Mari Carmen Campoy & María José Luzón (eds)

Spoken Corpora in Applied Linguistics

Offprint

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

ISBN 978-3-03911-275-3

Hochfeldstrasse 32, Postfach 746, CH-3000 Bern 9, Switzerland
info@peterlang.com, www.peterlang.com, www.peterlang.net

YUKIO TONO

The Roles of Oral L2 Learner Corpora in Language Teaching: the Case of the NICT JLE Corpus

1. Introduction

Corpora can provide a wide variety of linguistic information which could be useful in many different fields of language studies. Until two decades ago, most corpora were based on written texts, and with a few exceptions (e.g. the *London-Lund Corpus*, the *Spoken English Corpus*) very few spoken corpora were made available. Although more and more spoken corpora were constructed in the past twenty years, it is not until recently that digitized audio files and their transcripts in an orthographical format are integrated into a corpus in a sophisticated manner.

Today, there is a growing awareness that oral or spoken corpora could serve as useful resources for studying characteristics of human speech, thus being also beneficial for teaching oral skills of a language. Especially corpora of learner language (also known as *learner corpora*) have attracted much attention from researchers working in the field of second/foreign language learning and now there is a growing demand for spoken learner corpora to study oral proficiency skills of second language learners.

This study will first situate spoken learner corpora in learner corpus research and then introduce the *NICT JLE Corpus*, the biggest oral learner corpus in the world as of the beginning of 2007. Next, I will briefly summarize a series of studies using the *NICT JLE Corpus* focusing on the description of various aspects of Interlanguage performance. Finally, the future prospect of research and teaching based on oral learner corpora will be discussed.

2. A brief historical overview on learner language studies

The study of learner language is not new. In the late 19th century, child language acquisition researchers examined the interaction protocol data between a mother and a child. In second language acquisition research, it was not until the middle of the 20th century when Pit Corder wrote the seminal paper called "Significance of learner's errors", in which he stressed the importance of shifting our research focus from mere comparisons of source and target languages, which was then called Contrastive Analysis (CA), to the study of the independent system of learner language by investigating the systematic nature of learner errors (Corder 1967).

This lead to the growing body of research called Error Analysis (EA) in the 1970s. Despite the large number of studies, most research was lacking in the notion of treating the production data in its entirety. In other words, most researchers in those days analyzed the production data in such a way that they looked at only those points that they wanted to examine, thus discarding the data after the data analysis[1].

It was not until the beginning of the 1990s when people started to collect the language production data with a view to sharing it with others in the same research community. In the U.S., child language acquisition data started to be gathered for the project CHILDES. In Europe, the influence of corpus linguistics was seen in several different ways. One was the initiative taken by publishers such as Longman, which started to gather second language learners' writings as a corpus in the early 1980s, which later led to the *Longman Learners' Corpus* (LLC). Second, the *International Corpus of English* (ICE) project was launched in 1990, where they decided to collect not only regional varieties of English, but also a corpus of learner English as one of the varieties of English (Greenbaum 1996), which eventually became one of the first initiatives of collecting learner corpora worldwide, i.e., the *International Corpus of Learner English* (ICLE). Third, a series of conferences such as *Teaching and Language*

1 For further detail, see Tono (2002).

Corpora (TALC) provided the opportunities for corpus linguists and researchers in the field of language teaching and learning to discuss the potential of corpus-based approach in various aspects of language education. The present author was one of those who benefited greatly from those meetings. By the beginning of the Millennium, learner corpus projects became the mainstream of corpus applications for language teaching and learning.

3. Major learner corpus projects

Nowadays, we can see a growing number of researchers working on learner production data using a corpus-based approach. The number of major learner corpus projects is somewhat limited, however, because the corpus compilation takes a long time, tremendous effort and a large sum of money. Here I will introduce some of those major learner corpus projects.

There are two commercial learner corpora: the *Longman Learners' Corpus* (LLC) and the *Cambridge Learner Corpus* (CLC). LLC was launched around the mid-1980s, being thus one of the first learner corpora in the world. Now it has about 10 million words in size, and is composed of written compositions, exam scripts, and various other writings by more than 50 different nationalities.[2] CLC was a latecomer, but is growing very rapidly. It contains over 25 million words, and is composed of anonymous exam scripts written by students who took Cambridge ESOL exams all over the world. It currently contains scripts from over 85,000 students with more than 100 first languages and more than 150 nationalities. The unique feature of CLC is its learner error-coded scripts. According to their website,[3] approximately 13 million words (45,000 scripts) were tagged for errors.

---

2 For the use of LLC, see Gillard and Gadsby (1998).

3 <http://www.cambridge.org/elt/corpus/learner_corpus2.htm>

Another important learner corpus project in Europe is the *International Corpus of Learner English* (ICLE) (Granger 1998). ICLE started in 1990, aiming to describe the Interlanguages of homogeneous groups of English learners (the third- or fourth-year students majoring in English as a foreign language at a university level) with approximately 15 different mother tongue backgrounds. Their primary goal is to identify similarities and differences in overuse, underuse or misuse phenomena across different first language (L1) background groups, which, they hope, will clarify universal vs. L1-related developmental patterns in second language.

Granger also launched a project of compiling an oral corpus of learner English in 1995, called LINDSEI (*Louvain International Database of Spoken English Interlanguage*). There are currently 11 different groups of learners from different L1 backgrounds, each of which contains 50 transcripts of 15-minute oral interviews.

In Asia, various projects for compiling learner corpora arose and disappeared in the last decade, and now a few of them survived. The *HKUST Learner Corpus* is one of the first learner corpora built in Asian regions. It contains more than 25 million words of exam scripts and term papers written by Chinese-speaking learners of English. John Milton has developed the web learning materials as well as a concordancer (Word Pilot) based on this corpus (Milton and Chowdhury 1994, Pravec 2002). In Japan, the present author has been involved in developing two major learner corpora: the *NICT JLE Corpus* and the *JEFLL Corpus*. The *NICT JLE Corpus* is an oral learner corpus of more than 1200 Japanese-speaking learners of English, based on the oral proficiency interview test transcripts. The *JEFLL Corpus* contains more than 10,000 Japanese secondary school students' writings (approximately 0.7 million). These two corpora are the biggest second language developmental corpora in the world, in the sense that the subcorpora are controlled by the proficiency level guidelines (either the test grades or the school years).

4. The NICT JLE Corpus: its design criteria

This section describes in more detail the *NICT-JLE Corpus*, the first oral corpus of EFL (English as a Foreign Language) learners in Japan.[4] The project started in 2001, funded by the Japanese government and led by the National Institute of Information and Communications Technology (NICT). It contains close to 1300 examinees' interview transcripts, which is approximately 2 million words in size. What makes this corpus quite unique is the fact that each subject is tagged for his or her oral proficiency test score based on the *Standard Speaking Test*[5] (SST), thus making it possible to compare across groups of different proficiency levels. SST is modeled after ACTFL (The American Council on the Teaching of Foreign Languages) OPI. There are nine levels, ranging from Level 1 for Novice Low to Level 9 for Advanced. Each interview test lasts 15 minutes, in which there are five stages: (1) warm-up, (2) the task for eliciting simple present tense narration, (3) the task for eliciting questions or testing the ability to negotiate in English, (4) the task for eliciting simple past tense narration, and (5) wrap-up. The three stages in the middle involve tasks such as picture descriptions, role plays or story-telling. The interviewer will decide the tasks spontaneously as they interact with the subjects. Since SST is a test disguised as a natural conversation, it is by nature interactive and adaptive to the examinee's profile. The recorded interviews are then evaluated by at least two certified SST raters, following the SST evaluation scheme.

The project also provided added value to the corpus by supplying a partially error-tagged version and a comparable corpus, in which the same interview tests were administered to native speakers of English. It also provides the version of a back-translation corpus, the one whose original English interview transcripts were translated into Japanese in order to investigate the first language influence. They also developed some useful tools for the project: (a) *TagEdit*, an editor for

4 For general introduction, see Tono (2001).

5 The Standard Speaking Test was developed by ALC Press based on ACTFL OPI.

supporting oral data transcription (Fig.1) and error annotation, which manages transcriptions and tag insertions, and (b) *Analyzer*, a tool for basic concordancing and frequency and statistical analyses (Fig. 2).

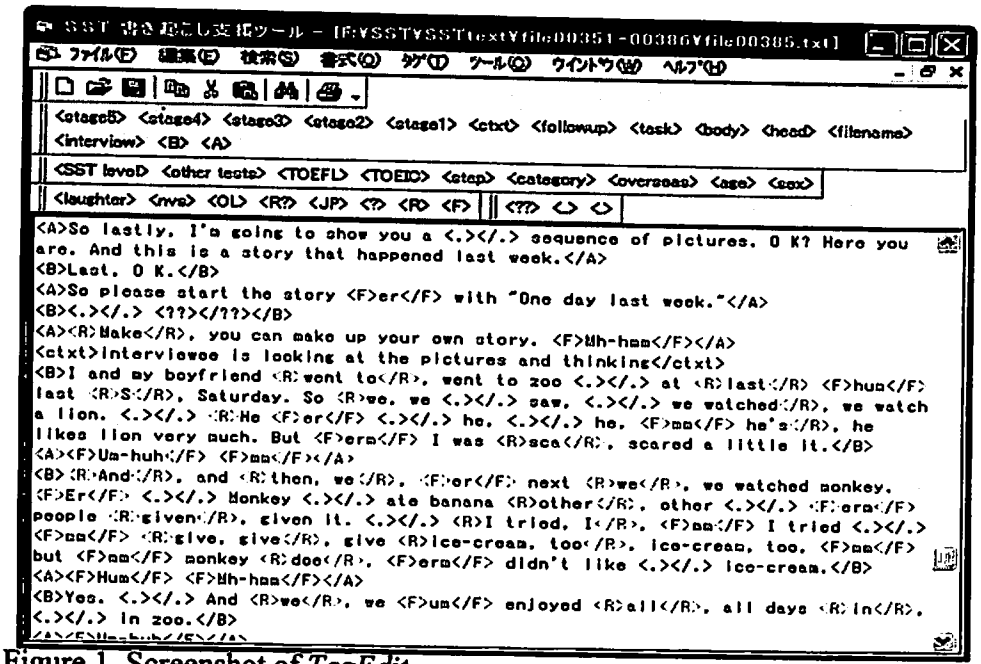

Figure 1. Screenshot of *TagEdit*.

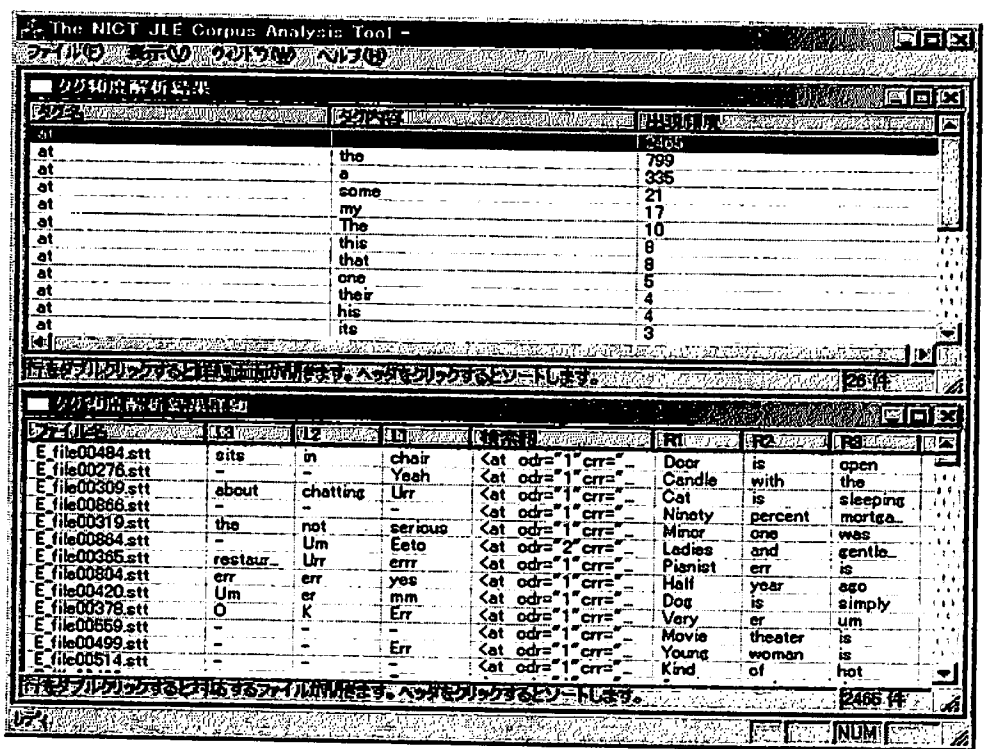

Figure 2. Screenshot of *Analyzer*.
(Window showing the analysis of article errors.)

5. Some findings using the NICT JLE Corpus

This section will describe on-going research into a spoken learner language, using the NICT JLE Corpus. Since the corpus was developed in collaboration between humanities researchers and Natural Language Processing (NLP) researchers, it has involved an interesting mixture of psycholinguistic, computational and pedagogical research. I will report mainly on the following two aspects: first, various research projects to identify the characteristics of spoken Interlanguages at different proficiency levels. Secondly, I will briefly summarize NLP people's research into (a) automatic error detection and identification, and (b) automatic identification of speakers' proficiency levels.

5.1. What basic text characteristics tell us

Tono (2004) examined basic text characteristics of the different proficiency level groups of the NICT JLE Corpus.[6] The measures include (a) average corpus size (the total size of the subcorpora (Level 1-9) divided by the number of subjects in each level), (b) Standardized Type/Token Ratio (STTR), and (c) Mean Length of Utterances (MLU). See Table 1 for the results. The mean sample corpus size for Level 1 was 338 words, and it increased to 790 words at Level 3, 1412 words at Level 6, and 1715 words at Level 8 respectively. This indicates that the total amount of speech made in the given time (15 min) is a strong indicator of the proficiency levels. In STTR, there was a sharp increase between Level 1 and 2, then followed by a gradual increase from Level 2 up to Level 7, which shows that STTR, an index of lexical density, distinguishes the low-mid groups effectively while it is not useful for discriminating the upper levels. MLU also shows a similar tendency as STTR. It increases constantly from Level 1 to Level 6, and then seems to reach the ceiling. It seems that in order

6 The version used for Tono (2004) was a pre-release version of the corpus, and it contained 1,313,293 words (1,201 examinees' utterances only).

to discriminate the upper levels (Level 7-9), we need to identify different types of Interlanguage features.

	1	2	3	4	5	6	7	8	9
Mean sample corpus size	338	457	790	1060	1298	1412	1505	1715	1632
Standardized TTR	36.67	42.03	42.48	44.20	46.22	47.58	49.02	49.00	49.35
MLU	3.09	4.04	5.90	7.44	8.44	9.00	9.14	9.25	9.24

Table 1. Basic text characteristics of the NICT JLE subcorpora.

5.2. Patterns in word/POS n-grams

One way to examine overall text characteristics is to obtain n-gram statistics for words and parts-of-speech. Figure 3 below shows the bird's-eye view of the top 100 trigrams of each level of the NICT JLE Corpus.[7]

Let me summarize some interesting findings from this diagram:

➢ The use of fillers is very frequent in the lower-proficiency groups, which gradually decreases to the minimum at the upper levels. The types of the fillers were also different; at the lower levels, fillers sound very Japanese (e.g. EETTO) but later sound more like English (e.g. erm, Uh-huh), and lexically more target-like and sophisticated (e.g. well, how can I say).

➢ The trigram patterns of "AT (article) + X" and "ADJ (adjective) + X" increased in number and types, which shows that learners tend to use more complex noun phrases.

➢ The trigram patterns of "PP*(personal pronoun) + X" and "V (verb) + X" also increased in number and types, which shows that learners have a tendency to use more complex predicate patterns, especially the use of modal auxiliaries and verb patterns.

7 This diagram was made by extracting the top 100 trigrams from each level, then sorted alphabetically. The patterns containing punctuations in the middle were deleted for the sake of simplicity.

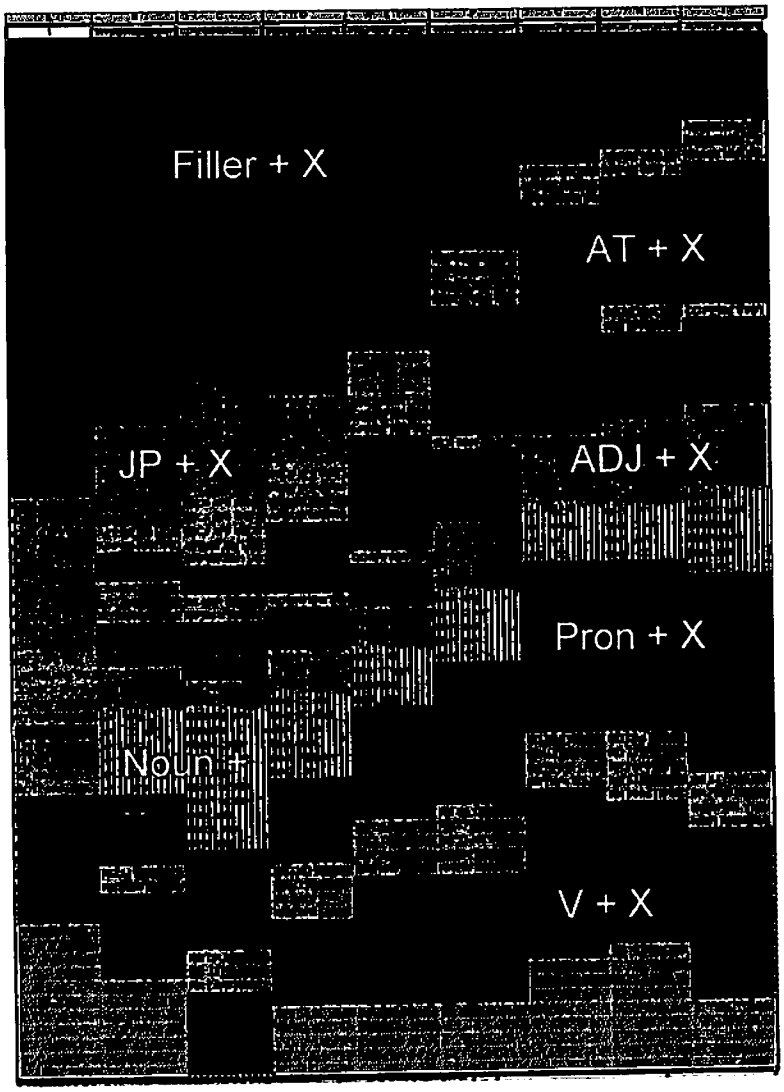

Figure 3. Trigram transition patterns across the NICT JLE Corpus.

These changes of trigram patterns are found to be much more frequent in oral learner corpora than written ones (Tono 2007).[8] This may be partly due to the fact that in the case of written essays, learners can spend more time monitoring their use of language as compared to speech, which helps them produce relatively more complex sentences from the beginning. It is worth noting, therefore, that attention has to be paid to different linguistic characteristics in evaluating oral and written performance.

8 Fillers are, of course, not usually observable in writings.

Patterns in word or POS n-grams are quite revealing, for it provides an overall picture of transition patterns of lexical as well as syntactic combinations across different stages of learning. By carrying out a thorough investigation on the similarities and differences in n-gram patterns between oral and written learner corpora, we will be able to describe the transitional structures of Interlanguage in a more systematic way.

5.3. Acquisition of verb subcategorization patterns

N-gram analysis is useful, but the technique is somewhat limited when it comes to examining a syntactic development. The analysis of verb complementation or subcategorization patterns, in particular, is a difficult job, for it involves grammatical categories such as noun phrases or prepositional phrases as a constituent, which cannot be captured by a simple n-gram analysis.

We conducted a preliminary investigation about the use of the verb "*get*" to compare the oral and written corpora, NICT JLE and JEFLL[9]. All the instances of the verb "*get*" were retrieved and classified according to the following verb categories:[10]

a. Basic structures (get + N; get + Adj; get + Part/Prep; get + Past Participle; get + Ving; get + to do)
b. Get + object + verb form (get + N + Ving; get + N + to do; get + N + Past Participle)

The classification was partly done automatically by searching for the POS patterns following the verb. The identification of noun phrases in learner production data was too difficult to automate, thus we categorized complex patterns by hand.

Tables 2 and 3 show the normalized frequencies (per 100,000) of subcategorization patterns for the verb "*get*".

---

9 I would like to thank Rie Suzuki, my postgraduate student, for her initial analysis of this work. Correspondence analysis was performed by the author.

10 The classification was based on Swan (2005).

JEFLL Level	Verb patterns of "get"							
	get + N	get + Adj	get + Part/Prep	get + p.p.	get + to do	get+ N to do	get + N p.p.	Active Margin
JH1	27.992	2.896	21.236	.000	.965	.000	.000	53.089
JH2	95.505	5.847	56.524	.650	.650	.000	.000	159.176
JH3	41.788	18.619	98.885	1.241	.000	.000	.414	160.947
SH1	34.886	18.660	36.509	2.434	.000	.000	.000	92.489
SH2	57.606	18.055	42.989	2.293	.573	.000	.000	121.516
SH3	108.982	26.439	46.430	2.579	2.579	.645	.000	187.655
Active Margin	366.759	90.516	302.573	9.197	4.768	.645	.414	774.872

Table 2. Correspondence Table. Patterns for the verb "*get*" (JEFLL).

SST Level	Verb patterns of "get"							
	get + N	get + Adj	get + Part/Prep	get + p.p.	get + to do	get + N to do	get + N p.p.	Active Margin
Level 3	17.694	1.083	13.000	1.083	.000	.000	.000	32.861
Level 4	27.559	4.039	17.106	2.970	.119	.000	.119	51.911
Level 5	35.876	10.250	24.297	4.176	.190	.000	.190	74.979
Level 6	48.062	22.491	24.955	6.162	2.773	.616	.000	105.059
Level 7	46.991	21.834	33.226	10.917	1.424	.000	.475	114.866
Level 8	61.078	32.888	42.285	12.920	8.809	.000	1.175	159.155
Level 9	53.215	38.225	61.460	11.243	5.996	2.249	3.748	176.134
Active Margin	290.476	130.810	216.328	49.471	19.311	2.865	5.705	714.966

Table 3. Correspondence table. Patterns for the verb "*get*" (NICT JLE)[11].

The first four patterns (get + N, get + Adj, get + Part/Prep, get + p.p.) show a steady increase in frequencies as the stages go up in both written and spoken corpora. In the case of the pattern "get to do", JEFLL shows an increase only in the last year (Year 12) while NICT JLE shows a gradual increase across the levels. Correspondence analysis was performed for the data from the NICT JLE Corpus in order to capture the relationship between proficiency levels and the use of different subcategorization patterns (see Figure 4 below).

In Figure 4, Dimension 1 explains 61.1% of the relationship between the two variables (SST level x Verb patterns of "get") and Dimension 2 explains 25.1%. As the diagram shows, Dimension 1 seems to indicate the proficiency levels, for the dots showing Level 3

11 Level 1 and 2 were omitted because there was only one occurrence of the verb *get* in these levels.

to Level 9 are plotted along the horizontal dimension. Dimension 2 can be interpreted as "the complexity of the verb subcategorization patterns". The two dots "get + N+ p.p." and "get + N + to do" are far away from all the groups, further down in the bottom right of the space, which indicates that these two verb patterns are seldom used by any of the groups. Level 9 is the only group slightly closer to these two categories. Thus, we can say that the presence of these two patterns in spoken data indicate that the speakers are very competent in using the verb "*get*".

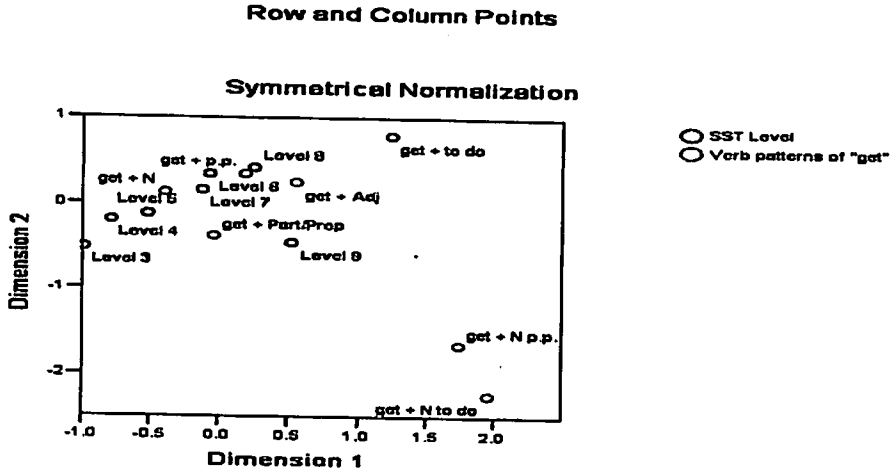

Figure 4. Correspondence analysis (NICT JLE).

In Figure 4, the use of the subcategorization pattern "get + N" is more closely associated with lower levels while such patterns as "get + Adj" and "get + Part/Prep" are plotted closer to upper levels. While all three patterns steadily increase along the proficiency levels, the frequent use of "get + Part/Prep" (i.e. phrasal verbs) or "get + Adj" patterns demonstrates that the speakers acquire the colloquial use of the verb "*get*", which resulted in closer plots between these patterns and upper levels. This is not the case with JEFLL; in the case of written data, the use of "get + N" is more prominent than other patterns. We can see that secondary school students have a rather limited repertoire of subcategorization patterns, as compared to the adult learners taking the SST. However, the JEFLL data shows that the pattern "get + N" is used with increasing frequency, which indicates that the learners become able to use different nouns within this pattern. In other words,

the patterns of use in spoken and written learner corpora are different; in speech, a wider range of subcategorization patterns become available as the level goes up. In writing, on the other hand, the repertoire of lexical choices within the simple pattern "get + N" becomes wider, whilst the subcategorization patterns are rather limited throughout the different developmental stages.

5.4. Automatic detection of learner errors

The NICT JLE Corpus has been used also by NLP researchers in order to apply natural language processing techniques to solve practical problems regarding learner language analysis. One of the areas is automatic detection of learner errors. Izumi, Uchimoto, and Isahara (2004), for example, investigated the possibility of identifying three types of learner errors (omission, replacement, insertion) based on the error-tagged data of the NICT JLE Corpus. They considered error detection as similar to text categorization and applied the machine learning model to solve this task. For this, they selected the Maximum Entropy (ME) model. By using the original error-tagged corpus as the training data, they obtained the recall of article errors to be approximately 35 percent and the precision to be around 48 percent, which was not very encouraging. In order to improve the results, they added corrected sentences and artificially-made errors, which led to better recall and precision (43 percent and 68 percent respectively).

5.5. Automatic identification of learners' proficiency levels

Another area of application in NLP using the NICT JLE Corpus is to automatically identify the proficiency levels of learners using linguistic features in the speech data. Since every individual file in the NICT JLE Corpus has proficiency level information as meta-data, it would be interesting to explore the possibility of how well computers can identify proficiency levels by looking at features of each speech data. Sakata et al. (2007) used the measurement called BLEU, which is proposed as a method of automatic evaluation of machine transla-

tion. It basically compares n-grams of machine translated texts against human translation and yield the rating (0.0 to 1.0 scaling) based on the overlapping ratio of n-grams between the two.

In the same vein, Sakata et al. used word n-grams of the speech data labeled with different proficiency levels in the NICT JLE Corpus in place of human translation, and proposed two ways of automatic evaluation using BLEU: one is to use word n-grams only, and the other is to use word and POS n-grams, whose weight was optimized by Support Vector Regression. The results show that the method using word and POS n-grams as feature vectors outperformed the word only method. Using this method, the accuracy rate of identifying the proficiency levels is approximately 65%.

If this type of research is carried out further, it would be possible to see the computer system for the future, which can determine the learner's proficiency level based on the input text features. This will be potentially very useful in developing the CALL system, where learners are guided to different levels of tasks based on their language proficiency levels. The same thing can be possible with human speech in the future, if speech recognition system is fully implemented. The NICT JLE Corpus can serve as an invaluable resource for such a direction of research.

6. Pedagogical implications

So far I have presented some findings based on the NICT JLE Corpus. It is exciting to see a growing number of studies conducted on various aspects of oral learner corpora; description of various linguistic features at different stages of L2 development, comparison between oral and written performance of the learners, computational analysis of learner corpora for automatic identification of errors and proficiency levels. Each of these findings will surely lead to a new paradigm of foreign language teaching and learning. To conclude this paper, let me discuss pedagogical implications of the studies mentioned above.

First, as we know more about the learning process, we will be able to adjust the learning environment to an appropriate learning level. Most foreign language teaching materials thus far do not take into account a scientific analysis of Interlanguage process. They largely rely on teachers' own experience and their forerunners' wisdom. I used the frequency data from the British National Corpus for developing my TV English conversation program for NHK (Nihon Hosou Kyokai: Japan Broadcasting Corporation), which ran from April, 2003 to March, 2006 and more than one million people watched the show. The book based on the program, called *Corpus Renshu-cho* (Corpus Drill Book) became a best-seller. I realized that people knew the value of such resources, if they were properly packaged and presented. Since I did not use learner corpora for my program, I will definitely produce something in the future, using spoken learner data. The analysis of oral learner corpora against written ones, together with the comparison with native speakers' corpora, will shed light on the learning path and the related problems on the way. This will benefit various areas such as syllabus construction, textbook writing, task design and creation, among others.

Second, teachers' attitude toward learner performance will change as more findings will be provided regarding the Interlanguage process. Examples of some features of the spoken learner corpus shown above (e.g., the transition of patterns in fillers or verb subcategorization) will provide teachers with a different viewpoint. The way a teacher evaluates students will be different with such insights into their language use.

Third, learners themselves can benefit from accessing the oral corpus directly. After my corpus-based TV programs, many teachers started to use corpora in the classroom. They make exactly the same comments: "We need corpora which are more accessible and easy to read." Corpora tuned to novice-intermediate learners are still hard to find. Oral corpora are also beneficial for learners because they are usually simpler than written corpora and students can learn useful colloquial expressions from them. Most spoken corpora, however, are often too difficult or too natural to understand the contexts fully. Therefore, I would like to propose an alternative: oral learner corpora with their corrected counterparts. Learners can access the oral learner

corpora and if they wanted to know the proper way to say it, they can access the corrected version of the corpus. This will be an excellent way to provide resources for writing classes or for preparing the speech or debate classes. I hope that research into oral learner corpora and their applications in the classroom should go hand in hand so that teachers and learners of foreign languages will get maximum benefit from this wonderful resource.

7. References

Corder, Pit 1967. The significance of learners' errors. *International Review of Applied Linguistics*, 5, 161-170.

Gillard, Patrick / Gadsby, Adam 1998. Using a learners' corpus in compiling ELT dictionaries. In Granger, Sylviane (eds) *Learner English on Computer*. Addison-Wesley: Longman, 159-171.

Greenbaum, Sydney 1996. *Comparing English Worldwide.* Oxford: Clarendon Press.

Izumi, Emi / Uchimoto, Kiyotaka / Isahara, Hitoshi 2004. Jido Eigo Ayamari Kensaku System No Kaihatu. (Developing an automatic error detection system for English learner corpora). In Izumi, Emi/ Uchimoto, Kiyotaka / Isahara, Hitoshi (eds). Nihonjin 1200 Nin no Eigo Speaking Corpus (A spoken corpus of 1200 Japanese-speaking learners of English), Tokyo: ALC Press. 141-153.

Milton, John and Chowdhury, Nandini 1994. Tagging the Interlanguage of Chinese learners of English. In Flowerdew, Lynne / Tong, Anthony K.K. (eds) *Proceedings of Joint Seminar on Corpus Linguistics and Lexicology*, Guangzhou and Hong Kong, 19-22 June, 1993. Hong Kong: Language Centre, HKUST, 127-143.

Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal,* 26, 81-114.

Sakata, Kosuke / Shimbo, Hitoshi / Matsumoto, Yuji 2007. Automatic estimation of language proficiency levels based on corpora.

(original in Japanese). ANLP. *Proceedings of the annual meeting of the Association for Natural Language Processing*, March 2007, 793-796.

Swan, Michael 2005. *Practical English Usage*. Oxford: Oxford University Press.

Tono, Yukio / Kaneko, T./ Isahara, Hitoshi / Saiga, Toyomi. / Izumi, Emi 2001. The Standard Speaking Test Corpus: a 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In Lee, S. (ed.) *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary. The Second Asialex International Congress*, August 8-10, 2001, Yonsei University, Korea. 7-17.

Tono, Yukio 2002. *The Roles of Learner Corpora in SLA Research: The Multiple Comparison Approach.* Unpublished Ph.D. thesis. Lancaster University.

Tono, Yukio 2004. On the use of productive vocabulary in the NICT JLE Corpus. In Izumi, Emi / Uchimoto, Kiyotaka / Isahara, Hitoshi (eds) *Nihonjin 1200 Nin no Eigo Speaking Corpus (A spoken corpus of 1200 Japanese-speaking learners of English)*, Tokyo: ALC Press, 97-112.

Tono, Yukio 2007. Comparing the NICT JLE Corpus with the JEFLL Corpus: Analyzing the word/POS-tag sequences. Paper given at the 29th JAECS (Japan Association for English Corpus Studies) conference. Dokkyo University, Kyoto, Japan, 28 April, 2007.