Japanese courses. We also classified these items as Low, Mid, High, and Superior according to the difficulty levels of linguistic features and mora length. This item creation process for these sixty items required only several hours, which was approximately the same length of time needed to generate eight items through the manual item engineering method mentioned in 3.1. We administered the EI test to 239 JL2 learners in our Fall 2010 semester; most (approximately 93%) are native speakers of English, and the rest native speakers of Korean, Spanish, Chinese, or Japanese. They represent all proficiency levels (i.e., courses numbered 100, 200, 300 or above, and native). Due to poor recording quality eight participants were excluded from the evaluation. Using Julius we processed the audio recordings of the responses to produce dictation texts of the EI responses and then converted the results to binary scores with the mora alignment algorithm mentioned in 2.1. To produce human-generated scores, two native speakers of Japanese transcribed the EI responses manually. The transcriptions were also decomposed into morae with MeCab and manually corrected when necessary. The binary scores were produced with the aforementioned mora alignment algorithm and used as the gold standard for the comparison with ASR-generated scores.

Figure 4 shows the effectiveness of the corpus-based EI items to classify subjects based on their proficiency represented by their course levels. The interaction plot in Figure 4(a) indicates the relationship between subject groups and item difficulty levels. While native subjects' scores were stable throughout all the item levels, the non-native subjects' performance was greatly hindered as the item difficulty increased. Table 4(b) shows the unique patterns of each subject group's performance according to the item difficulty levels. As indicated, the low-level items differentiated the 100-level group from the other subject groups, and the 200-level group as well as the 100-level group were separated from the others by the mid-level items, and the high- and superior-level items made a complete distinction between the native subjects and the others. Thus, the results clearly indicate that these corpus-based EI items are eminently capable of classifying subjects based on their proficiency levels.

	System I	System II	System III
Agreement (%)	84.8	83.8	86.0
Unweighted κ	0.686	0.669	0.713
Rater Bias	0.550	0.441	0.500
Item-Level r	0.9024	0.8940	0.9088
Subject-Level r	0.9815	0.9799	0.9852

Table 1: IRR and Correlation Statistics with Human-Generated Scores of Three Grading Systems

Table 1 shows the scoring accuracy of three systems. The first three rows indicate the inter-rater reliability (IRR) statistics against human scores, and the last two rows indicate the item-by-item and subject-by-subject correlation coefficients between human and ASR scores. As shown, the mora agreement, Cohen's κ, and the rater bias indicate that the EI scores produced by System III were the closest to those by humans among the three. The correlation results also indicate System III's scores were most highly correlated with the humans'. Figure 5 shows the regression analysis and score distributions of human- and ASR-generated (System III, subject-level) EI scores. Figure 5(a) visualizes the strong correlation to human scoring through-



(a) Interaction plot

Group Pair \ Item Level	Low	Mid	High	Superior
L100—L200	15.78***	14.15***	5.39	3.76
L100—L300	27.10***	40.46***	36.50***	30.92***
L100—Native	33.96***	58.07***	66.81***	65.05***
L200—L300	11.32***	26.31***	31.12***	27.15***
L200—Native	18.18	43.92***	61.43***	61.29***
L300—Native	6.86	17.60	30.31**	34.14**

(b) Tukey post-hoc test

Figure 4: Differences in EI scores based on subject groups and item levels (**$p < 0.01$, ***$p < 0.001$)

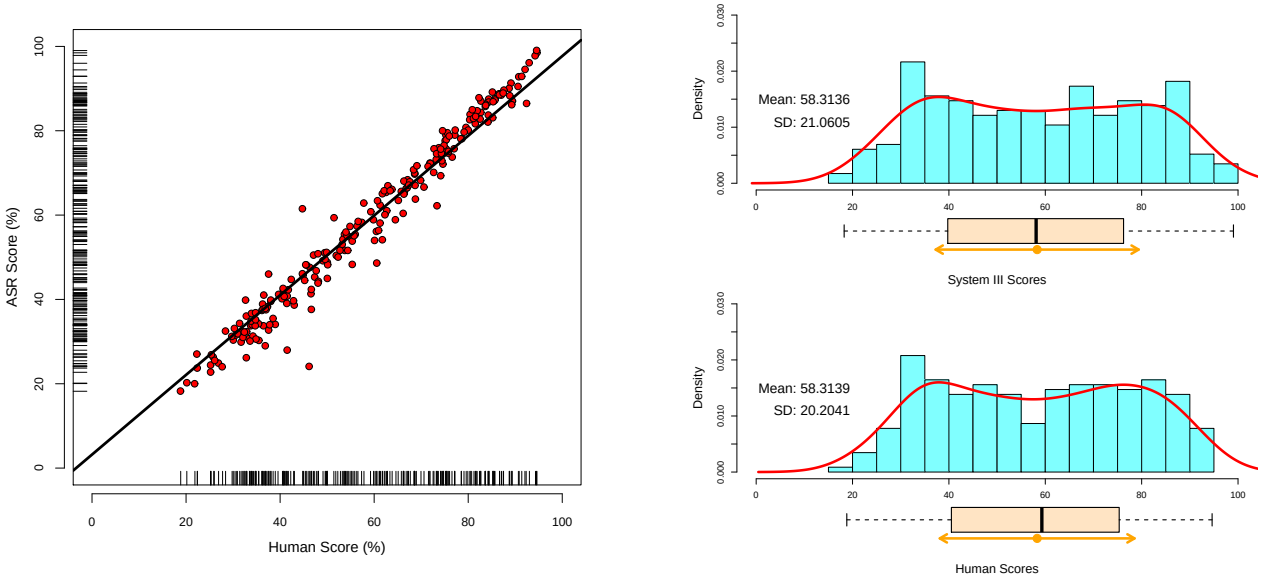
out all the proficiency levels. It also indicates that the inconsistencies of human and ASR scores of particular proficiency groups observed in System I and II were rectified by System III. Figure 5(b) also shows the close similarities between human and System III scores according to the exhibited score distribution patterns. Therefore, it is safe to say that the bootstrapped AM-generated corpora have significantly enhanced the grading capability of System III, which cannot be attained by the grammar-based or CSJ-only previous systems. Note that this success was achieved while only using 20% of the entire transcription data; further enhancement is possible by incorporating more EI error transcription data into the AM modeling process. System III, with its carefully engineered EI items through the corpus-based approach based on targeted linguistic concepts, and scored automatically using language models that contain bootstrapped learner language along with L1 language usage, gives a high-value assessment on learners' proficiency in terms of L2 accuracy.

5. Conclusions and future work

This study presents and evaluates an approach to measuring comprehensive Japanese L2 oral proficiency using EI testing and scoring methods. Similar techniques have been used by others for other languages (Müller et al., 2009; Bernstein et al., 2010) but our approach is innovative for Japanese; we have also applied our approach to other languages in the past. We have shown that corpus-based test items effectively classify learners according to their language proficiency and how Japanese EI recordings can be automatically scored using various levels of ASR technology to produce a metric of proficiency that correlates well with human grading. We also showed how a language modeling technique was used to successfully bootstrap a corpus of learner errors.

Several further directions could be pursued in this work. Our EI test focuses on the accuracy of learner speech, whereas the other widely accepted dimension for analysis is fluency, involving such features as hesitation patterns, speech burst length, turn taking, length of narration, and discourse management (ACTFL, 1999). Ongoing work involves using another type of speech test to address these aspects of spontaneous speech and thus increase face validity of the EI test.

In addition, we are not currently measuring pronunciation accuracy explicitly, since that is beyond the current ASR capabilities, highly subjective, and somewhat ill-defined even for human graders. Though our scoring was post-hoc,



(a) Regression analysis

(b) Score distribution

Figure 5: Regression and score distribution analyses of System III

it would be more desirable to grade EI items in real time during the test administration. In fact, this would enable adaptive testing, as we have shown for English EI testing (Lonsdale & Christensen, 2011).

Test score calibration was done against our university's Japanese class levels as an approximation to OPI scores, but not in a direct comparison to OPI scores. The correlation, while very good, is not perfect. Further statistical analyses must be carried out to increase the reliability and validity of the EI test and produce scores closely comparable with such oral tests.

We also envision creating tools to help on our EI item transcription and design processes to assist in targeting pertinent morphological, syntactic, and semantic structures. Such tools would enable access to various NLP and language resources including corpora, lexical databases, and vocabulary lists.

Many EI utterances contain snippets of L1 English embedded in the Japanese responses. By adding English content to the language models, we might improve recognition, since the ASR engine tends to be misled by English words and phrases.

6. References

ACTFL. 1999. *Oral Proficiency Interview Tester Training Manual*. American Council on the Teaching of Foreign Languages, New York.

Bader, M. & Bayer, J. 2006. Introducing the human sentence processing mechanism. In *Case and Linking in Language Comprehension*, volume 34, pages 19–47. Springer Netherlands.

Bernstein, J., Van Moere, A., & Cheng, J. 2010. Validating automated speaking tests. *Language Testing*, 27:355–377.

Carroll, D. W. 2008. *Psychology of Language*. Thompson Wadsworth, Belmont, CA, 5th edition.

Christensen, C., Hendrickson, R., & Lonsdale, D. 2010. Principled construction of elicited imitation tests. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D., (Eds.), *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10)*, pages 233–238. European Language Resources Association (ELRA), Valetta, Malta.

Comrie, B. 2010. Japanese and the other languages of the world. *NINJAL Project Review*, 1:29–45.

Erlam, R. 2006. Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27:465–491.

Graham, C. R., Lonsdale, D., Kennington, C., Johonson, A., & McGhee, J. 2008. Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*, pages 57–67, Marrakech, Morocco.

Hughes, A. 2003. *Testing for Language Teachers*. Cambridge University Press, Cambridge, UK, 2nd edition.

Jessop, L., Suzuki, W., & Tomita, Y. 2007. Elicited imitation in second language acquisition research. *The Canadian Modern Language Review*, 64(1):215–220.

Kobayashi, N., Ford-Niwa, J., & Yamamoto, H. 1996. Nihongo no atarashii sokuteihoo SPOT *[SPOT: A new method for measuring Japanese ability]. Sekai no Nihongo Kyooiku* [Japanese Education in the World], 6:201–218.

Lee, A. & Kawahara, T. 2009. Recent development of open-source speech recognition engine Julius. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

Lonsdale, D. & Christensen, C. 2011. Automating the scoring of elicited imitation tests. In *Proceedings of*

the ACL-HLT/ICML/ISCA Joint Symposium on Machine Learning in Speech and Language Processing.

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, Tokyo.

Matsushita, H. & LeGare, M. 2010. Elicited imitation as a measure of Japanese L2 proficiency. In *Paper presented at Association of Teachers of Japanese (ATJ)*, Philadelphia, PA.

Matsushita, H. & Tsuchiya, S. 2011. The development of effective language models for an EI-based L2 speaking test: Capturing Japanese interlanguage phenomena with ASR technology. In *Paper presented at American Association for Applied Linguistics (AAAL)*, Chicago, IL.

Matsushita, H., Lonsdale, D., & Dewey, D. 2010. Japanese elicited imitation: ASR-based oral proficiency test and optimal item creation. In Weir, G. R. S. & Ishikawa, S., (Eds.), *Corpus, ICT and Language Education*, pages 161–172. University of Strathclyde Publishing, Glasgow, UK.

McCready, E. & Ogata, N. 2007. Evidentiality, modality and probability. *Linguistics and Philosophy*, 30:147–206.

Müller, P., de Wet, F., van der Walt, C., & Nielser, T. 2009. Automatically assessing the oral proficiency of proficient L2 speakers. In *Proceedings of SLaTE 2009*.

Nakayama, M. 2002. Sentence processing. In Tsujimura, N., (Ed.), *The Handbook of Japanese Linguistics*, pages 398–424. Blackwell Publishing, Malden, MA.

NINJAL. 2006. Nihongo hanashikotoba kopasu no kochikuho *[The construction of the Corpus of Spontaneous Japanese]*. `http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/csj_report/CSJ_rep.pdf`, National Institute for Japanese Language and Linguistics.

Sawa, T. 2005. Comprehension of the relative clause structure in Japanese: Experimental examination for Japanese sentence processing by the self-paced reading method. *Proceedings of Tokyo Gakugei University*, 56:329–333.

Sawasaki, K. 2009. Nihongo gakushuusya no kankeisetsu rikai: Eigo, kankokugo, chuugokugo bogo wasya no yomi jikan kara no koosatsu *[Processing of relative clauses by learners of Japanese: a study on reading times of English/Korean/Chinese L1 speakers]*. *Daini Gengo to shite no Nihongo no Shuutoku Kenkyuu* [Acquisition of Japanese as a Second Language], 12:86–106.

Shikano, K., Ito, K., Kawahara, T., Takeda, K., & Yamamoto, M. 2007. *Onsei Ninshiki Shisutemu [Speech Recognition System]*. Ohmsha, Tokyo, Japan, 7th edition.

Skousen, R., Lonsdale, D., & Parkinson, D. B. 2002. *Analogical Modeling: An exemplar-based approach to language*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Tomita, Y., Suzuki, W., & Jessop, L. 2009. Elicited imitation: Toward valid procedures to measure implicit second language grammatical knowledge. *TESOL Quarterly*, 43:345–349.

Tsujimura, N. 2007. *An Introduction to Japanese Linguistics*. Blackwell Publishing, Malden, MA, 2nd edition.