Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Application of Speech Recognition in a Japanese Dictogloss System

Satoru KOGURE[a*], Toshiaki NAKAHARA[b], Yasuhiro NOGUCHI[a], Tatsuhiro KONISHI[a], Makoto KONDO[a] & Yukihiro ITOH[c]
[a]*Faculty of Informatics, Shizuoka University, Japan*
[b]*Graduate School of Integrated Science and Technology, Shizuoka University, Japan*
[c]*Shizuoka University, Japan*
*kogure@inf.shizuoka.ac.jp

Abstract: A learning method called dictogloss is focused in second language learning. We have constructed a dictogloss environment that enables self-directed study for second language learners. In the present study, we focused on the speaking skill, which was not supported in our previous dictogloss environment. We thus added to the system "pronunciation evaluation of the learner's reconstructed sentences which they read aloud to the system" and "error identification conducted by the learner of the CLA's reconstructed text" by utilizing speech recognition technology. In the present study, we implemented the proposed system and evaluated its effectiveness using participant questionnaires and interviews.

Keywords: Second language learning, dictogloss, speech recognition

1. Introduction

One such potential learner support system, which has been receiving increasing amounts of attention from researchers' second language education, is the dictogloss method. Dictogloss is a learning method in which learners can cooperatively learn the four language skills of listening, speaking, reading and writing (Wajnryb, 1988, 1990). The activities of dictogloss include the following three stages: (1) the dictation stage, (2) the reconstruction stage, and (3) the evaluation and feedback stage. Learners first listen to a short text read by the teacher in the dictation stage several times while taking notes. Next, the learners work collaboratively in small groups to reconstruct the text from their note in the reconstruction stage. Finally, the teacher evaluates the learners' reconstructed text and provides feedback. Several previous studies have focused on the effect of using the dictogloss activity on each language skill (Jibir-Daura, 2013; Sari Dewi, 2014; Lindstromberg et al., 2016).

However, dictogloss on its own is not suitable for self-directed study because learners must work collaboratively in groups and the teacher must be involved in the first and final stages of the activity. Therefore, we have constructed a dictogloss environment that enables self-directed study for second language learners (Kogure et al., 2015; Kogure et al., 2016; Kondo et al., 2012; Tashiro et al., 2013). In the dictogloss environment, we created a cooperative learner agent (CLA) and teacher agent (TcA) in the system, where the learner inputs the reconstructed text into a computer. However, the system does not cover the speaking skill. The learners compare the reconstructed texts written by the CLA and themselves and if the learner determines that the reconstructed text is wrong, he or she clicks on the wrong word in the reconstructed text written by the CLA. The system automatically generates a message to the CLA identifying the mistakes (Kondo et al., 2012; Tashiro et al., 2013). Learners can also explain to the CLA why they think the word is wrong (Kogure et al., 2015). This environment has about 330 minutes learning contents from 22 different texts (Kogure et. al., 2017).

In the present study, we focused on the speaking skill, which was not supported in our previous dictogloss environment. We thus added to the system (a) pronunciation evaluation of the learner's reconstructed sentences (which they read aloud to the system) and (b) error identification conducted by the learner of the CLA's reconstructed text (utilizing speech recognition technology).

In the present study, we implemented the proposed system and evaluated its effectiveness using participant questionnaires and interviews.

2. The Existing Japanese Dictogloss Environment

Figure 1 displays a screenshot of the existing Japanese dictogloss environment (Kogure et al., 2017).

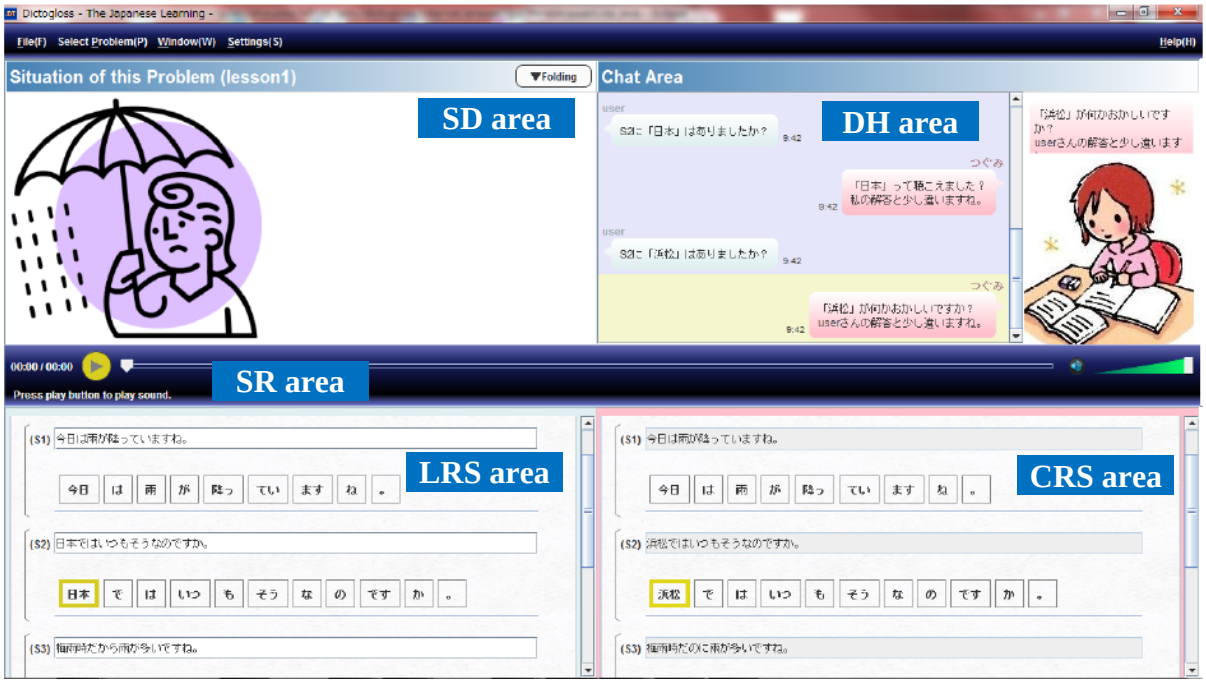

Figure 1. Screenshot of the Existing Japanese Dictogloss Environment

In Figure 1, the upper left part (SD area) shows the situation diagram and the upper right part (DH area) shows the dialogue history with the CLA. The middle part (SR area) shows the speech reproduction interface of the task sentences, the lower left part (LRS area) shows the reconstructed sentences input by the learner, and the lower right part (CRS area) displays the CLA's reconstructed sentences. The rough flow of learning is as follows. First, learners click the play button in the SR area to listen to the task sentences. Learners can listen to them up to five times. Based on the heard speech, the learners then reconstruct the sentences using the form in the LRS area. Next, the system evaluates whether the learner's reconstructed sentences are correct. Based on the evaluation results and the educational procedure outlined in Table 1, the system generates the CLA's reconstructed sentences and displays the sentences in the CRS area. The learner compares their own reconstructed sentences with the CLA's reconstructed sentences. If they find a different word, they click on the word that they think is wrong in the CRS area, and the system generates a learner's question that asks the CLA whether the clicked word was correct and displays the sentences in the DH area. Next, the system generates a CLA utterance that responds to the learner's indication based on the correctness of the learner and CLA's words, and the system displays this utterance in the DH area. The system generates different texts depending on which of the four categories; focused language forms (LFs), keywords, other LFs, and the other forms. See Kondo et al. (2012) for further discussion.

In some LFs, learners can point out the correctness of words through their explanations. We classified the explanations into three types: grammatical explanations, contextual explanations, and situational explanations. A grammatical explanation is one that the learner can explain in terms of grammatical rules concerning the correctness of a word. A contextual explanation is one in which the learner cannot explain the correctness of a word except in terms of its surrounding sentences. A situational explanation is one that the learner cannot explain except in terms of situational knowledge that does not appear in the reconstructed sentences. To scaffold for grammatical

explanations, a teacher defines the basic knowledge required to do so in advance. The knowledge is associated with an explanation's attribute and value. The teacher also predefines peripheral explanatory knowledge.

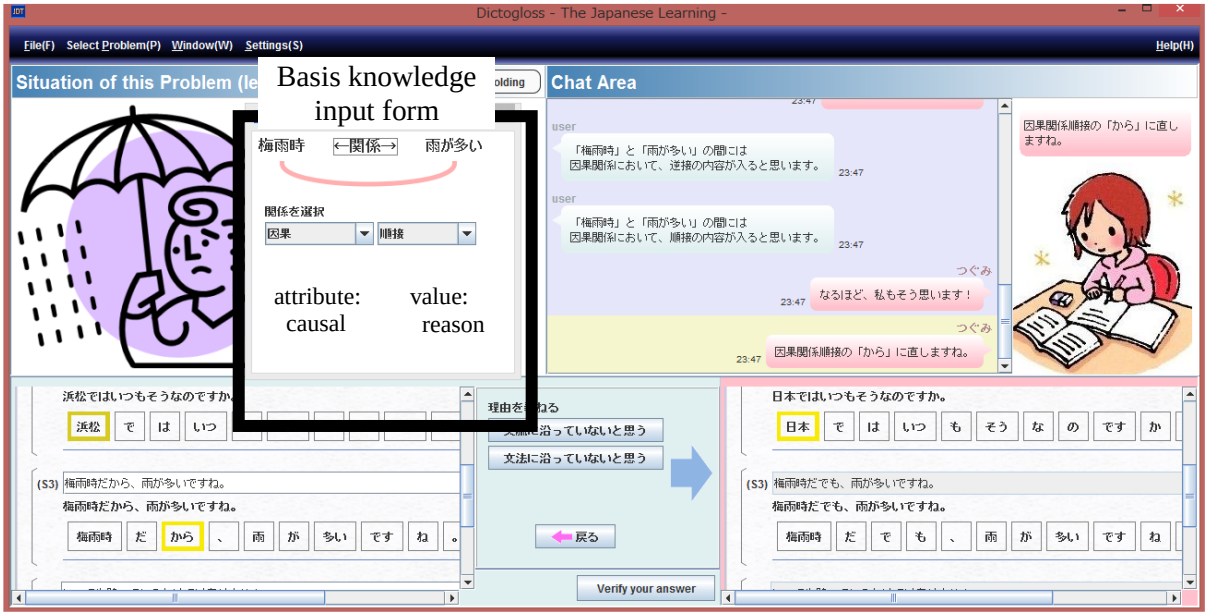

Figure 2. The Screenshot of Discussion for Conjunction Correctness

Figure 2 shows a screenshot of a discussion of conjunction correctness with the CLA. For the situation, the teacher focuses on the conjunction form (causal relationship) as an LF. In S3, the learner correctly reproduces the text and the CLA generates incorrect text following the procedure described in Table 1. If the learner clicks the conjunction word *kara* (lower yellow box in the LRS area), the system makes the learner provide an explanation (peripheral knowledge) for why his or her own answer is correct. If a learner selects a contextual explanation, the agent requires the learner to indicate the basic knowledge. Regarding basic knowledge, the learner selects a basis attribute and a basis value. If the learner selects a correct attribute and value, the system generates the user utterance, "I think there is a word indicating the reason for the causal relationship between *rainy season* and *it rains every day*." Then, the CLA tries to correct its own reconstructed text.

3. Utilizing Speech Recognition

The act of speaking involves three processes; (a) thinking about what you want to say, (b) thinking how to say, and (c) actually saying it. In the initial stage of language learning, learners find process (b) to be particularly challenging. In a typical preliminary speaking lesson, the learner intensively practices process (c). In our proposed system, as with the typical method, we will adopt a learning method that the learner first speaks in process (c). In this study, we focused on two points. The first was pronunciation evaluation by reading out the learner's reconstructed text and the second was oral error identification of the CLA's reconstructed text utilizing speech recognition technology.

3.1 Pronunciation Evaluation

There have been many studies on pronunciation learning or pronunciation evaluation methods (Strik & Cucchiarini, 1999; Terguieff, 2012; Kasahara et al., 2014). In pronunciation evaluation, a teacher evaluates the learner's pronunciation based on four aspects; incorrect vowels/consonants, incorrect beat/rhythm, incorrect accent, and incorrect intonation.

It can be difficult for beginners to pronounce vowels in their target language because the number of vowels varies depending on the language; moreover, there are silent vowels in some languages. For example, Arabic has three vowels but Japanese has five vowels. Japanese learners

whose native language is Arabic often find it hard to distinguish between '*i*' and '*e*', and '*u*' and '*o*'. Consonants are similar to vowels. For example, learners whose native language is Korean may find it difficult to pronounce the consonants 'za, zi, zu, ze, and zo', as these sounds do not exist in Korean. Thus they may replace them with 'ja, ji, ju, je, and jo'. Regarding beat/rhythm, learners make acoustic mistakes with long vowels, syllabic nasal sounds and double consonants in particular. In Japanese, words corresponding to syllable strings depend on the accent used. It is difficult for beginners to accurately understand differences in meaning created by accent. For example, the syllables 'ni-ho-N' means *Japan* if the set of pitch used are low, low, and high, but it means *two* if the set of pitch used is high, low, and low. It is also difficult for beginners to distinguish the meanings of sentences based on intonation. The difference in intonation affects syntax or semantic analysis. Consider a Japanese sentence 'ki no u na ku shi ta ka gi ga mi tsu ka ri ma shi ta' (I found a lost key yesterday). If we pronounce it '[high] ki no u na ku shi ta ka gi [low] ga mi tsu ka ri ma shi ta', a word 'ki no u' (yesterday) modifies `na ku shi ta' (lost). On the other hand, if we say 'ki no u [high] na ku shi ta ka gi [low] ga mi tsu ka ri ma shi ta', 'ki no u' modifies 'mi tsu ka ri ma shi ta' (found). The [high] and the [low] labels indicate changes in intonation.

In the present study, in the interest of simplifying implementation, we focused on a framework that evaluated pronunciation based on errors of incorrect vowels/consonants, and incorrect beat/rhythm.

3.2 Usage Points of Speaking in the Dictogloss

We propose the addition of speaking practice at two points in the dictogloss activity; (1) reading one's own reconstructed text in the reconstruction stage, and (2) reading out word errors to the CLA in the reconstruction stage. Both emphasize the speaking skill (i.e., process c). At the first point, learners read aloud their own reconstructed text that they input using the keyboard. At the second point, learners are required to think about how to identify and explain errors in the CLA's text (i.e., process b). Therefore, by presenting the template to learners, we allow them to focus on speaking (i.e., process c).

4. Design and System Implementation of a Dictogloss Support Environment Using Speech Recognition

First, we discuss the speech recognition function for learners' reconstructed texts. After the learner inputs the reconstructed sentences in the reconstruction stage, before engaging in the dialogue with the CLA, the system asks the learner to read their reconstructed sentences aloud. The system then activates the speech recognizer. When the recognition of the learner's speech is finished, the recognition result is retained. To evaluate the learner's pronunciation, we registered both correct and incorrect pronunciation in the vocabulary dictionary used for speech recognition. The feedback for the actual pronunciation errors is provided in the evaluation and feedback stage. We programmed the system to disable the reading function when the learner's reconstructed sentence is far from the correct text. In the speech recognizer we used, we needed to describe "word notation" and "syllable string corresponds to reading" in the vocabulary dictionary. We entered the "corresponding correct pronunciation" in the word notation of the vocabulary dictionary for erroneous pronunciation when registering incorrect pronunciation. The tendency of pronunciation error differs depending on the learner's native language. With reference to some Japanese language-learning research (Aida et al., 1997) and Japanese learning textbooks (The Japan Foundation, 2012, 2014), we prepared a pronunciation dictionary for learners with Korean as their native language.

Second, we discuss the oral error identification function for the CLA's reconstructed test. In our previous study, when the learner identified the error in the CLA's reconstructed sentence, the learner clicked on the wrong word of the CLA. In the proposed system of the present study, the learner identifies the CLA's error through speech. We provided the following template to the learner; "*Did you have* ☐ *in S*☐*?"* Enter "incorrect word" in the first box and "sentence number" in the second box. We added a button for the leaner to click to start the oral error identification function. The system activates the recognizer when the button is clicked. Then, the system interprets

the recognition result according to the template, and performs the identification function of the wrong word.

Third, we discuss a feedback function for pronunciation errors. The system provides feedback using the results recognized during the reconstruction stage. Since the recognition result recognized by "wrong pronunciation" includes "pronunciation of corresponding correct answers," the system can return feedback according to the following template; *I heard it pronounced WORD in S as "error pronunciation". The correct pronunciation is "correct pronunciation". Let's check it again.* When the confidence measure of the recognition result is low, the system generates a feedback sentence using the following template; *It may have been pronounced WORD in S as "error pronunciation". Please check "correct pronunciation" of correct pronunciation.*

For speech recognition, we used the module mode (server) of the large vocabulary continuous speech recognition engine *julius* (Lee & Kawahara, 2009). *Julius* has a grammar version that defines the context free grammar (CFG) and a statistical language model version that uses statistical information from the corpus. We used the grammar version. We inserted a T label for correct word pronunciation and an F label for erroneous word pronunciation (e.g., a dictionary for *natsu* (summer in English) has three pronunciations; `natsu_T_na-tsu n a ts u`, `natsu_F_na-chu_na-tsu n a ch u`, and `natsu_F_na-su_na-tsu n a s u`).

5. Experimental Evaluation

First, we conducted an experiment to evaluate the interface of the system. We evaluated our system by asking four Japanese university students to participate in the present study. The subjects conducted three tasks using the system. The subjects pretended to be international students during the experiment. The system usage time was about 1 hour. After completion of the experiment, the subjects answered a questionnaire asking about the usability of the interface, reading out the learner's reconstructed sentences and the interface's provision of feedback and pronunciation evaluation. We obtained the results of 1.75 points for the interface of reading reconstructed sentences and 4.25 points for the feedback interface (possible scores ranged from 1 to 5 points). The evaluation of the interface for reading sentences was low. Therefore, we interviewed the subjects individually and improved the interface accordingly. We improved two points of the interface; the first improvement is that ``adjusting the window display position so that the on/off state of speech recognition can be confirmed easily,'' and the second one is that ``relocating the reconstructed sentence reading button to where it could more easily be clicked.''

Second, we conducted experiments to evaluate the learning system and evaluate incorrect pronunciation. We conducted the same experiment outlined above with a Japanese-language learner whose native language was Korean. After completion of the experiment, the subject also completed the questionnaire asking about the usability of the interface, reading the reconstructed sentences, and the interface's provision of feedback and pronunciation evaluation. We obtained the results of 5 points for the interface of reading reconstructed sentences, and 4 points for the feedback interface (possible scores ranged from 1 to 5 points). In addition, we asked the subject two more questions. The first question was 'do you think that reading the reconstructed sentences out loud will lead to an improvement in learners' speak?'. The second question was 'do you think that the feedback provided by the system on pronunciation can help to improve learners' speaking?'. We obtained a full score of 5 points for both questions. Finally, we also asked the subject to describe freely any additional features that would be beneficial for the system. We received answers such as "I want to be able to listen to my own utterances" and "I want assistance with the words I cannot read.'

We also obtained results regarding the Korean participant' recognition rate at false pronunciation. For three lessons, we prepared eight words with a dictionary for erroneous pronunciation. The system required a subject to pronounce all eight words. The discrimination rate was 62.5%: a value obtained by dividing the sum of the number of correctly identified incorrect pronunciations (2 times) and the number of correctly identified correct pronunciations (3 times) were correctly uttered by the total number (8 times).

We did not obtain a high discrimination rate. In the future we will investigate various identification technologies for pronunciation evaluation and implement them in the system.

6. Conclusion

We examined ways to utilize speech recognition in an existing Japanese language dictogloss training environment. We implemented the functions of reading reconstructed text, identifying and explaining errors in the CLA's text, and evaluating the pronunciation of the reconstructed sentences. We conducted an evaluation experiment and obtained a favorable evaluation from a Korean Japanese-language learner. On the other hand, we found the discrimination rate of the pronunciation evaluation needs improvement.

In the future, we will implement a Japanese dictogloss training environment focusing on speaking and listening abilities. We will also consider how to integrate accent and intonation pravice into the system, which we did not implement this time.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 17K00483 and 26350273.

References

Aida, K., Lee, K, & Shirai, K. (1997). Pronunciation CAI System Based on Japanese Speech Recognition, *Bulletin of Waseda University Japanese Research and Education Center, 9*, 111-131 (in Japanese).

Basterrechea, M. & García, M. & Lesser, M. (2014). Pushed Output and Noticing in a Dictogloss: Task Implementation in the CLIL Classroom. Porta Linguarum.

Jibir-Daura, R. (2013). Using dictogloss as an interactive method of teaching listening comprehension. *Advances in language and literary studies, 4*(2), 112-116.

Kasahara, S., Kitahara, S., Minematsu, N., Shen-P., H., Makino, T., Saito, D. & Hirose, K. (2014). Improved and robust prediction of pronunciation distance for individual-basis clustering of World Englishes pronunciation, *Proceedings of ICASSP*, 3240-3244.

Kogure, S., Miyagishima, K., Noguchi, Y., Kondo, M., Konishi, T. & Itoh, Y. (2016). A teachable agent for the Japanese dictogloss learning support environment. *Proceedings of ICCE2016*, 88-90.

Kogure, S., Okugawa, K., Noguchi, Y., Konishi, T., Kondo, M. & Itoh, Y. (2017). Improvement of the Situational Dialog Function and Development of Learning Materials for a Japanese Dictogloss Environment, *Proceedings of ICCE2017*, 104-106.

Kogure, S., Tashiro, A., Noguchi, Y., Kondo, M., Konishi, T. & Itoh, Y. (2015). An answer support environment based on grammar, context and situation for a dialogue to learner agent on Japanese dictogloss system. *Proceedings of ICCE2015*, 94-96.

Kondo, M., Sano, R., Tashiro, A., Noguchi, Y., Kogure, S., Konishi, T. & Itoh, Y. (2012). Development of a dictogloss system oriented for focus on form. *Proceedings of ICCE2012*, 1-8.

Lee, A. & Kawahara, T. (2009). Recent Development of Open-Source Speech Recognition Engine Julius *Proceedings of APSIPA ASC*.

Lindstromberg, S., Eyckmans, J. & Connabeer, R. (2016). A modified dictogloss for helping learners remember L2 academic English formulaic sequences for use in later writing. *English for specific purposes*, *41*, 12-21.

Sari Dewi, R. (2014). Teaching writing throught dictogloss. *Indonesian Journal of English Education*, *1*(1), 66-76.

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature, *Speech Communication, 29*, 225-246.

Tashiro, A., Noguchi, Y., Kogure, S., Kondo, M., Konishi, T. & Itoh, Y. (2013). Evaluation of an improved dictogloss system oriented for focus on form. *Proceedings of ICCE2013*, 110-114.

Terguieff, E. (2012). The English Pronunciation Teaching in Europe Survey: Finland. *Journal of Applied Language Studies, 6(1)*, 29-45.

The Japan Foundation, (2012). Education teaching to talk (2nd edition), *Hitsuji Shobo* (in Japanese).

The Japan Foundation, (2014). Education teaching speech (4-th edition), *Hitsuji Shobo* (in Japanese).

Wajnryb, R. (1988). The dictogloss method of language teaching: A Text based communicative approach to grammar, *English teaching forum, 26*, 35-38.

Wajnryb, R. (1990). Grammar Dictation. *Oxford: Oxford University Press.*