사전 편찬학 연구 제11집 2호(2001)
pp. 7~17

The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography

TONO, Yukio & KANEKO, Tomoko
ISAHARA, Hitoshi, SAIGA, Toyomi, IZUMI, Emi
NARITA, Masumi & KANEKO, Emiko

연세대학교 언어정보개발연구원

The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography

TONO, Yukio[*] & KANEKO, Tomoko[**]
ISAHARA, Hitoshi, SAIGA, Toyomi, IZUMI, Emi[***]
NARITA, Masumi[****] & KANEKO, Emiko[*****]

Abstract

This paper will report on the progress of a project to compile a one-million word spoken corpus of Japanese learners of English. In 1999, we launched a project to compile a new learner corpus in collaboration with ALC Press Inc. and Communications Research Laboratory. The major characteristic of the project is that the corpus data is entirely based upon audio-recordings of an English oral proficiency interview test called ACTFL-ALC Standard Speaking Test. One of the unique features of the corpus is that each speaker's data include his or her proficiency profile based on the SST evaluation schemes. This makes it possible for corpus users to study Japanese learner English across different proficiency levels. In this study, we will describe the project by summarizing the data

[*] Department of Foreign Languages and Cultures, Meikai University 8 Akemi, Urayasu, Chiba 279-8550, JAPAN / y.tono@meikai.ac.jp

[**] Showa Women's Univeristy. 1-7 Taishido Setagaya-ku, Tokyo 154-8533 / JAPAN kaneko@swu.ac.jp

[***] Communications Research Laboratory 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, JAPAN / {isahara, saiga, emi}@crl.go.jp

[****] Software Research Center, RICHO 1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN / m-narita@nts.ricoh.co.jp

[*****]ALC Press Inc. 2-54-12, Eifuku, Suginami-ku, Tokyo 168-8611, JAPAN ekaneko@alc.co.jp

collection procedure, text format, transcription guidelines, and annotation schemes as well as the theoretical and pedagogical implications of the project. We will focus especially on how learner corpora data can be exploited from the viewpoint of pedagogical lexicography.

1. Introduction

The purpose of this paper is to report on an on-going project to compile a 1-million-word spoken corpus of Japanese learners of English and to provide possible implications for pedagogical lexicography as one of the potential areas of application. Recently, second language acquisition researchers and language teaching professionals have begun to realize the importance of learner corpora as resources for teaching and research, and major dictionary publishers such as Longman and Cambridge University Press have already compiled their own learner corpora in order to enrich their dictionary content by providing information on common learner errors. Projects such as ICLE (International Corpus of Learner English; see Granger 1998) and JEFLL (Japanese EFL Learner) Corpus (see Tono 2000) both aim to compile learner corpora to describe the interlanguage of particular L2 learner groups using a corpus-linguistic methodology. To date, however, most of these learner corpus projects are composed of written data only, while the few spoken learner corpora that do exist (e.g. LINDSEI project at Louvain, described in De Cock, et al. 1999), are rather small in size.

We launched our Standard Speaking Test (SST) Corpus project in 1999. This project is a joint collaboration between Communication Research Laboratory and ALC Press, with a few advisory members from universities. While English is taught over six years in secondary schools in Japan, many of us feel that Japanese learners still cannot function properly in English for communication purposes. It has been argued that one of the problems of English language teaching in Japan is that no serious attempt has been made to systematically record and describe the acquisition process of Japanese learners of English in our EFL context. It is very important to know objectively how much English we have acquired after six years of instruction, and the progression of standards. Mere adoption and application of teaching methods from foreign countries will not always work in our country. We need to gather data on how Japanese learners learn English and to describe the developmental path of their interlanguage. Thus, one of the main purposes of this project is to identify the features of interlanguage at different stages of L2 acquisition and construct a model of interlanguage development. We hope to identify the mechanisms of development from one stage of interlanguage to another, which we hope will lead to the improvement of teaching methods and more rigorous empirical research on the effect of learning methods on the transition process.

2. The Standard Speaking Test

The Standard Speaking Test (SST) is collaboration between the American Council on the Teaching of Foreign Languages (ACTFL) and ALC Press. It is based on the *ACTFL Proficiency Guidelines* for speaking and the Oral Proficiency Interview (OPI). The ACTFL-OPI was first developed in 1982 and since then it has been one of the most influential speaking tests in the world[1)] despite the fact that there has been some criticism against the empirical bases of the guideline (see, for example, Bachman and Savignon 1986; Chalhoub-Deville 1997). ACTFL and ALC Press worked together to develop a new speaking test for Japanese learners of English. The proficiency guideline defines 9 different proficiency levels (Level 1 is the most basic). Each level is defined specifically in terms of the following criteria: (a) text type, (b) accuracy, (c) pronunciation, (d) fluency, and (e) overall task & function.

The SST takes the form of a 10- to 15-minute tape-recorded conversation between a trained interviewer and a test candidate. The SST utilizes interview techniques and picture prompts to simulate natural conversation to the maximum extent possible in a testing situation.

The tape-recorded interview is scored by a trained Rater (a certain percentage of tapes are second-rated by Master Raters). In the SST, the elicitation and scoring of speech samples are separate procedures, unlike the OPI where both tasks are performed on-the-spot by the interviewer. The SST interview process elicits speech samples through the application of the following five-stage format:

1) Warm-up and initial assessment
2) Single picture prompt with level checks and probes
3) Role-play
4) Single or picture sequence prompt with level checks and probes
5) Wind-down

Although the interviewer is not responsible for the formal rating of the

[1] At the time of writing, the ACTFL-OPI is available for 32 differentlanguagearoun the world

test candidate, the interviewer must be able to conduct an on-going informal assessment of the speaker's proficiency in order to tailor the questions, prompts, and role plays most suited to the test candidate's interests and level of speaking proficiency. If the speech sample is poorly elicited via prompts inappropriate to the speaker's level of proficiency, the rating of the speech sample may become invalid.

The SST serves to discriminate spoken English at Novice to Intermediate High levels of proficiency utilizing a shorter interview than the OPI. The SST discriminates more finely at Intermediate proficiency levels than do other existing standardized measurement instruments of speaking proficiency. The SST can also serve as a potential screening interview for speakers who might be ready for, and benefit from taking, the OPI.

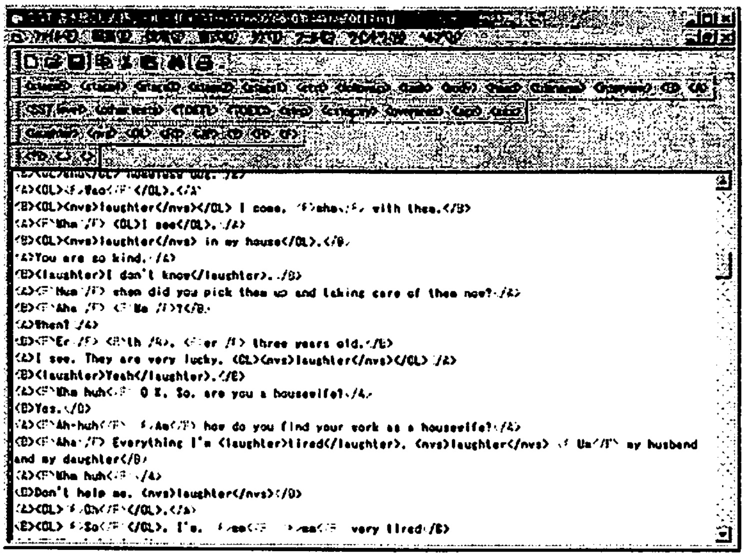
ALC Press possesses large archives of audio recordings of this test. We saw this data as potentially very useful spoken resources, as most learner corpora to date consist of written data only. For this reason, we decided to launch the present project to transcribe these archived recordings and convert them into a spoken learner corpus. The strength of this corpus project is that each file/piece of data has specific information on the examinee's oral proficiency level, as assessed by the professional examiner. Whilst there are some developmental interlanguage corpora available (e.g. JEFLL), the labelling or determination of learner proficiency levels is often based on external factors such as school years. Thus, a comparison between subcorpora based on school years sometimes causes a problem. In the Longman Learner's Corpus, learner proficiency for each file is encoded in its header, but judgements about the proficiency levels seem to be entirely up to the teachers who donated the composition data, and are thus not entirely reliable, since we have no information about the possibly varied criteria or standards used by the different teachers. Compared with these other learner corpora, therefore, SST data have more reliable information on learners' proficiency levels, which will help to make comparative research based on proficiency subsections of the corpus more valid.

3. The SST Corpus Project

The SST Corpus has been compiled as part of a larger project called 'Research & Development of Congruent Communication Technologies' led by the Telecommunications Advancement Organization of Japan (TAO) and Communication Research Laboratory (CRL). The primary goal of this project is to develop natural language processing technologies that can handle human errors in speech or writing and their application in such areas as support systems for writing in English, machine translation, and applications in education (e.g. learner error databases, computer-assisted self-access language learning systems).

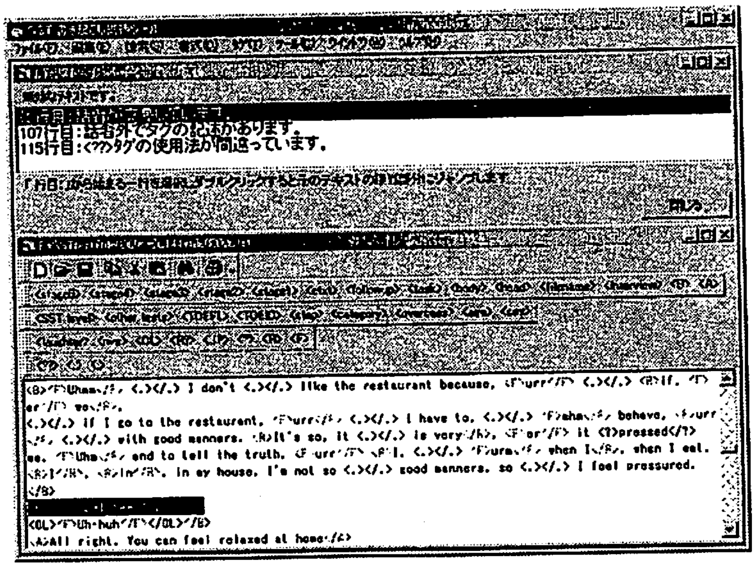
The SST learner corpus data have various types of error (lexical, syntactic, phonetic, etc.). Information on the types and rates of error that learners make at various proficiency levels, the contexts in which particular learner errors occur, etc. will serve as the input for machine learning of interlanguage grammar. CRL has created an automatic machine learning system (consisting of a lexicon and a grammar) which can be potentially very useful for testing and evaluating the interlanguage grammar model that the machine learns automatically. The results can be used for NLP and educational purposes.

The initial phase of the project includes the development of the following tools and guidelines: (1) transcription guidelines, (2) tagging schemes, (3) a tag editor, (4) error tagging schemes, and (5) error tagging support tools. At the time of writing (April 2001), the first version of a set of transcription guidelines has been prepared for transcribing the data of about 140 subjects, using a "TagEditor" (see Figure 1). *TagEditor* has been developed specially for this learner corpus project. Besides the usual functionalities common to text editors, it has unique features such as the ability to insert customized XML tags, validate tags (see Figure 2), show tags in colour, do simple grep searches & concordancing, use file templates, check file formats, do automatic 'search and replace', etc. A similar tool was developed at Louvain for the ICLE project, but our tool has several new features, including tag validation and a simple concordancer.



```
<A><OL><F>Wao</F></OL>.</A>
<B><OL><nvs>laughter</nvs></OL> I come, <F>ehe</F> with them.</B>
<A><F>Uha</F> <OL>I see</OL>.</A>
<B><OL><nvs>laughter</nvs> in my house</OL>.</B>
<A>You are so kind.</A>
<B><laughter>I don't know</laughter>.</B>
<A><F>Mum</F> when did you pick them up and taking care of them now?</A>
<B><F>Aha</F> <F>Me</F>?</B>
<A>Then?</A>
<B><F>Er</F> <F>Uh</F>, <F>er</F> three years old.</B>
<A>I see. They are very lucky. <OL><nvs>laughter</nvs></OL></A>
<B><laughter>Yeah</laughter>.</B>
<A><F>Uha huh</F> O K. So, are you a housewife?</A>
<B>Yes.</B>
<A><F>Ah-huh</F> <F>Am</F> how do you find your work as a housewife?</A>
<B><F>Aha</F> Everything I'm <laughter>tired</laughter>. <nvs>laughter</nvs> <F>Um</F> my husband
and my daughter</B>
<A><F>Uha huh</F></A>
<B>Don't help me. <nvs>laughter</nvs></B>
<A><OL><F>Oh</F></OL>.</A>
<B><OL><F>So</F></OL>. I'm <F>em</F> <F>em</F> very tired</B>
```

<Figure 1> A screenshot of TagEditor



107行目:話者外でタグの記述があります。
115行目:<??>タグの使用法が間違っています。

「行目:」から始まる一行を選択しダブルクリックすると元のテキストの該当部分にジャンプします。

```
<B><F>Uhmm</F> <.></.> I don't <.></.> like the restaurant because, <F>urr</F> <.></.> <R>If, <F>
er</F> we</R>,
<.></.> if I go to the restaurant, <F>urr</F> <.></.> I have to, <.></.> <F>ahm</F> behave, <F>urr
</F> <.></.> with good manners. <R>It's so, it <.></.> is very</R>, <F>er</F> it <?>pressed</?>
me. <F>Uhm</F> and to tell the truth, <F>urr</F> <R>I, <.></.> <F>urm</F> when I</R>, when I eat.
<R>I</R>, <R>In</R>. In my house, I'm not so <.></.> good manners, so <.></.> I feel pressured.
</B>
<OL><F>Uh-huh</F></OL></B>
<A>All right. You can feel relaxed at home</A>
```

<Figure 2> Tag validation function of TagEditor

Currently, the team is working on error tagging schemes. It is extremely difficult to define a generic error tagset that can be applied to any type of learner corpus. We will implement an error tagset which is as general as possible, but at the same time focused mainly on lexical and syntactic errors only. At a later stage, however, we will shift our focus to other types of errors, e.g. phonetic, pragmatic, and discoursal errors. The data will be made publicly available after the three-year project ends in 2003.

4. Learner Corpus Data and Pedagogical Lexicography

There is a growing awareness that L2 learner corpora can provide useful information about learner errors. A systematic analysis of the interlanguage process will shed light on the learning process and help identify the specific learning difficulties of L2 learners. This type of information will be very useful for designing foreign language syllabuses and teaching materials. The compilation of learners' dictionaries should also benefit from such findings. Longman, for example, produced the first learners' production dictionary called *Longman Language Activator,* which was designed to support the L2 learner's encoding activities by supplying synonymous words with usage notes and illustrative examples. The essential edition of this *LLA, Longman Essential Activator*, was published a few years later, and contained special columns for common errors based upon the error information extracted from the Longman Learners' Corpus. Tono (1996) demonstrated the value of learner error information provided in learner's dictionaries by examining errors in the use of basic verb patterns by Japanese-speaking learners of English.

Learner corpus data can help ascertain and delineate problematic areas for learners such as the following:

(a) Pronunciation & intonation
(b) Lexical collocation (e.g. '*strong* (NOT *powerful*) *tea*')
(c) Grammatical collocation (e.g. verb patterns; noun patterns; adjective patterns, etc.)
(d) Number agreement (e.g. Uncountable versus Countable; grammatical morphemes such as plural -s, etc.)
(e) Article system
(f) Tense & aspect (especially morphological errors)
(g) Discourse & pragmatic errors (e.g. misuse of connectors, fillers, etc.)

It should be noted that not all the above types of information can be supplied by existing learner corpora. Since most learner corpora available to date are collections of written essays, information such as (a) and (g) are difficult (if not impossible) to gather with such corpora.

Spoken learner corpora such as the SST Corpus have the potential to provide very interesting data for pedagogical dictionary-making. For example, the detailed analysis of learners' pronunciation errors can supply important information for the writing of usage notes addressing the typical difficulties faced by learners. Information about items (b) - (f) can be obtained from both written and spoken data, but the error frequencies in speech data can be very different from those in writing. It would be desirable for learners to be informed about the different problematic areas in writing and speech. Unfortunately, however, almost no information on learner errors in terms of 'speech versus writing' has yet been supplied in learner's dictionaries. In pedagogical dictionary- making, it was not until five years ago that information on the differences in usage between written and spoken language by native speakers was first systematically described (cf. LDOCE 1995). The next- generation of learner's dictionaries will integrate earner corpus findings on English vocabulary and grammar into their entries.

References

Bachman, L.F. and S. Savignon (1986) The evaluation of communicative language proficiency: A critique of the ACTFL oral interview, *Modern Language Journal 70*, 380-390.

Chalhoub-Deville, M. (1997) Theoretical models, assessment frameworks and test construction, *Language Testing 14*, 3-22.

De Cock, S., S. Granger and S. Petch-Tyson (1999) The Louvain International Database of Spoken English Interlanguage (LINDSEI) Project. An internal report at Catholic University of Louvain.

Tono, Yukio (1996) Using Learner Corpora for L2 Lexicography. LEXIKOS 6 (AFRILEX SERIES 6) Stellenbosch: Universiteit van Stellenbosch, pp. 116-132.

Tono, Yukio (2000) A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. Lewandowska-Tomaszczyk, B. and J. P. Melia, (eds.) *PALC'99: Practical Applications in Language Corpora*. Frankfurt am Main: Peter Lang, pp.323-340.