

The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques

Emi Izumi^{1,3} Toyomi Saiga^{1,2} Thepchai Supnithi^{1,2,4}

Kiyotaka Uchimoto¹ Hitoshi Isahara^{1,2,3}

Communications Research Laboratory, Japan¹

Telecommunications Advancement Organization of Japan²

Kobe University, Japan³

Information Technology R&D Division, NECTEC, Thailand⁴

1. Introduction

To keep up with the information-driven society, it must be one of the most important things to acquire foreign languages, especially English for international communications. In order to develop a computer-assisted language teaching and learning environment, we have been compiling a large-scale speech corpus of Japanese learner English, which provides a lot of useful information to construct a model of the developmental stages of Japanese learners' speaking ability. In this paper, first, we will present the overview of this project by introducing the activities done so far such as its data collection procedure, annotation schemes including error tagging, and the development of the original software tool which makes data collection process easier. Secondly, we will explain how this corpus can be exploited for second language acquisition research or system development by introducing several experiments we have done so far, such as a linguistic analysis on how the English article system is acquired by Japanese learners of different proficiency levels and the attempts to automatic processing of learners' language in collaboration with Natural Language Processing (NLP) techniques.

2. SST Corpus

First, we would like to give some details about our corpus named "SST Corpus". We will do this first by explaining what the Standard Speaking Test (SST), which is the main speech resource of this corpus, is like and showing the data collection procedure such as transcribing, and tagging, then by introducing our original software tool, named the "TagEditor", which was developed for transcribing and tagging this corpus.

2.1 The characteristics of SST corpus

The main characteristics of the TAO corpus are:

- The corpus data is entirely based upon the audio-recorded data of the interview test, the "Standard Speaking Test".
- The interviewees are Japanese learners of English.
- The data is divided into nine groups depending on the learner's proficiency levels based on the SST evaluation scheme.
- Compared with other learners' speech corpora, the scale of the corpus is relatively large (300 hours of data, totaling one million words)

2.2 The Standard Speaking Test (SST)

Now, we would like to give a general description of the SST. This 15-minute interview test was jointly developed for Japanese learners by the American Council on the Teaching of Foreign Languages (ACTFL) and a famous Japanese ELT publishers, ALC Press. This test is based on the ACTFL's Oral Proficiency Interview (OPI) exam. The SST is a face-to-face interview between one examiner and a

test-taker. In most cases, the examiner is a native speaker of Japanese who is officially certified as a SST examiner. There might be some criticism to the effect that the interview data is biased because a non-native English speaker conducts it, but we consider that sometimes it must be good for making interviewees relaxed. The SST consists of five stages as shown in Figure 1 below.

Stage1	Warming up	3-4 min.
Stage2	Picture Description	2-3 min.
Stage3	Role playing	1-4 min.
Stage4	Story telling	2-3 min.
Stage5	Winding down	1-2 min.

Figure 1. Five stages in SST

The interview starts with an informal chat on general topics such as the interviewee's job, hobbies, family, and so on. From stage 2 to stage 4, the interviewee is asked to do three task-based activities, namely picture description, role-playing, and story telling. Each stage consists of two parts; a task part and a follow-up part. After the actual task part, the interviewer asks some questions which have to do with the task in the follow-up part. The interview ends with, again, an informal chat. All the interviews are audio-recorded, and judged by two or three raters based on the SST evaluation scheme (SST level 1 to 9).

We consider the SST to be a very useful spoken resource, because most existing learner corpora consist of written language only. Another benefit of choosing the SST as the main resource of the corpus is that each file of data has specific information on the examinee's oral proficiency level, as assessed by the professional examiner. There are some developmental learners' corpora available, but in most cases, they determine the learners' proficiency level based on external factors such as school years. A comparison between sub-corpora based on school years sometimes is not reliable because it ignores other parts of the background of students, how they have been learning English and from what kind of teachers. On the other hand, the SST data contains more reliable information on learners' proficiency levels, which will help to make comparative research based on proficiency subsections of the corpus more valid. (Tono, 2001)

2.3 Transcribing and tagging

There are some general rules for transcribing the speech data. Even though a word is mispronounced, it is transcribed with a correct spelling as long as the transcribers can understand which word the speaker produced. If acronyms are pronounced as sequences of letters, they must be transcribed as a series of upper case letters, which are separated by spaces. It is not allowed to use either Roman or Arabic numerals. All numbers must be transliterated as words. The transcribers are allowed to insert the phrase and sentence boundaries with commas and full stops on their own decision. Some information on non-verbal behaviors or concurrent events such as remarkable noises is also inserted.

There are two kinds of tags used in this corpus; the basic tags for such as filled pauses or repetitions, and the error tags for the analysis of the learners' errors. In this section, we would like to explain about the basic tags. The error tags will be explained in the section 2.4.

The tags in this corpus are based on the XML syntax although there is one exception that crossings of different tags are allowed. There are many advantages of using the tags based on XML syntax. XML can clearly identify the structure of the text and it is quite beneficial when the corpus data is exploited for the web-based pedagogical tools or database as a hypertext.

There are more than 30 basic tags and they are divided into four categories: tags for representing the structure of the entire transcription file, tags for the interviewee's profile, which is attached as a header of the file, tags for speaker turns, and tags for representing several phenomena in an utterance

such as fillers and repetitions.

We have developed the original software tool for SST transcription, the “TagEditor”. We believe that it must be good to use an editor designed especially for transcribing and tagging because manual transcribing and tagging are time-consuming, and it is inevitable that the transcribers will make a lot of mistakes if they work on just a simple text editor. Another reason for developing the TagEditor is that since we use a non-standard XML style tagset, it would not be a good idea to use the existing format checkers for standard XML texts. We believe that the TagEditor improves the efficiency of transcribing. The system can be easily updated whenever the tagset is revised.

Although the TagEditor is still under development, it is already accommodating a lot of functions. The transcribers can insert tags just by clicking on buttons. If the transcribers highlight the letters, then click on the button indicating a certain tag, the selected letters are wrapped with a tag. Even people who are not good at working on computers can easily do this. For those who are advanced computer users, it might be easier for you to input by keyboard than to click on a mouse. Tags can also be inserted with short-cut keys. Users can assign a key to each tag as they like.

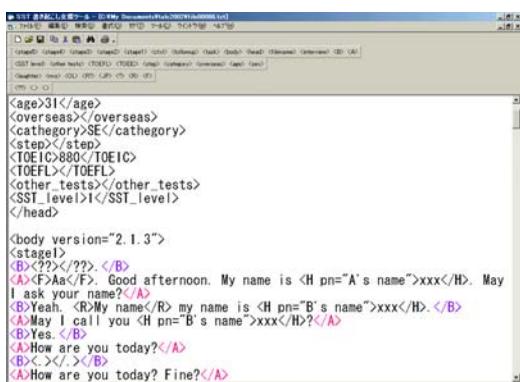


Figure 2. Screen dump of the TagEditor

2.4 Error tagging

The most remarkable difference between corpora of native and non-native (learners) speakers would be that a lot of errors are contained in learners' corpora. It has been said that analyzing errors produced by learners is quite efficient for finding out the learners' developmental stages and decide what is the best teaching method for them. In this project, we decided to analyze errors mainly by error tagging to construct a model of Japanese learners' English across different proficiency levels. We are aware that it is quite difficult to design a consistent error tagset because the learners' errors extend across various linguistic levels including grammar, lexis and phonetics. To do this, it is necessary to have a robust error typology.

We designed our original error tagset for learners' grammatical and lexical errors which are relatively easy to categorize, compared with other types of errors such as discourse errors or the ones related to more communicative aspects of learners' language. Our error tags contain three pieces of information; part of speech, a grammatical/lexical system and a corrected form. For some errors which cannot be categorized to any word class, such as the misordering of words, we prepare special tags. Our error tagset currently consists of 45 tags. The structure of the error tags is as follows. (Figure 3)

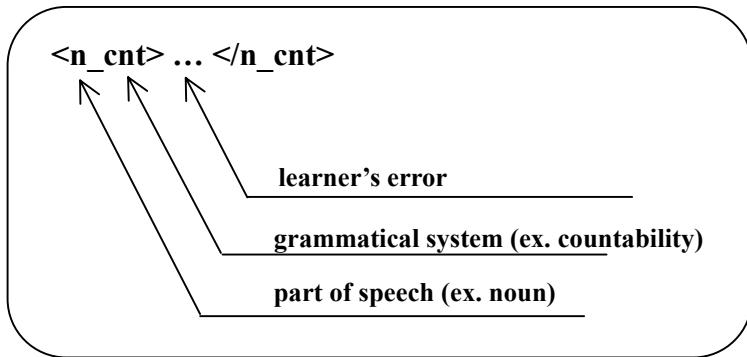


Figure 3. The structure of an error tag

I will show several examples of error-tagged data below.

- *She has been waiting here for many <n_num crr="hours">hour</n_num>.
(The wrong number of a noun)
- *The car looks <aj_inf crr="better">more good</aj_inf> now that it has been waxed.
(The wrong inflection pattern of an adjective)
- *There <v_agr crr="is">are</v_agr> a computer.
(The wrong subject-verb agreement)
- *If you <v_tns crr="go">will go</v_tns> to Vancouver, be sure to take a day trip to Victoria.
(the wrong tense of a verb)

2.5 Sub corpora

We have also compiled sub-corpora for comparison. For example, if we have the native speakers' speech data, it would be useful for comparing the utterance of native speakers and Japanese learners. We could do this by collecting the speech data of native speakers' taking the similar kind of interviews as SST. We have also compiled a back-translation corpus. It can be compiled mainly by guessing what the learners intended to say in the interview, and translate it into correct Japanese. We believe it's not so difficult to do this for Japanese native speakers or people who are familiar with English that Japanese learners speak. With the back-translation corpus, we could study how L1 (Japanese) transfer interferes the second language acquisition, or what kind of things are difficult for Japanese learners to express in English. As stated above, we perform error tagging only for grammatical and lexical errors. These sub corpora may cover what we cannot examine only by error tagging.

3. Linguistic analysis

In this chapter, we would like to show an example of how this corpus can be exploited to examine the learners' language.

3.1 Analysis on Japanese learners' acquisition of English article system

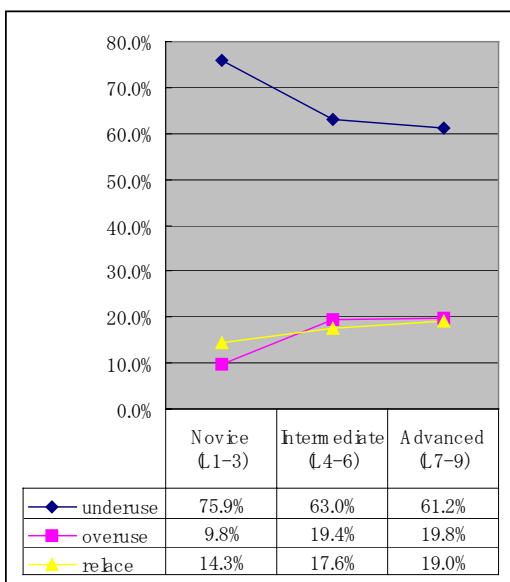
The English article system is one of the most difficult item for Japanese learners. We examined the article errors made by Japanese learners by using this corpus. For this analysis, we used the error-tagged part of the SST Corpus. We wanted to compare the tendencies of each SST level, but since we have not got enough data of certain levels (L1 and L9), we decided to divide them into three groups; novice (L1-3), intermediate (L4-6), and advanced (L7-9), and make a comparison between them.

Figure 4. The number of article errors in error-tagged data of SST Corpus (Izumi, 2003)

	The number of the interviews	The number of article errors	The number of nouns in learners' utterances	The number of article errors per 100 nouns
Novice (L1-3)	37	569	3158	18
Intermediate (L4-6)	35	817	4818	16
Advanced (L7-9)	30	547	5927	9

As seen in Figure 4, the number of the article errors per 100 nouns is getting lower and lower as the level gets higher. By examining all these article errors, we found that there were three types of errors: overuse, underuse, and replacement of articles. We compared these types across different proficiency levels. (Figure 5)

Figure 5. Three types of the article errors



We also found an interesting tendency in the replacement errors. In all proficiency levels, the replacement of “a/an” with “the” is the most frequent. This might be because if the learners cannot decide which is better, “a/an” or “the”, they tend to choose “the”, which can be put to a singular, plural and other abstract nouns. The learners often use this kind of communication and learning strategy of simplifying a rule to make communication carry on even if they do not understand the rule properly.

We found that the most frequent errors were the underuse errors in all proficiency levels. The reason for this might be that Japanese language does not have an article system. However, the frequency of the underuse errors decreases as the level goes up, and the overuse and the replacement errors increase. Most of the underuse errors are related to the syntactic rules, on the other hand, the overuse and the replacement errors are related to the semantic and pragmatic rules in most cases. It is often said that the semantic and pragmatic rules are more strongly influenced by the L1 transfer, which means they are more difficult to understand properly for learners than the syntactic rules. (Mizuno, 2000)

Figure 6. Four types of the replace errors

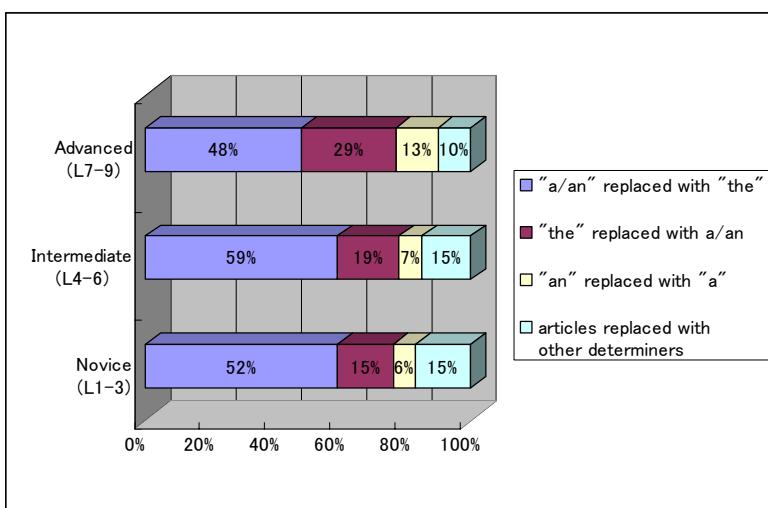
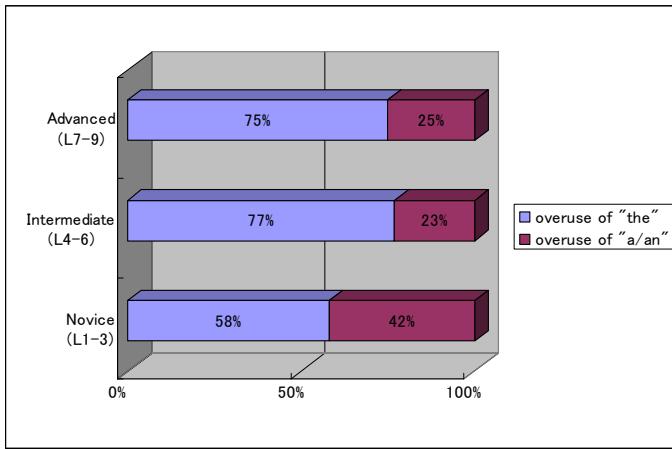
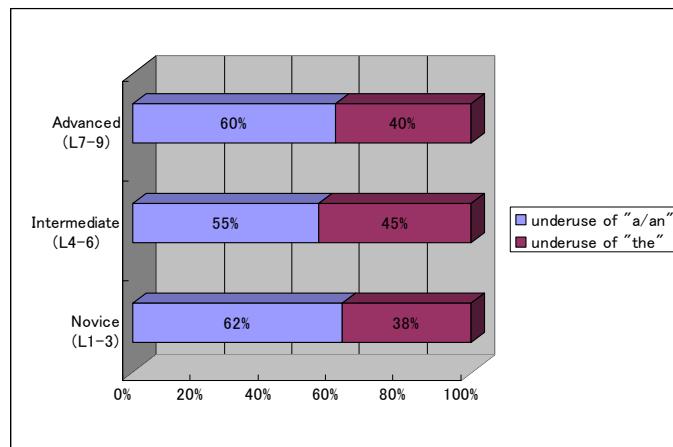


Figure 7. Two types of the overuse errors



In the overuse errors, we found that the overuse of “the” was more frequent in the intermediate and the advanced levels than in the novice level. This result might imply that understanding the meaning of zero articles is more difficult than understanding how to distinguish “a/an” from “the”. Although there is not an article system in Japanese, “no articles” in Japanese is not equal to zero article in English.

Figure 8. Two types of the underuse errors



In the underuse errors, the underuse of “a/an” was more frequent than the underuse of “the”. This might be occurred because L2 learners tend to produce an L2 output by translating the sentence in L1. An English article “the” is usually learned as the counterpart of a Japanese pronoun “sono”, but Japanese teachers often tell their students not to translate “a/an” into Japanese. Therefore, the learners tend to forget to put “a/an” when they translate a Japanese sentence into English. It can be said that this might be a type of errors which has been induced by the way of teaching.

There has been much research on the learners’ acquisition order of various linguistics phenomena, such as negation, tense and aspect, and so on. Most of them are examined based on the data that was collected on purpose for inducing the learners’ errors on an intended grammatical rule. This must be a good way to get results that the researchers have intended, from a small amount of data. However, it can be said that the results extracted from the more spontaneous and large-scale data are more reliable, although it would sometimes be rather difficult to formulate the extracted results.

4. Applications in collaboration with NLP

In this chapter, we would like to introduce an experiment we have done to examine how this corpus can be applied to develop the system in collaboration with NLP.

4.1 Automatic level check system

We examined the efficiency of judging learners’ proficiency levels automatically, and what kind of information in learners’ data is the most effective for it. We thought the following information is useful for the automatic level check based on SST evaluation scheme. (Figure 9)

<u>Vocabulary</u>
1) The number of each vocabulary
2) The number of 1) which appears more than once
3) The number of 1) which is a content word
4) The number of 3) which appears more than once
5) The number of the sequences of two words (Bi-gram)
6) The number of 5) which appears more than once
7) The number of the words produced only by learners (excludes the words also produced by interviewers)
8) The frequency of each level when assigning the words to the twelve levels of ALC Press's Standard Vocabulary List (SVL)
<u>Grammar</u>
9) The frequency of each part of speech
<u>Fluency</u>
10) The number of fillers, repetitions, self-corrections and pauses
11) Duration per word
12) The total number of words
13) Average sentence length
14) The number of sentences
15) Duration of an entire interview
<u>Communicative skills</u>
16) The volume of interviewers' vocabulary and the frequency of each word
17) The number of speaker turns

Figure 9. The features which can be used for the automatic level check (Supnithi, 2003)

The automatic level check can be considered as the similar task to the automatic text categorization. For automatic text categorization, the machine learning method called SVM (Support Vector Machines) is often used. We used this method for the level check. SVM is defined by Cristianini and Shawe-Talor (2000) as follows: “Support Vector are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory. This learning strategy ... is a principled and very powerful method that in the few years since its introduction has already outperformed most other systems in a wide variety of applications.”

In the next step, we tried to decide which features in Figure 9 are the most effective as the features for the automatic level check based on SVM. We did this by the depth-first search. We first did the level check by using all the information (the full set of all the features) in Figure 9. We repeated the same thing by using the set of 8)-17) and one feature picked up from 1)-7). Then, we found the set of 8)-17) and 7) has the highest accuracy rate. Then we decided to use those features as a root of searching for the level check. From the searching result, we found that the features 7), 8), 9), 10), 11), 13), and 15) are the most effective information.

Figure 10 shows the result of our experiment for the automatic level check.

Figure 10. The result of automatic level checking (Supnithi, 2003)

SST Level	The number of interviews	Accuracy rate (0)	Accuracy rate (± 1)
1	3	0%	100%
2	29	65.52%	100%
3	198	68.69%	99.50%
4	452	83.23%	99.78%
5	226	44.69%	98.67%
6	127	47.24%	92.13%
7	52	26.92%	82.69%
8	24	12.50%	70.83%
9	9	0%	11.11%
All	1,121	65.57%	96.52%

As seen in Figure 10, overall accuracy rate using our most effective information (65.57% for exact match, and 96.52% for allowing one-level gap) is not bad. However, the accuracy rates are still relatively low in the levels of which we have not got the enough data. Therefore, we are planning to improve our framework by collecting more data of those levels, and using other kinds of information such as learners' errors.

5. Conclusion

In this paper, we have presented an overview of the SST Corpus by explaining the data collection procedure such as transcribing and tagging including error tagging for error analysis. We also showed to what extent this corpus can be exploited first by showing our analysis on learners' acquisition of the English article system, which might be useful for constructing a model of the learners' language, and by introducing the framework for the automatic level check.

We are planning to make this corpus publicly available so that teachers and researchers in many fields can use it for their own research interests, such as second language acquisition research, syllabus and material design, or the development of computerized pedagogical tools by combining it with NLP technology.

Acknowledgements

We have benefited greatly from cooperation and discussions with Assistant Prof. Yukio Tono of Meikai University.

References

- Cristianini, N & Shawe-Talor, J 2000 An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, pp.7
- Izumi, E 2003 Error tag tsuki nihonnjinn eigogakushusha hatsuwa corpus wo mochiita gakushusha no kanshi shuutoku keiko no bunseki. (In Japanese) In *Proceedings of the ninth annual meeting of the Association for Natural Language Processing*, Japan
- Mizuno, M 2000 Chukan gengo bunseki, (In Japanese) Kaitaku-sha pp.74-75
- Supnithi, T 2003 Eigogakushusha hatsuwa corpus wo mochiita shuuujukudohanntei. (In Japanese) In *Proceedings of the ninth annual meeting of the Association for Natural Language Processing*, Japan
- Tono, Y 2001 The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In *Proceedings of ASIALEX2001*, Korea, pp.257-262