Reading Skill Test to Diagnose Basic Language Skills in Comparison to Machines

Noriko H. Arai (arai@nii.ac.jp)
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Naoya Todo (ntodo@niad.ac.jp)
NIAD-QE, 1-29-1 Gakuen-nishimachi, Kodaira-shi, Tokyo 187-8587, Japan
Teiko Arai (arai-teiko@g.ecc.u-tokyo.ac.jp), Kyosuke Bunji (bunji@p.u-tokyo.ac.jp)
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Shingo Sugawara (shingo.sugawara.77@hosei.ac.jp)
Faculty of lifelong Learning and Career studies, Hosei University, 2-17-1 Fujimi, Chiyoda-ku, Tokyo 102-8160, Japan
Miwa Inuzuka (m_inuzuka@mail.tais.ac.jp)
Faculty of Humanity, Taisho University, 3-12-4 Nishi-sugamo, Toshima-ku, Tokyo 170-8470, Japan
Takuya Matsuzaki (matuzaki@nuee.nagoya-u.ac.jp)
Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
Koken Ozaki (koken@gssm.otsuka.tsukuba.ac.jp)
Graduate School of Business Sciences, University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan

Abstract

A reading skill test to diagnose basic language skills is introduced. The test is designed to measure six component skills relevant to reading in comparison with those of state-of-the-art natural language processing technologies. The results of the first large-scale experiments using the test are reported. Surprisingly, almost half of Japanese junior high school students do no better than machines in dependency analysis. More than half of 7th grade students do no better than making random choices on questions involving inferences and definition understanding.

Keywords: Reading Skills, Language Comprehension, Test Theory

1. Introduction

Artificial intelligence (A.I.) armed with machine learning technologies often surprises us by demonstrating its power. Arai et al. developed A.I. systems that were capable of passing the entrance examinations of more than half the universities in Japan (Arai & Matsuzaki, 2014). On the other hand, teachers are facing the problem that many students come into their classrooms without the requisite knowledge, skills, or disposition to read and comprehend the materials placed before them (RAND, 2002).

This situation raises a natural question. Will there be any economic returns to education when A.I. is smart enough to "learn" better than most of us? Do we have to set different goals for education in the age of A.I.?

Before jumping to any conclusions, we must carefully study the performance of human beings in comparison with those of machines, especially of the skills and expertise that are believed to be acquired only through education. Reading comprehension is, of course, one such example.

In this paper, we introduce a new reading skill test (RST) for assessing an examinee's basic language skills involved in the comprehension of texts consisting of sentences taken from junior high and high school textbooks and dictionaries. It is a major version-up from the prototype developed in (Fujita et al., 2016). A unique feature of the RST is that it is designed to analyze language skills of both human beings and machines. Consequently, the test results will tell us not only an examinee's language skills relative to others, but also to machines. It will also reveal what kinds of sentences (i.e. lexical, structural, thematic) are harder than others to comprehend (process) for human beings (for machines).

The RST contains six different types of question. The first two types are designed to measure an examinee's ability to analyze intra- and inter-sentential relations among words: dependency analysis and anaphora resolution. Statistical algorithms often achieve precisions around 80%-90% in parsing sentences and 60%-70% in anaphora resolution (Nivre et al., 2007; Pradhan et al., 2012), which indicates that not only examinees but also A.I. may be able to perform syntactic analysis of a sentence without understanding its meaning. The second two are designed to measure an examinee's inferential skills based on appropriate amounts of vocabulary and common sense. They are closely related to tasks called textural entailment recognition or synonymy recognition in the field of natural language processing, and both of them are known to be very hard (Dagan et al., 2013). The last two are designed to measure how examinees can map texts into meanings. They require high-level symbol grounding and abstract thinking, and neither a practical algorithm to solve them nor a theory to formalize them has been proposed yet.

If an examinee does equally well on the six different types of question, we can assume that he/she reads differently from machines. On the other hand, we had better doubt that an examinee reads like a machine if he/she does well on the first two types of question: he/she appears to understand the meaning of the texts, but actually may not. In other words, human-machine comparison and error analysis of machines may allow us to diagnose why many readers read poorly.

The results of the first large-scale investigation involving 1758 students from six public junior high schools are reported. Surprisingly, in a country like Japan where education is compulsory up to the end of junior high school, and which is among the top countries in PISA tests (OECD, 2016), more than half of the 7th grade students did no better

than random choice on the third two types of question. These results lend support to the concerns expressed in (RAND, 2002). The performance of a popular Japanese dependency structure analyzer on dependency analysis questions is also reported for comparison.

2. Design of RST

2.1 Six Component Skills and their Measurement We define six component skills relevant to reading. Each skill is measured separately in the RST. We do not claim that basic language skills consist exclusively of these six. We plan to add new types as necessary.

1. Dependency Analysis (DEP): The skill of recognizing the dependency relations between words and phrases in a given sentence.

2 Anaphora Resolution (ANA): The skill of anaphora resolution. ANA is comprised of two elements: Demonstrative Anaphora Recognition (DANA) and Zero Anaphora Restoration (ZANA).

DANA: The skill of recognizing the anaphoric relation between a demonstrative pronoun in a sentence and its antecedent.

ZANA: The skill of restoring and recognizing a noun phrase implicitly omitted in a context.

3. Paraphrasing (PARA): The skill of recognizing that a sentence is the same in meaning as another one. PARA is comprised of three elements which are Lexical Paraphrasing (LeP), Structural Paraphrasing (SP), and Logical Paraphrasing (LoP). The participant reads two sentences and judges whether they are synonymous. The examinees are asked to choose "Yes" or "No".

LeP: The skill of recognizing the synonymy between words or short phrases.

SP: The skill of recognizing the synonymy between two sentences written in different voices (active/passive).

LoP: The skill of recognizing logical equivalency of two sentences.

4. Logical inference (INF): The skill of reading a sentence and determining what can be inferred from a proposition in the sentence, what conflicts with it, and what does not relate to it. Here, two sentences are presented to the examinees. The instruction asks the examinees whether the proposition in the second sentence (task sentence) can be inferred from the proposition in the first sentence (presented sentence). The examinees are asked to choose "Yes" if the sentence can be inferred, "No" if the first and the second propositions cannot hold true at the same time, and "Not known" if the propositions are not related to each other.

5. Representation (REP): The skill to represent an image (figure or table) by comprehending a sentence of the textbook. The participant reads a sentence and chooses the images correctly representing the sentence out of four (multiple responses).

6. Instantiation (INST): The skill to understand how to use a term correctly according to a given definition of the term. The participant reads a definition sentence and chooses correct usages from four sentences (multiple responses).

2.2 Test settings Each RST question requires a considerable amount of concentration. We designed the RST so that examinees would not get confused or become exhausted. As a result, each examinee randomly takes three of six types of questions in the current setting. After answering two sample questions of a type, examinees are asked to answer questions randomly chosen from an item pool as precisely and quickly as possible in four minutes.

We intend to change the design of the test so that he/she takes all six types when we are ready to calculate *b*, the difficulties of the questions, and θ, the ability of the examinee in Item Response Theory (IRT; Lord & Novick, 1968; Hambleton & Swaminathan, 1985) with fewer questions.

2.3 Interface RST is conducted as a Computer Based Test (CBT) or Paper Based Test (PBT). Figure 1 shows a screenshot of an REP question. For the details of the design, the reader should refer to (Fujita et al., 2016).

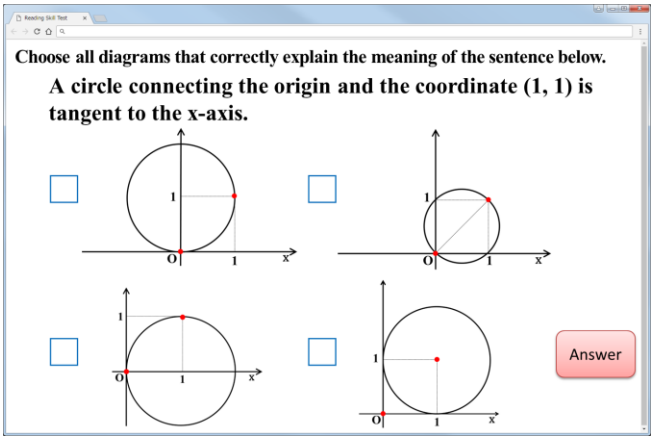

Figure 1: Question REP 39 shown in CBT

2.4 Materials We created all of the questions, except for the INST questions, on the basis of textbooks that have been approved by the Ministry of Education, Culture, Sports, Science and Technology and are being used in Japanese junior high and high schools. The INST questions were created using terms and definitions appearing either in the textbooks or in Japanese dictionaries.

3. Psychometric Properties of RST

An examinee's score is usually assessed by the sum score of all items to which he/she responded. However, because in the setting of RST, each examinee responds to different items, the sum score is not appropriate for an examinee's assessment. That is, the sum score is "item dependent", which means that the assessment result depends on the difficulties of the items that the examinee responded to as well as the examinee's characteristics.

Therefore, in this project, IRT is used for each examinee's assessment. One of the distinctive features of IRT is that it

is not item dependent. The reason is that an item's difficulty and an examinee's characteristics are treated as different parameters. An item *j*'s difficulty parameter is denoted as b_j. The higher b_j is, the more difficult the item is. An examinees *i*'s characteristic is denoted as θ_i. The higher θ_i is, the better the examinee's characteristic is, which is reading skill in this study. For the details of IRT, the reader is referred to the above references.

In the near future, we will start computerized adaptive testing (CAT, van der Linden and Glas, 2010). In CAT, each examinee answers items shown on a PC display or tablet. If the examinee correctly answers an item, the next item is more difficult, whereas if he/she incorrectly answers an item, the next item is easier. Note that CAT requires an item pool, which is a set of items whose item parameters have already been estimated. In CAT, an appropriate item for each examinee is selected from the item pool. Therefore, IRT is suitable for the CAT framework. This is another reason why IRT is used in the analysis.

The R software (version 3.1.0) was used to fit the IRT model. Estimations were performed for each component. Therefore, if an examinee took all six different types of tests, he/she would have six θ values.

Before going to the next analysis where θ is used, inappropriate items were detected and deleted and the IRT analysis was done once more. Inappropriate items were detected using item analysis, in particular, a trace line plot.

Figure 2 shows trace line plots of appropriate (left) and inappropriate (right) items. The horizontal axis of this figure is θ. All the examinees who responded to these items were divided into four groups in accordance with θ. The vertical axis of the figure is the ratio of the examinees who selected options 1 to 4 for each θ group. For both items, option 2 (bold line) is the correct one. Note that "s" in this figure means 'skipped the item'.

The left item is appropriate because the higher θ is, the higher is the rate of the examinees correctly answering the item. This item will be examined in detail in the Results section. On the other hand, the right item is inappropriate because the higher θ is, the lower is the rate of the examinees correctly answering the item. Therefore, the right item was deleted.

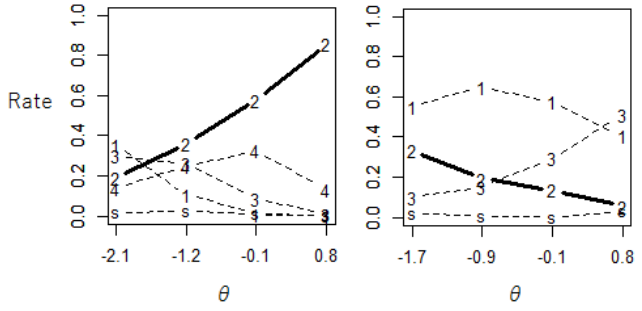

Figure 2: Two trace plots

The three deletion criteria described below were applied to items responded by more than one hundred examinees. Items applied to more than one criterion were deleted.

1. The rate of the selecting correct option is almost one hundred percent for all of the four θ groups.
2. The higher θ groups do not have higher rates of selecting the correct option (right of Figure 2).
3. The highest θ group is most likely to select an incorrect option (right of Figure 2).

Table 1 shows the numbers of deleted items and the numbers of remaining items.

Table 1: The number of deleted and the number of remaining items

	DEP	ANA	PARA	INF	REP	INST
Deleted	13	0	0	1	10	28
Remained	121	37	36	34	35	85

To examine the validity, reliability, and one-dimensionality of each test, correlations between the six θs, ω coefficients (McDonald, 1999) and the factor loadings in categorical factor analysis were estimated. Table 2 shows the results. Most of the correlations between the six θs are above 0.5, which means that the six tests all measured different aspects the same trait (reading skill). This shows that the tests have enough validity. Moreover, all six ω coefficients are very high, which shows that the tests have enough reliability. Finally, the means of the factor loadings are not small, which shows the one-dimensionality of each test, which is required in IRT.

Table 2: Correlations, omega coefficients, and mean of the factor loadings

	DEP	ANA	PARA	INF	REP	INST
ANA	0.657					
PARA	0.652	0.455				
INF	0.541	0.575	0.541			
REP	0.515	0.534	NA	0.620		
INST	0.178	0.354	0.607	0.126	NA	
ω	0.990	0.963	0.962	0.967	0.927	0.964
mean loadings	0.644	0.620	0.610	0.421	0.489	0.589

Note: NA means that the correlation could not be calculated because there was no examinee who took the two tests.

4. Related work in cognitive science

To answer the RST items, examinees need to parse sentences with unfamiliar content. In this situation, the literature suggests that human parsers tend to make errors with ambiguous sentences (Frazier & Rayner, 1982). On the other hand, readers can construct coherence between sentences through automatic inferences (McKoon & Ratcliff, 1992). However, studies on the human parsing process are mainly based on data collected from adult readers. Some studies suggest that there are different characteristics in the sentence processing of younger children (Otsu, 1994) and older adults (Baota et al., 2001), but there seems to be no evidence on sentence processing of young students. Moreover, despite that some school teachers recognize the possibility that the difficulties in parsing and building coherence between sentences are larger

than expected, achievement tests remain mainly concerned with higher levels of discourse.

5. Results of junior high school students

5.1 The appropriateness of RST Response data from six public junior high schools' students were analyzed, to show the appropriateness of the RST. These schools are in City A, whose schools are known to perform well (the best in the prefecture in 2016) in national standardized achievement tests. The responded included 613 students in grade 7, 537 in grade 8, and 608 in grade 9. The students responded to questions (items) taken from sentences from junior high and high school textbooks and from Japanese dictionaries.

The analyses calculated two statistics: the Correct Answer Rate (CAR) and the Rate of Students who may respond by Guessing (RSG). CAR is the percentage of items that a student correctly answered, while RSG is the rate of students who were not statistically significant in a one-sided hypothesis test assessing whether each student's correct answer rate is greater than that by guessing (null hypothesis). For example, in the PARA test, whose items have two alternatives, the expected correct answer rate by guessing is 0.5.

First, we calculated CARs for each student in the six component tests. Although each examinee responded to different items as noted above, because these items were selected randomly, the CARs can be assumed to be comparable. The mean CAR was calculated for each grade (Table 3).

Table 3: CAR means of each grade in the six component tests

Grade	DEP	ANA	PARA	INF	REP	INST
7	0.613	0.611	0.728	0.548	0.278	0.247
8	0.646	0.653	0.746	0.576	0.303	0.281
9	0.703	0.739	0.798	0.621	0.384	0.383

Table 3 indicates that in all the component tests, as the grade goes up, the mean CAR also increases. Generally speaking, reading skills improve as the grade goes up.

Table 4: Means of θ and RSG of each component skill in each grade

	Means of θ			RSG		
Skills	grade 7	grade 8	grade 9	grade 7	grade 8	grade 9
DEP	-0.595	-0.502	-0.295	0.376	0.302	0.188
ANA	-0.558	-0.425	-0.106	0.365	0.260	0.110
PARA	-0.551	-0.440	-0.228	0.107	0.069	0.020
INF	-0.470	-0.443	-0.200	0.660	0.531	0.423
REP	-0.450	-0.436	-0.103	0.522	0.339	0.255
INST	-0.154	-0.072	0.232	0.583	0.505	0.312

Next, to examine the relationships between the six component skills and grades, the means of six θs in each grade and the RSGs for each grade were calculated (Table 4). Including city A's, we collected responses from more than 13000 participants, which are elementary-school students to adults. The θs were estimated using the responses of all the examinees and the mean of the θs was set to 0 for all six components. The means of the six θs for the junior high school students therefore tend to be negative in this table. The table shows that like CARs, for all component skills, as the grade goes up, the means of θ also increase and the RSGs decrease.

Finally, to determine whether the relationships between six component skills, RSGs, and grades differ among schools, we calculated the means of the six θs and the RSGs of each grade in the six schools. The results indicate that the six junior high schools showed almost all the same tendencies as Table 4. That is, in all schools, as the grade goes up, the means of the six θs tended to increase and the RSGs of the tests decreased.

All these results are evidence of the validity of the test.

5.2 Assessment of students' reading skills

It is a good sign that RSGs decrease as the grade goes up. However, the RSGs of the 7th grade students on INF, REP and IST exceeded 50%. In other words, more than half of them failed to make inferences correctly based on the knowledge given in the textbooks, map the texts into the correct images, or understand the definitions. Our statistics show that at least one fourth of students graduate from junior high school without the ability to read and comprehend textbooks at a level better than guessing. As far as we know, this is the first large-scale investigation revealing this inconvenient fact.

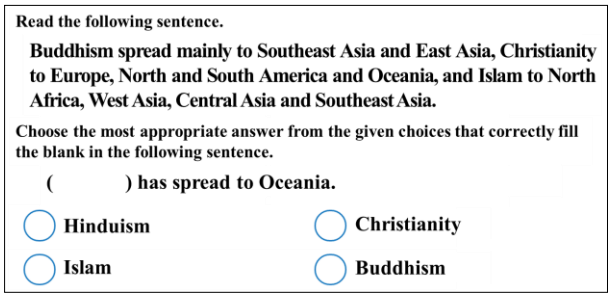

Figure 3: Question DEP 103

Now, let us examine three items as to whether or not the items were tricky or too difficult for them to answer (Table 5). In DEP 103, given in Figure 3, one can choose the correct answer, Christianity, without knowledge of the four religions. Figure 2 shows the trace plot of DEP103. It shows the item was neither tricky nor inappropriate. Still, about 40% of 7th graders, 50% of 8th graders, and 33% of 9th graders were not able to choose the correct answer.

Table 5: Percentage of correct answers to the three questions for each grade

Question	grade 7	grade 8	grade 9
DEP103	0.609	0.516	0.676
REP39	0.070	0.281	0.298
REP38	0.250	0.419	0.492

Moreover, all the 8th grade students had learned the words appearing in REP39, in Figure 1, (i.e., circle, origin, x-axis, tangent to) in the 7th grade. The gap between the CARs of the 7th and 8th grades (0.070: 0.281) might be explained by the unfamiliarity of these words to the 7th graders. Then, how can we explain that only 28.1 percent of the 8th grade students were able to choose the correct image of the text?

One may explain that unskillful readers fail to monitor when they are checking more than one condition. Here, we asked the following simpler question as REP38: "The circle passes through the origin O". The gap between CARs of REP 38 and 39 of the 8th and 9th grades might be explained by monitoring failure. Still half of the 9th grade students failed to answer correctly. We could not find relevant literature to explain this phenomenon.

5.3 Correlation with schools' characteristics We calculated correlations between these statistics and the schools' characteristics; Distances from the nearest station (Dis), the Number of Students (NS), and Rates of Students receiving Financial Help for school attendance (RSFH) in each grade (Table 6).

Table 6: Correlations of means of θ and RSG with school characteristics

	Means of θ			RSG		
Skills	Dis	NS	RSFH	Dis	NS	RSFH
DEP	-0.534	0.302	-0.540	0.434	-0.324	0.491
ANA	-0.315	0.104	-0.451	0.098	0.018	0.359
PARA	-0.294	0.101	-0.313	0.236	-0.105	0.412
INF	-0.262	0.251	-0.288	-0.001	-0.114	-0.016
REP	-0.235	0.143	-0.291	0.347	-0.209	0.408
INST	-0.156	0.279	-0.310	0.160	-0.345	0.237

Table 6 shows that in all the tests, the correlations of the means of θ with Dis and RSFH are negative, but positive with NS, and that in almost all of the tests, the correlations of RSG with Dis and RSFH are positive, but negative with NS. These results imply that students whose schools are near a station, are large, and offer less financial support tend to have higher component skills and therefore may respond to items not by guessing. We will continue to investigate these findings.

We asked examinees to answer a questionnaire including items on their attitudes toward reading and likes and dislikes of school subjects. City A conducts standardized achievement tests every year. We are planning to assess the relationship between the results of the RST, the responses to questionnaires and the scores of the achievement tests.

6. Comparison of performances with automatic dependency structure analyzer

We processed the test sentences of the RST dependency analysis questions (DEP) with the CaboCha parser (Kudo & Matsumoto, 2002) and analyzed the errors. We hoped that the analysis of errors made by a machine would help us to understand the human errors. CaboCha is a dependency parser based on Support Vector Machine. It was trained only on a news corpus, and its accuracy on news text is around 90% at the dependency relation level and 50% at the sentence level. The comparison with the human responses provided here is hence preliminary in that we expect the parser's accuracy will improve by retraining it on textbook data.

We analyzed the items on which we collected the responses from more than 100 students. DEP is a set of multiple-choice questions that ask for a phrase that stands in a certain grammatical relation to a phrase in a test sentence. We chose the answer based on CaboCha's output. The rate of correct answers by CaboCha was 66%. For example, CaboCha parsed DEP103 (Figure 3) correctly.

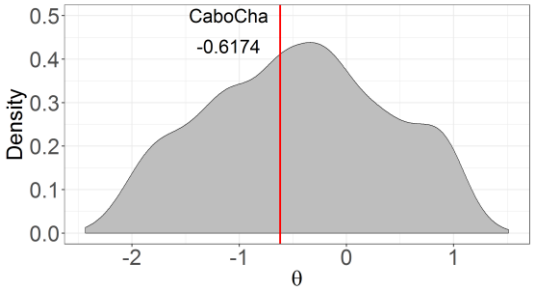

Figure 4: DEP θ value of humans and CaboCha

Figure 4 shows the distribution of human θ and the estimated θ of CaboCha on the DEP questions. It reveals the mode of human θ is only slightly above that of CaboCha.

The most common error types made by CaboCha were as follows (the numbers in parentheses are the fractions of the errors of these types).

1. When the test sentence includes a phrase inside parentheses (7%)
2. When the sentence is long (11%)
3. Unusual use of a comma or no use of comma (2%)
4. Choice of the attachment site of a subordinate or parallel verb phrase (60%): CaboCha made mistakes most frequently on the sentences including more than one subordinate or parallel verb phrase (VP). It corresponds to a sentence in the form of "… Verb … VP1 … VP2 …" in English, where VP2 has two possible attachment sites, Verb (matrix verb) and VP1 (another subordinate VP), as in:

   Adaptive immunity [$_{\text{Verb}}$ includes] humoral immunity in which B cells [$_{\text{VP1}}$ form proteins called antibodies] [$_{\text{VP2}}$ to remove extracellular pathogens], and …(snip).

   There is no syntactic clue to choose between the two possibilities. Thus, it should be judged by meaning, and hence, it is difficult for CaboCha.
5. Wrong word segmentation (5%)

In Japanese, words are not separated by whitespaces as in English. CaboCha often fails to segment technical terms correctly.

The errors of type 1, 3, and 5 would be reduced by retraining the parser on textbook data. On the other hand, the errors of type 4 require context and meaning to fix them. Table 7 lists the rate of correct answers by the human examinees on the questions on which CaboCha made mistakes. It suggests the choice of subordinate or parallel VP attachment is also difficult for humans. While Table 3 indicates that students gradually acquire the skill and knowledge to do it, it would remain a hard problem for an automatic parser since it requires some understanding of the meaning and context of a sentence.

We would like to confirm and extend these findings by examining more diverse samples collected through RST. Of special interest is a further analysis of the errors of human and automatic parsers on the basis of the cognitive studies on sentence processing（Mitchel 1994）such as the garden-path theory (Frazier & Fodor, 1978; Frazier & Rayner, 1982) and minimalist hypothesis (McKoon & Ratcliff, 1992).

Table 7: Human CARs on the questions on which the automatic analyzer made mistakes

Error type of CaboCha	Human CAR
Parenthesized phrase	0.584
Long sentence	0.572
Unusual use or no use of comma	0.615
Attachment of subordinate VP	0.549
Word segmentation	0.786

7. Conclusion

We developed a new reading skill test (RST) to measure six component skills relevant to reading. By analyzing the responses to the RST, we confirmed that it has enough reliability and validity. In addition, we analyzed response data of Japanese junior high school students to the RST, and the results implied that, surprisingly, the six component skills might be lower than expected. Finally, we compared the performances of the students with those of a Japanese dependency parser. The results implied that students do no better than a machine in dependency analysis.

References

Arai, H. N., & Matsuzaki, T. (2014). The impact of A.I. on education - Can a robot get into the University of Tokyo?. *Proceedings of the 22nd International Conference on Computers in Education* (pp. 1034-1042).

Balota, D. A., Cortese, M. J., & Wenke, D. (2001). Ambiguity resolution as a function of reading skill, age, dementia, and schizophrenia: The role of attentional control. In Gorfein, D. S. (Ed). *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 87-102). Washington, DC, US: American Psychological Association.

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. 2013. Recognizing Textual Entailment: Models and Applications. Morgan & Claypool.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6(4)*, 291-325.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178-210.

Fujita, A., Todo, N., Sugawara, S., Kageura, K., & Arai, N. H. (2016). Development of a Reading Skill Test to Measure Basic Language Skills. *Proceedings of the 8th IEEE International Conference on Technology for Education* (pp.156-159).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhof.

Kudo, T., & Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. *Proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 63-69). Association for Computational Linguistics.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates, Mahwah, NJ.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99, 440-466.

Mitchell, D. C. (1994). Sentence parsing. *Handbook of psycholinguistics*, 375-409.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. (pp. 915-932).

OECD (2016). *PISA 2015 Results in Focus*. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Otsu, Y. (1994). Early acquisition of scrambling in Japanese. In: Teun Hoekstra & Bonnie D. Schwartz (eds.) *Language Acquisition Studies in Generative Grammar*, 253-264. Amsterdam: John Benjamins Publishing.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. *Proceedings of the Joint Conference on EMNLP and CoNLL – Shared task*. (pp. 1-40).

RAND Reading Study Group (2002). Reading for understanding: Toward an R&D program in reading comprehension. Santa Monica, CA: RAND Education.

van der Linden, W. J., & Glas, C. A. W. (eds.). (2010). *Elements of Adaptive Testing*, New York, NY: Springer.