Developing a One Million Word Spoken EFL Learner Corpus

Yukio Tono
Meikai University
Tomoko Kaneko
Showa Women's University
Hitoshi Isahara
Communication Research Laboratory
Emi Izumi
Communication Research Laboratory
Toyomi Saiga
Communication Research Laboratory
Emiko Kaneko
ALC Press

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 the 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 the learner language across different proficiency levels, a feature which has not often been available for other learner corpora. In this paper, we will describe the project by summarizing the data collection procedure, text format, transcription guidelines, and annotation schemes (especially error tagging schemes), as well as the theoretical and pedagogical implications of the project.

本研究は１００万語規模の日本人英語学習者の話し言葉コーパス構築プロジェクトの中間報告である。１９９９年に我々は出版社アルクおよび通信総合研究所の協力を得て新しいコーパスを作成する試みを開始した。このプロジェクトの最大の特徴は、コーパスがACTFL-ALC Standard Speaking Testと呼ばれる英語会話能力面接試験の録音テープに基づいている点である。統一されたテスト判定結果がデータ内に話者属性として与えられているため、学習者グループを能力別に分けてサブコーパス検索などが可能になる、という大きな特徴を持っている。本論では、データ収集方法、フォーマット、書き起こしガイドライン、注釈付け（特にエラータグの仕様）などについて概説し、このプロジェクトの理論的、教育的意義について述べる。

The paper reports on an on-going project to compile a one-million-word spoken corpus of Japanese learners of English and provides possible implications for SLA and ELT research as well as language engineering. Recently, SLA 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.

The Standard Speaking Test

The *Standard Speaking Test* (SST) is a 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 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: 'Novice-Low' through Level 9: 'Advanced'). Each level is defined specifically in terms of the following criteria: (a) context/content area, (b) text type, (c) global task & function, (d) accuracy, which includes grammatical accuracy, fluency, and pronunciation.

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 two different Raters (in case of disagreement, the interview is rated by a Master Rater). 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 with level checks and probes
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 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.

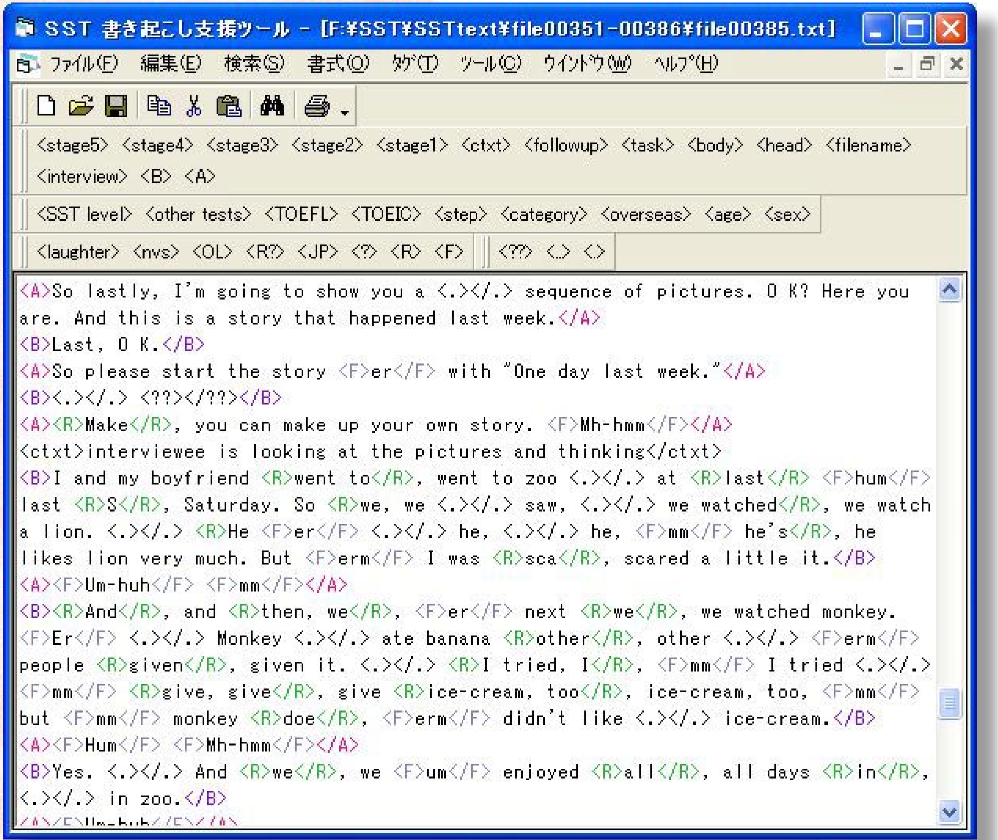
ALC Press possesses large archives of audio recordings of this test. We saw this data as potentially very useful spoken resources, as most learner corpora to date consist of written data only. For this reason, we decided to launch the present project to transcribe these archived recordings and convert 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.

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.

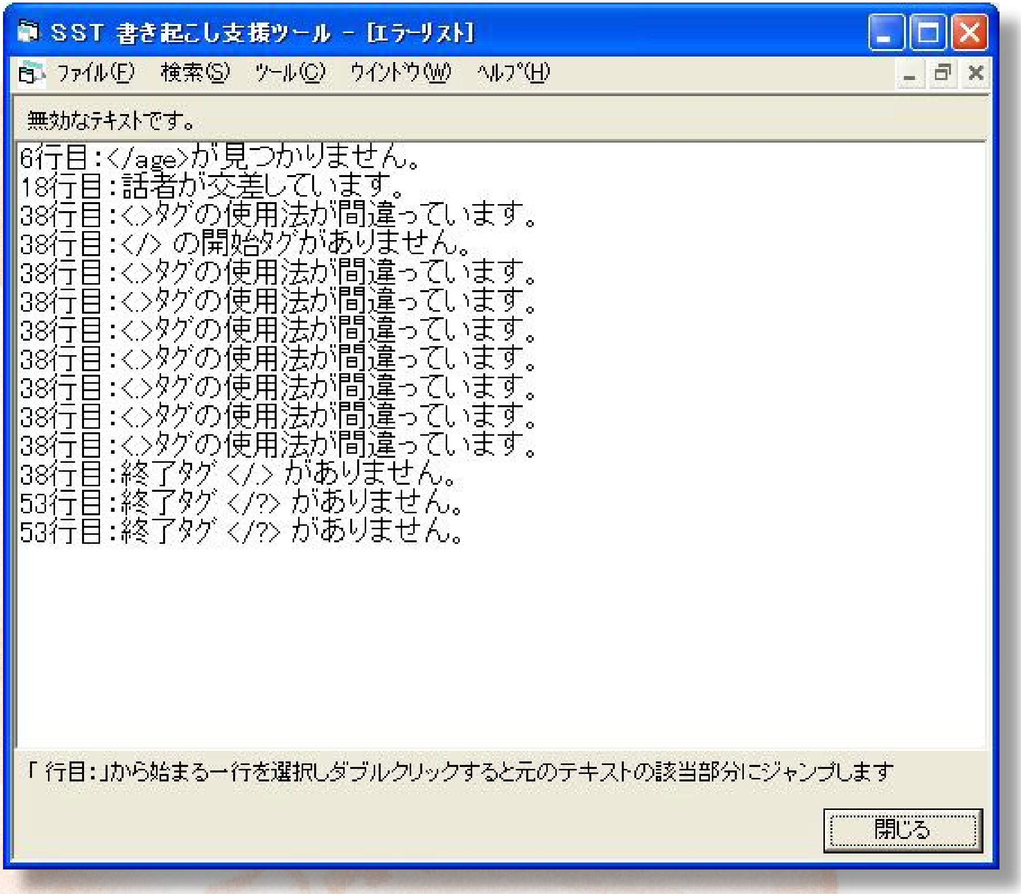
The initial phase of the project includes the development of the following tools and guidelines: (1) transcription guidelines, (2) tagging schemes, (3) a tag editor, (4) error tagging schemes, and (5) error tagging support tools. At the time of writing (December 2001), the second version of a set of transcription guidelines has been prepared for transcribing the data of about 510 subjects, using a "TagEditor" (see Figure 1).


SST 書き起こし支援ツール - [F:¥SST¥SSTtext¥file00351-00386¥file00385.txt]
ファイル(F) 編集(E) 検索(S) 書式(O) タグ(T) ツール(C) ウインドウ(W) ヘルプ(H)
<stage5> <stage4> <stage3> <stage2> <stage1> <ctxt> <followup> <task> <body> <head> <filename>
<interview> <B> <A>
<SST level> <other tests> <TOEFL> <TOEIC> <step> <category> <overseas> <age> <sex>
<laughter> <nvs> <OL> <R?> <JP> <?> <R> <F>
<??> <.> <>

```
<A>So lastly, I'm going to show you a <.></.> sequence of pictures. O K? Here you
are. And this is a story that happened last week.</A>
<B>Last, O K.</B>
<A>So please start the story <F>er</F> with "One day last week."</A>
<B><.></.> <??></??></B>
<A><R>Make</R>, you can make up your own story. <F>Mh-hmm</F></A>
<ctxt>interviewee is looking at the pictures and thinking</ctxt>
<B>I and my boyfriend <R>went to</R>, went to zoo <.></.> at <R>last</R> <F>hum</F>
last <R>S</R>, Saturday. So <R>we, we <.></.> saw, <.></.> we watched</R>, we watch
a lion. <.></.> <R>He <F>er</F> <.></.> he, <.></.> he, <F>mm</F> he's</R>, he
likes lion very much. But <F>erm</F> I was <R>sca</R>, scared a little it.</B>
<A><F>Um-huh</F> <F>mm</F></A>
<B><R>And</R>, and <R>then, we</R>, <F>er</F> next <R>we</R>, we watched monkey.
<F>Er</F> <.></.> Monkey <.></.> ate banana <R>other</R>, other <.></.> <F>erm</F>
people <R>given</R>, given it. <.></.> <R>I tried, I</R>, <F>mm</F> I tried <.></.>
<F>mm</F> <R>give, give</R>, give <R>ice-cream, too</R>, ice-cream, too, <F>mm</F>
but <F>mm</F> monkey <R>doe</R>, <F>erm</F> didn't like <.></.> ice-cream.</B>
<A><F>Hum</F> <F>Mh-hmm</F></A>
<B>Yes. <.></.> And <R>we</R>, we <F>um</F> enjoyed <R>all</R>, all days <R>in</R>,
<.></.> in zoo.</B>
<A><F>Um-huh</F></A>
```

Figure 1: A screenshot of *TagEditor* Version 1.2.

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-like tags, validate tags (see Figure 2), show tags in colour, do simple grep searches and concordancing, use file templates, check file formats, do automatic 'search and replace', and so on. 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.



無効なテキストです。

```
6行目:</age>が見つかりません。
18行目:話者が交差しています。
38行目:<.>タグの使用法が間違っています。
38行目:</> の開始タグがありません。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:<.>タグの使用法が間違っています。
38行目:終了タグ </.> がありません。
53行目:終了タグ </?> がありません。
53行目:終了タグ </?> がありません。
```

「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

Figure 2: Tag validation function of *TagEditor*

Currently, the team is working on error tagging. It is extremely difficult to define a generic error tagset that can be applied to any type of learner corpus. So far we have developed a generic error tagset and an accompanying error tag manual. We focus mainly on lexical and syntactic errors only at this stage. At a later stage, however, we will shift our focus to other types of errors, e.g. prosodic, pragmatic, and discoursal errors. The data will be made publicly available after the three-year project ends in 2003.

SST Data, SLA/ELT Research and Beyond

Using SST Corpus for SLA research

The influence of corpus linguistics has spread rather slowly in the fields of SLA research. Leech (1998) states two main reasons for this: firstly, it is extremely time-consuming and labour-intensive to create a corpus. Secondly, the intellectual climate in applied linguistics has been such that theory-driven experimental studies have prevailed for the last 20 years and the data-driven descriptive methods of corpus linguistics have not been preferred. Recently, however, there is a growing awareness that a computational analysis of a large body of language use data could provide useful data for language researchers as more focus is now on usage-based linguistic models as well as lexico-grammatical aspects of a language.

L2 learner corpora have great potentials for clarifying the L2 acquisition process. It serves us as better data than what we have mostly relied on so far and enables us to investigate the non-native speaking learners' language not only from a negative point of view ('What did the learner get wrong?') but also from a positive one ('What did the learner get right?') It also makes it possible to investigate L2 learners' avoidance behaviour by examining overuse/underuse phenomena. Undeniably, all types of SLA data have their strengths and weaknesses and one cannot help but agree with Ellis (1994:676) that 'Good research is research that makes use of multiple sources of data.' The learner corpora, if properly compiled, serve well as a valuable addition to current SLA data sources. They can be used for (a) verification of theories and hypotheses in SLA and (b) description of interlanguage development (developmental stages of linguistic features (lexico-grammatical/ discoursal/ pragmatic), L1 transfer effect, overuse/ underuse of particular features, universal/ L1 specific errors, native-like/ non-native-like performance (Tono 1998).

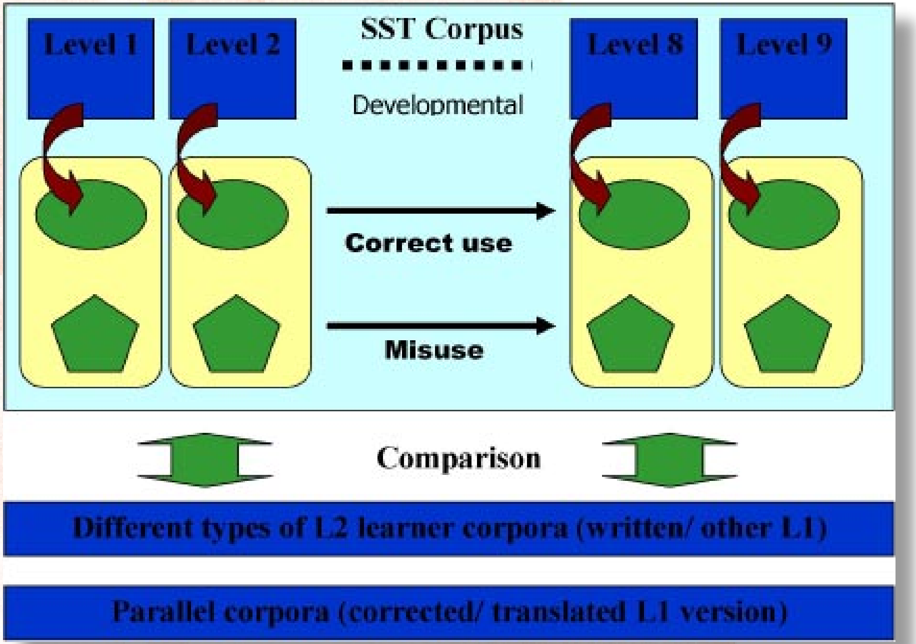

Level 1
Level 2
SST Corpus
Developmental
Level 8
Level 9
Correct use
Misuse
Comparison
Different types of L2 learner corpora (written/ other L1)
Parallel corpora (corrected/ translated L1 version)

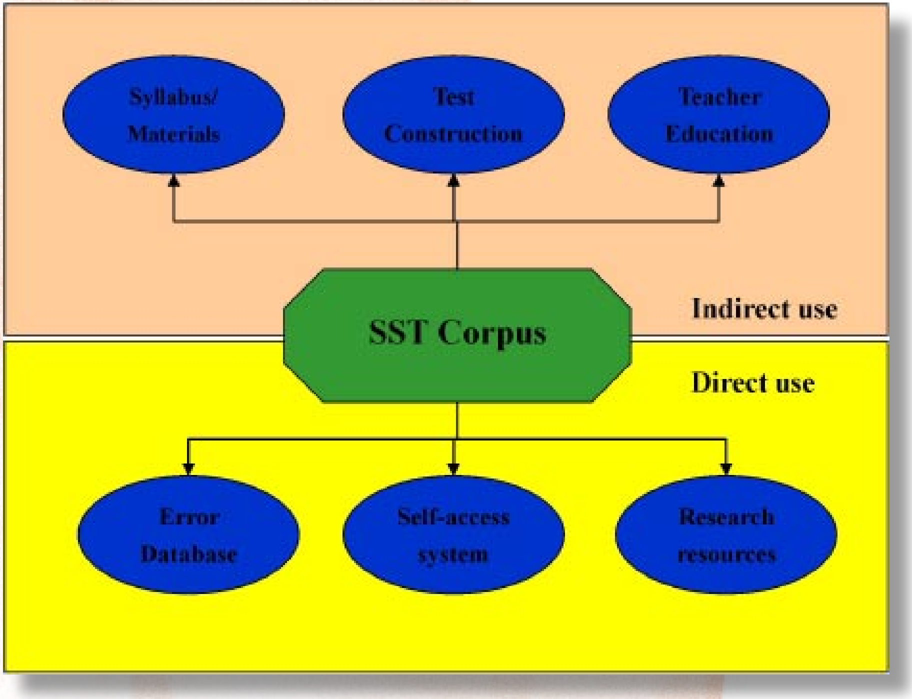
Figure 3: The potential of SST Corpus for SLA research

Figure 3 illustrates this point. If the patterns of use/misuse and overuse/underuse are identified across different stages of acquisition and a proper comparison is made between the SST data and different types of learner corpora such as written or parallel versions (with corrected errors), we will be able to identify various aspects of interlanguage grammar development. Ellis (2001) predicts an important role of corpora in SLA in the future and distinguishes three types of corpora as useful: (1) corpora of 'authentic' native speaker language use, (2) corpora of learner language use, and (3) corpora of native speaker language use with learners (i.e. samples of foreigner talk). The SST data serves as a useful tool for the second type.

The implications of the SST Corpus project for English language teaching

The exploitation of SST Corpus for English language teaching can be either direct or indirect. As shown in Figure 4, as indirect applications of the data, spoken learner corpus data could provide the information about L2 learners' acquisition process and possible error sources and patterns, which can be invaluable resources for (a) ELT syllabus and materials design, (b) test construction and (c) teacher education. For direct use, SST Corpus can be exploited as a part of on-line resources for L2 learners' data-driven learning or a part of research resources for postgraduate students in applied linguistics.

Figure 4: The potential of SST Corpus for English language teaching



Future technologies

Another interesting possibility is the collaboration with researchers in natural language processing (NLP) fields. They have considerable expertise in using language corpora for machine translation, information retrieval, and language modelling. Some computers can learn rules of a language statistically on the basis of a large amount of natural language data. If a learner corpus is fed into such a computer, it could possibly learn a grammar of interlanguage. Since SST Corpus has nine proficiency levels, it would be an interesting possibility to explore how a computer can statistically learn the grammars of interlanguages at different proficiency levels. This approach is in line with the one that Dan Jurafsky and his colleagues have taken recently in what they call 'computational psycholinguistics.' We could possibly develop an error detector or a language proficiency assessment system that can analyse the input from learners and identify which developmental stage a particular learner is at.

Learner data can also be used for the development of a noise-proof machine translation system. A future translation tool could hear the non-native speakers' erroneous sentences and interpret them properly. The learner corpus data can be useful resources for developing the knowledge base for such a system. Our team is ideal in the sense that we have both SLA and NLP researchers working together to fully exploit a corpus of Japanese-speaking learners of English.

We are hoping to finish the data collection within a year and release the corpus in the fiscal year 2003 so that it will be freely available for research and commercial purposes. We are looking forward to your input and comments on this exciting project.

References

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

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press: Oxford.

Ellis, R. (2001). Real Data and Real Pedagogy. Lecture given at the 2nd Learner Corpus Workshop at Showa Women's University, Tokyo.

Leech, G.. (1998). Preface. In S. Granger (ed.) *Learner English on Computer*. London: Longman.

Tono, Y. (1998). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *TALC (Teaching and Language Corpora) 98 Conference Proceedings*, Keble College Oxford.

Tono, Y. (2000) A corpus-based analysis of interlanguage development: analysing part-of-speech tag sequences of EFL learner corpora. Lewandowska-Tomaszcyk & J. Melia (eds.) *PALC'99: Practical Applications in Language Corpora*. Peter Lang GmbH, Frankfurt am Main, pp. 323-342.