Studies in Corpus-based Linguistics and Language Education I
International Symposium
Corpus and Variation in Linguistic Description and Language Education
(2009年3月)　抜刷　非売品

Variability and Invariability in Learner Language:

A Corpus-based Approach

Yukio TONO

Global COE Program: Corpus-based Linguistics and Language Education (CbLLE)
Tokyo University of Foreign Studies

Variability and Invariability in Learner Language: A Corpus-based Approach

Yukio TONO

1. Introduction

It is S. Pit Corder who first showed us the significance of learner errors in order to appreciate the systematic nature of learner language (i.e. Interlanguage) (Corder 1967). This revelation led to many works of Error Analysis (EA) in 1970s, but soon people became aware that looking at errors only would not be sufficient and that the description of learner language in its entirety would be appropriate. Thus so-called Performance Analysis (PA) became a central theme of Interlanguage research, in which both correct and incorrect usages by learners become the objects of the study. Since then, it is probably fair to say that there has been no drastic paradigm shift in methods and practices, and much of the landscape described by Pit Corder, Larry Selinker, and Jack Richards among others remains unchanged. Recently, however, we are beginning to have another major breakthrough in the study of Interlanguage. With the advent of computer technologies, it is increasingly realistic to store a massive amount of learner production data on computer and analyze the texts using corpus linguistic or natural language processing techniques. This new area of analyzing learner language on computer is called *learner corpus research.*

Learner corpus research is an exciting interdisciplinary area, where natural language processing and corpus linguistics meet second language acquisition and foreign language learning/teaching. Theoretically, there is a growing awareness that a natural language has properties which are probabilistic in nature. A vast amount of linguistic texts can produce various probabilistic information about the combinations of lexis and structures, sometimes overt (words and words) and sometimes covert (the combination of parts of speech, for example). This probabilistic information is useful not only for a theory construction but also for pedagogical applications.

I have been working in this new field of learner corpus research for quite some time, and joined TUFS in 2007. As I joined the faculty, the university just launched the government-funded five-year G-COE project called CbLLE. As a core member of the team, my research group has been actively involved in the compilation of a new learner corpus called

'International Corpus of Crosslinguistic Interlanguage' (ICCI), which aims to gather corpora of younger learners of English ranging from upper primary school to secondary school children across more than 10 different nations as their mother tongue backgrounds. To my knowledge, this type of learner corpus has not yet been compiled so far.

In this paper, first I will briefly summarize some of the previous learner corpus studies my research group conducted and showed the potential of this research area. Secondly, I will discuss the nature of learner language in terms of the theme of this symposium, "language and variation". To this end, I will show various aspects of consistencies of patterns of development shown in Interlanguage data as invariability and at the same time will also exemplify individual or variational features across modes of speech and levels of proficiency as variability. Finally, I will discuss theoretical as well as pedagogical implications of learner corpus research and close my paper with a brief introduction to the new project ICCI.

2. Learner corpus research

My research group has been working mainly in three major areas of research: (1) compilation of learner corpora, (2) analysis of learner corpora and SLA theory construction based on the analysis, and (3) applications of learner corpus data for pedagogies and practice. I will overview the first two in this section.

2.1. Compilation of learner corpora

In learner corpus research, like other corpus linguistic studies, compiling a learner corpus is a very important project in itself. I have been involved in two major corpus building projects for Japanese-speaking learners of English. One is called the NICT JLE Corpus, which is a 2-million word spoken corpus of Japanese learners of English. I initiated the project, but the funding came from the National Institute of Information and Communications Technology (NICT), so NICT took over the remaining work and completed it in 2004. It is a collection of more than 1,200 subjects' oral proficiency interview test transcripts. The test is called the Standard Speaking Test (SST) developed by ALC Press, which is a customized version of the ACTFL Oral Proficiency Interview (OPI). The SST consists of five parts in a 15 minute interview; warm-up, picture description, story-telling, role play and wind-down. Each interview script has an individual proficiency score, which has nine levels: beginner (level 1) to near-native (level 9). The corpus is now available as a book with a CD-ROM (Izumi et al 2004) and also in electronic format under license.

The other one is called the JEFLL Corpus, which is a collection of more than 10,000 Japanese secondary school students' English compositions. The corpus contains timed in-class free compositions in English on six different topics (argumentative or narrative). Each task was given as part of regular classroom activities, not homework. The subjects were not allowed to use a dictionary while writing. We encouraged spontaneous production in writing, and if there were any words they could not come up with in English, they were allowed to write them in Japanese. The average length of each essay is rather short (about 60-70 words), but with more than 1,000 compositions in each category, we could approximate the patterns of use and possibly the path of learning. The JEFLL Corpus is now publicly available via the Shogakukan Corpus Network (http://www.corpora.jp), where you can access the corpus via the web-based query tool.

2.2. Analysis of learner corpora

Another important area of research is the analysis of learner language using the corpora mentioned above. Some of the major objectives are (1) the description of Interlanguage development in terms of overuse vs. underuse as well as correct use vs. misuse of certain linguistic features, (2) the identification of criterial features which distinguish one proficiency level from another, and (3) the development of the list of language features which are very slow to learn and needs some revision or modification in teaching syllabi or methodologies. Theoretically we are interested to find those patterns of overuse/underuse/misuse which are 'specific to learners' L1 knowledge and those which are universal, applicable to every learner from different L1 backgrounds. In this way, we could possible redesign the syllabus adjusted to the L2 learning path and ask people working on action research in the classroom to test the effects of such modifications.

Figure 1 shows different types of corpus designs for different research questions. The European project such as the International Corpus of Learner English (ICLE), in which they compile corpora of university EFL learners' written essays from 20 different L1 backgorunds, focuses on the comparisons between L2 learners with different L1 backgrounds (IL-a, b, c, d, etc.). Their primary interest is to distinguish L1-related Interlanguage phenomena from universal ones. They are also interested in the comparison between non-native speakers and native speakers in order to describe the "foreign-soundingness" of learner language.

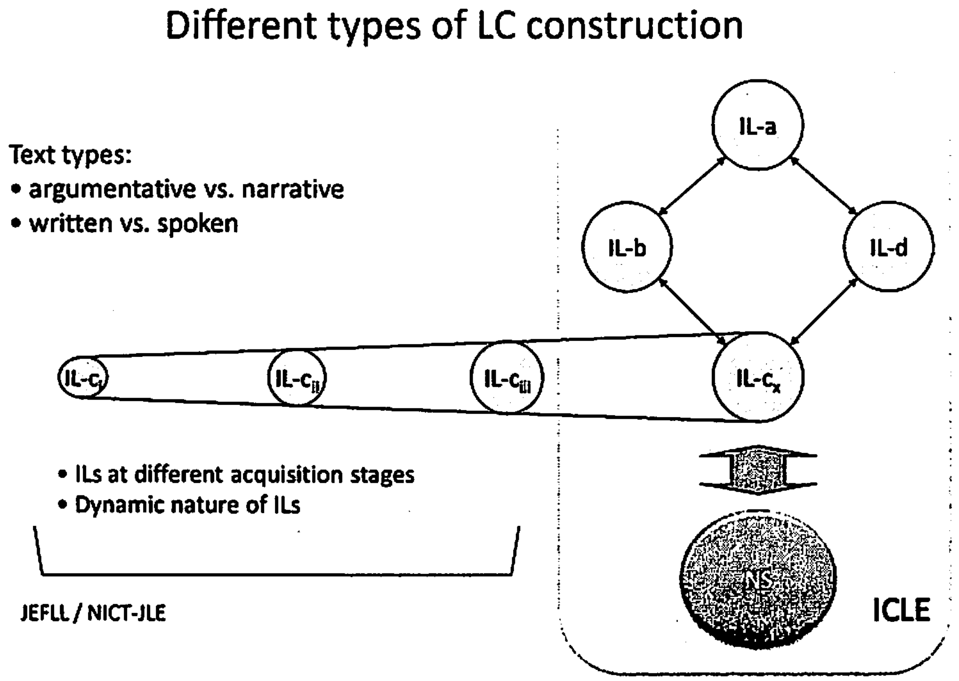
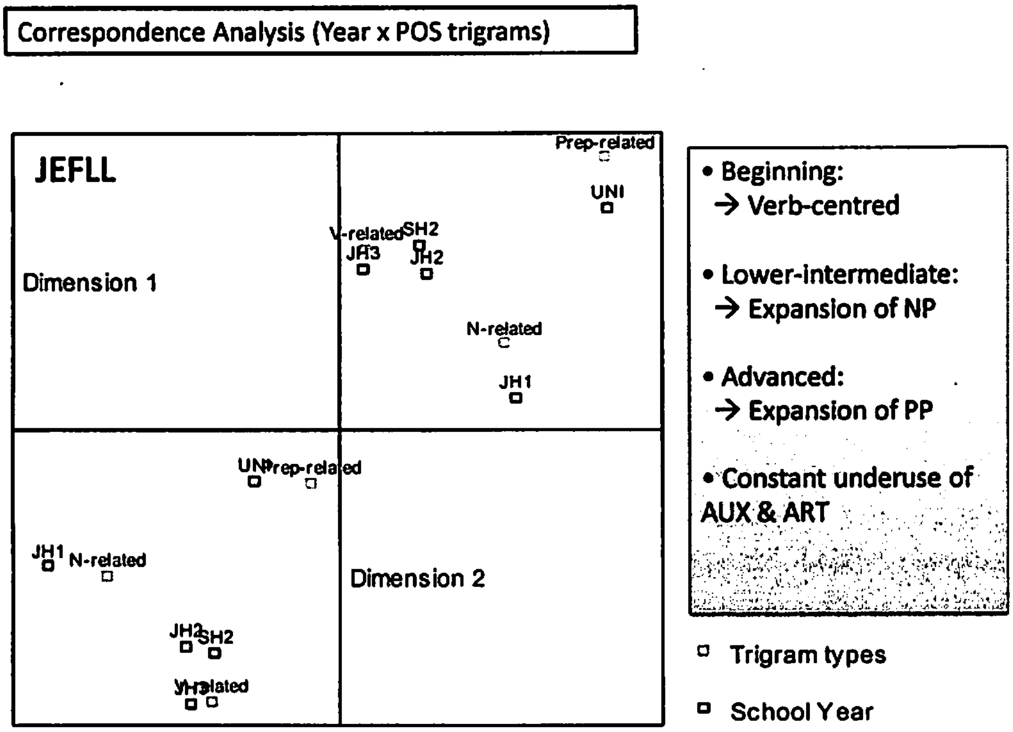


Figure 1. Different types of learner corpus construction

We are interested in different types of learner corpora (see the set of Interlanguage data arranged horizontally in Figure 1). We construct corpora representing Interlanguages at different acquisition stages in order to describe the dynamic nature of Interlanguage development. These corpora should also be designed in such a way that they represent different text types (e.g. argumentative vs. narrative in JEFLL) or modes of text (written in JEFLL vs. spoken in NICT-JLE). By using these different sets of learner data, we could possibly investigate various aspects of learner language in different modes at different learning stages.

We will have an overview of some of our research results. Our research group is currently working on the description of the Interlanguage features, especially focusing on the syntactic complexities across proficiency. There are mainly two threads of research going on at the moment; one is to compare the frequencies of complexity measures (sequences of part-of-speech tags or parsed units) across proficiency levels and the other is the transition of error patterns across proficiency. Let me first give a brief review of the first type of studies and then move on to the second.

Tono (2000) investigated the relationship between the subjects' school years and frequencies of part-of-speech (POS) tag sequences (three sequences of tags = trigrams) in the JEFLL Corpus. By looking at POS tag sequences, we can observe the frequent patterns of use in the sequences of part-of-speech categories, which helps to understand the process of acquiring the syntactic patterns in the target language. This was done by tagging the learner data with POS information and extracting the tag sequences automatically, and then performed a data reduction statistical procedure called Correspondence Analysis over the frequencies of tag sequences across different school-year groups. The results are shown in Figure 2. The analysis shows that the beginning level has a tendency to be more closely associated with verb-related patterns ('V-related' in the diagram), while N- or Prep-related trigrams are more closely associated with lower-intermediate and advanced learners respectively, which clearly shows that more advanced students have a tendency to have more complex noun phrases or prepositional phrases. We also observed constant underuse of auxiliaries and articles. This is one of the first corpus-based studies in learner corpus research to confirm the transition of different syntactic features characterizing different stages of acquisition.



Correspondence Analysis (Year x POS trigrams)

- Beginning:
  → Verb-centred
- Lower-intermediate:
  → Expansion of NP
- Advanced:
  → Expansion of PP
- Constant underuse of AUX & ART

Figure 2. Tono (2000): Analysis of POS tag sequences across proficiency

In the same vein, Kobayashi (2006) performed Correspondence Analysis over single POS tags across different proficiency groups in the NICT-JLE Corpus. As Figure 3 shows, there is a distinctive tendency that the lower proficiency level groups are more closely associated with POS tags belonging to noun categories (NN*, NP*) while upper-proficiency level groups are linked with POS tags featuring verbs (V*). Please note that this is a single tag distribution and not the trigram patterns shown in Tono (2000). He shows that at the beginning stages of acquisition, learners tend to rely on the use of nouns more, which is partly the tendency of spoken data, but it also coincides with the findings above in Tono (2000) that the beginning-level students tend to make more verb errors, especially agreement and omission errors. Gradually, however, advanced learners were found to become able to use more lexical verbs consistently in the utterances, which resulted in more occurrences of the verbs in the utterances compared to the lower levels (see the left circle of Figure 3).

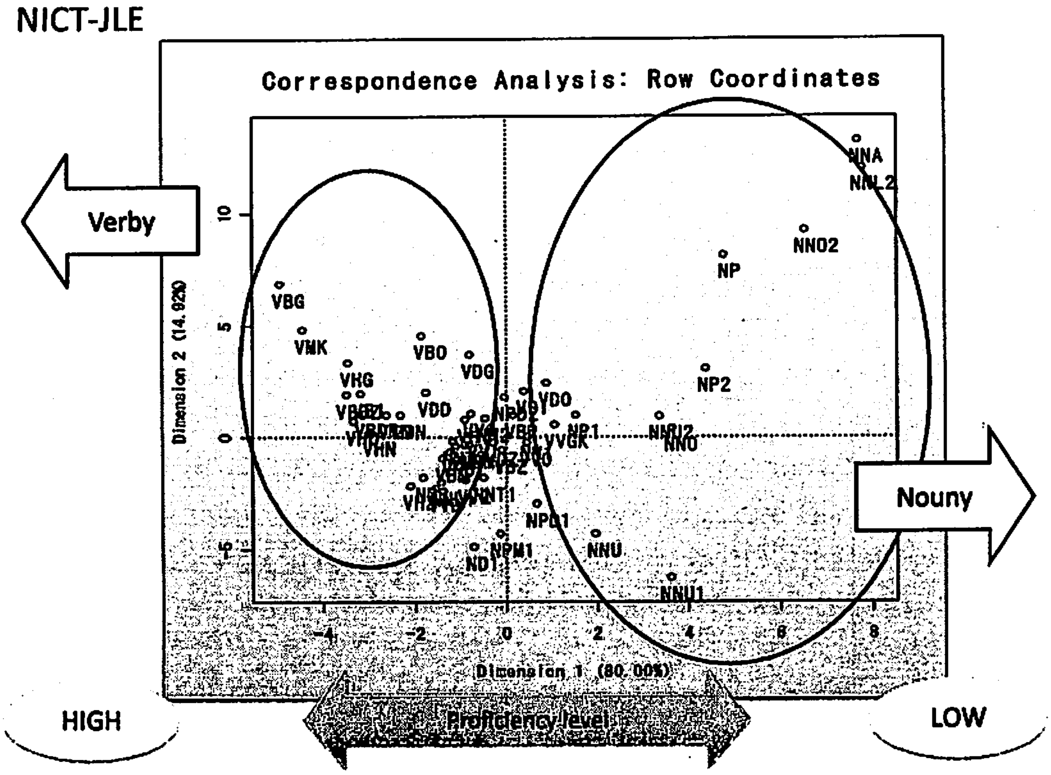


Figure 3. Kobayashi (2006): Correspondence analysis over single POS tags in NICT JLE

Kaneko (2006) also reported that the internal structures of noun phrases are closely related to the developmental stages. She manually parsed the data in the NICT-JLE Corpus for noun phrase boundaries and performed Corresponding Analysis to see the relationship between the frequencies of different types of noun phrases and the different proficiency groups (see Figure 4). The analysis shows that simpler NPs are more closely associated with lower-intermediate learners while the nouns followed by prepositional phrases or *that*-clause constructions are strong indicators of characterizing advanced learners. This is another evidence confirming the relationship between structural complexities and the acquisition stages.

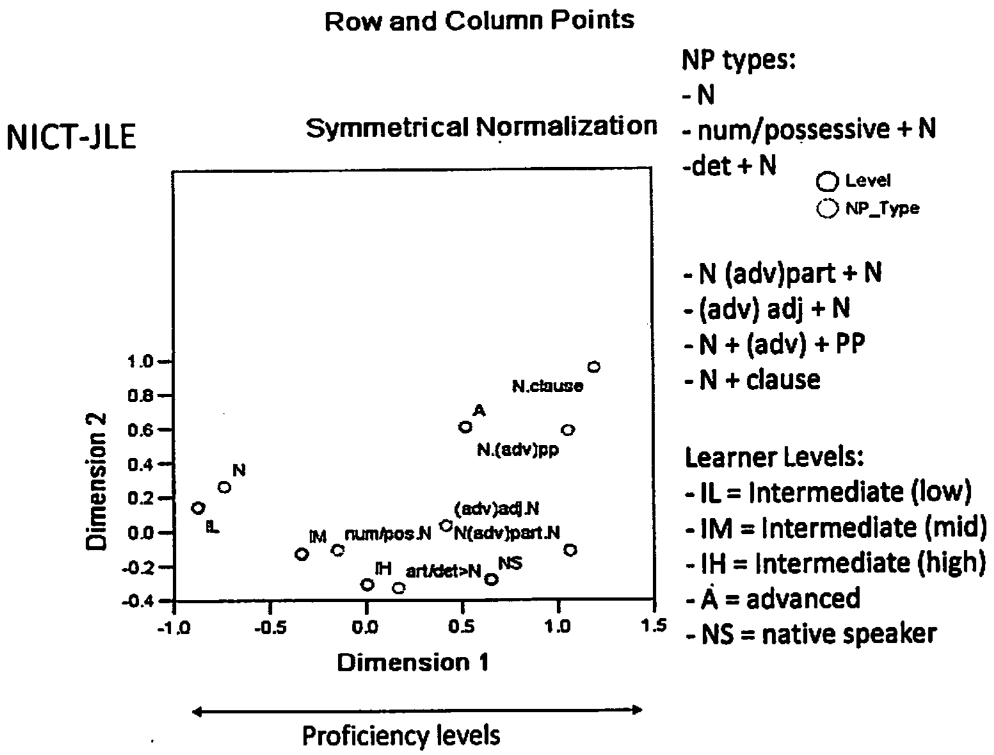


Figure 4. Kaneko (2006): Correspondence analysis over NP structures in NICT JLE

These findings seem to be rather obvious to some people, but the point is that this kind of thorough descriptions of learner language for different lexico-grammtical features in light of attested language use data will empirically verify the claims that we just trust by faith or from experience. By investigating each feature characterizing different stages of acquisition, we could come up with a better solution for profiling learner language.

Let me move on to the second area of studies, that is, the analysis of error frequencies across proficiency. Tono (2000) was one of the first learner-corpus-based studies of acquisition order of grammatical items. Tono replicated well-known English grammatical morpheme studies by Dulay & Burt (1972, 1974) and found that there was a certain degree of similarities in the order of acquisition. However, Japanese learners showed very distinctive tendencies that the article system is acquired the latest, and that possessive marker *–s* is acquired relatively earlier. In the so-called universal order of acquisition, the article system is supposed to be acquired in the middle of the acuiqisition order, while possessive *–s* is acquired very late. As is shown in this study, the article system is found to be a very difficult item for the Japanese, because there is no article system in our language. On the other hand, possessive –s seems to be relatively easy to acquire because we have a genitive marker "*-no*", which behaves in a very similar way as possessive *–s*. These findings coincide with some of the previous empirical studies on Japanese EFL learners (cf. Shirahata 1988).

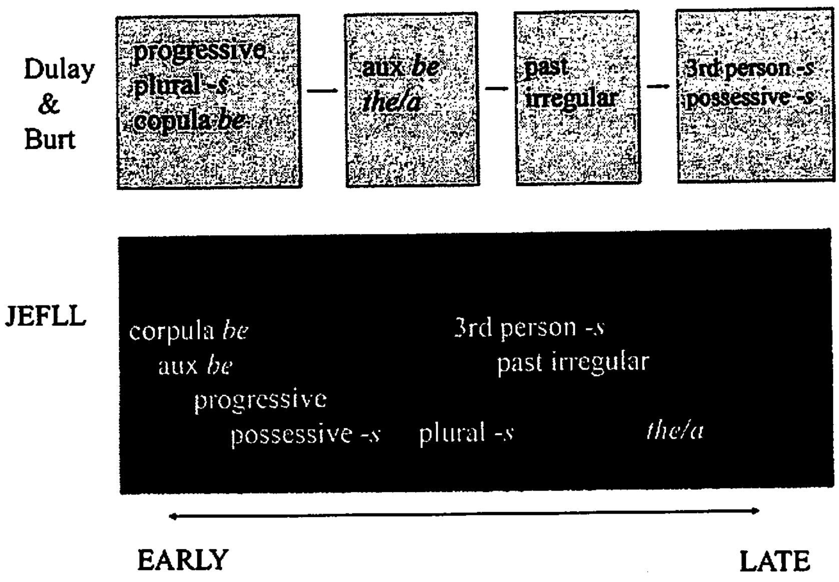


Figure 5. Tono (2000): Morpheme order study replicated using learner corpora

Abe and Tono (2005) investigated error distributions in the two corpora, JEFLL and NICT-JLE (see Figure 6). They made a manual annotation of errors over sampled essays and speech transcripts (10,000 tokens for each level) from JEFLL and NICT-JLE respectively. Then they classified verb errors into the followings : (a) tense, (b) aspect, (c) agreement, (d) inflection, and (e) lexical choice, and noun errors into (a') countability, (b') inflection, (c') case agreement, and (d') lexical choice. Performing Correspondence Analysis over the frequency counts of error tags in two modes of learner data (written vs. spoken) across proficiency levels (6 levels each for JEFLL and NICT JLE), they found that (1) in both written and spoken corpora, there was a distinct tendency that verb-related errors were closely related to lower-proficiency students while noun-related errors characterized higher-proficiency learner groups (cf. the two arrows shown on the right-hand margin), (2) spoken and written modes are plotted independently against each other, which shows the patterns of occurrences, especially error rates, were different in speech and writing, and (3), among more advanced learner groups, lexical choice errors of nouns are more significantly correlated with spoken modes and lexical choice errors of verbs, with written modes. This study was one of the first attempts at making a comparison of spoken and written learner corpora in terms of acquisition features acorss proficiency. While there is an issue of comparability between the two modes of corpora, it is worth noting that some of the error patterns, e.g. noun- vs. verb-related errors, exhibit very similar occurrence patterns across proficiency, which is an interesting finding in the sense that some errors occur persistently across different modes of performance. These persistent errors could be a good predictor for assessing proficiency levels.

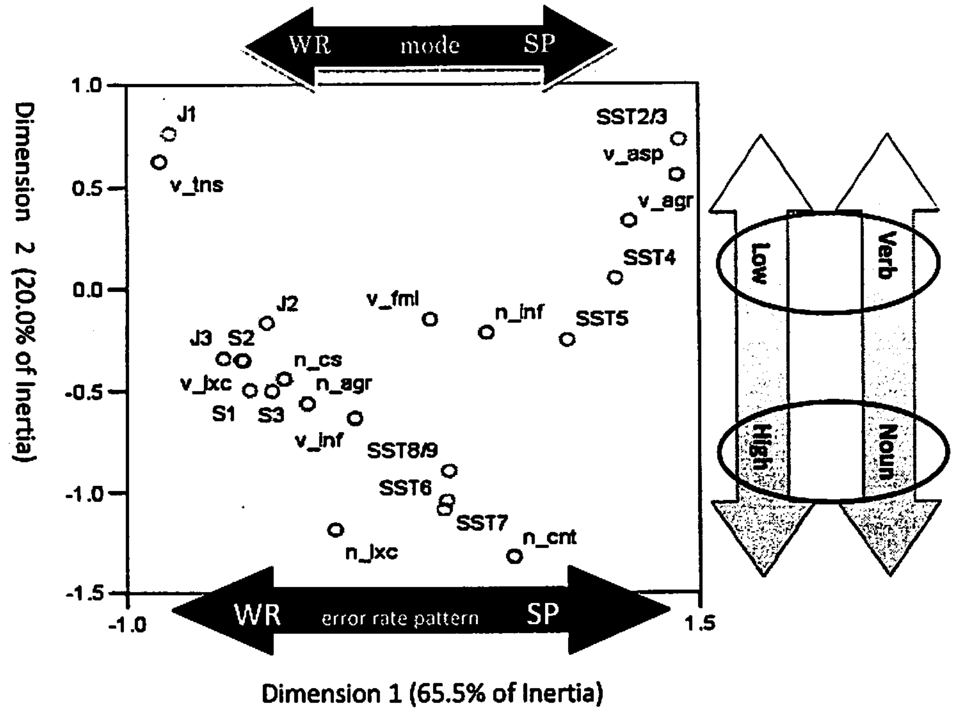


Figure 6. Abe and Tono (2005): Error distributions across proficiency in JEFLL and NICT JLE

3. Contributions of learner corpus analysis to SLA theory construction

So far we have seen some of our research group's findings using learner corpora. Whilst we need to investigate various aspects of Interlanguage features more extensively, these findings will surely contribute to a better understanding of SLA processes and its theory construction. In this section, I will describe the nature of variability in Interlanguage observed in our studies and its theoretical implications.

3.1. Interlanguage variability

Interlanguage variability is a notion first brought into light by Tarone (1979, 1983). She argues that learner utterances are systematically variable in two senses: linguistic context and the elicitation task. Following Labov's (1969) Observer's Paradox, she proposed the heterogeneous Interlanguage model, made up of a continuum of styles; there is a *careful* style on one hand, which is the style produced when the speaker pays the most attention to language form, and a *vernacular* style on the other end, which is defined as the style produced when the speaker pays the least amount of attention to language form (Tarone 1983: 152). In the analysis of learner corpora, this distinction between the vernacular and careful styles is worth noting, because we are sometimes confronted with seemingly conflicting results from different sources of data. Larsen-Freeman (1975) found that the rank orders of morphemes produced by second language learners varied across five tasks of 'speaking, listening, reading, writing and elicited imitation.' Tarone showed these phenomena as the effects of style-shifting in the Interlanguage system.

In our case, this is especially relevant when we make a comparison between spoken and written learner corpora. For example, more morphological omission errors are observed in spoken data than in written corpora, which is due to the performance task factor. Plural *–s* or third-person singular *–s* are the most notorious ones. Most Japanese EFL learners fail to properly supply them in speech even if they were at a quite advanced level. They perform differently, however, in writing. Usually advanced learners will supply these morphemes fairly accurately in writing. Thus, we need to take into account this type of variability of Interlanguage across task and proficiency.

3.2. Variability and invariability

Tarone discussed Interlanguage variability to refer to the stylistic changes of Interlanguage systems at a given stage of development. Variability, however, can be observed throughout different developmental stages of the Interlanguage system. As our research shows, there are some distinctive linguistic features characterizing each stage of L2 acquisition. Novice learners, for instance, tend to use more fillers in speech, and produce simpler sentence structures. They tend to omit verbs in their utterances, which results in a higher ratio of noun clusters per utterance.

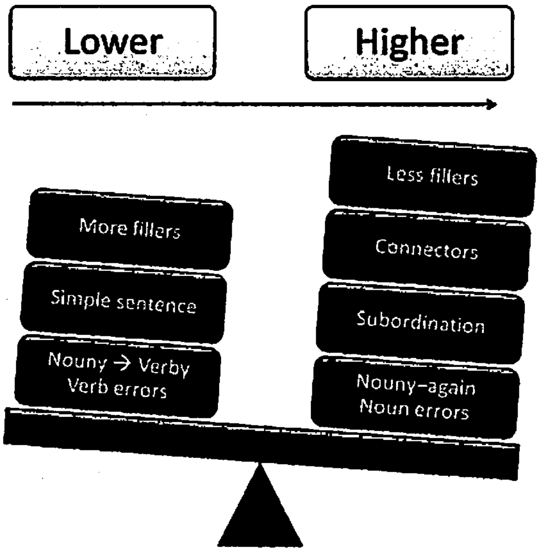


Figure 7. Variable and invariable nature of Interlanguage

At later stages, however, their utterances become more 'verby', as they learn to use verbs properly, which results in producing more verb morphological errors. Intermediate to advanced level learners, on the other hand, use less fillers and produce more complex sentences with subordinations and pre- and post- modifiers. Also the better use of connectors seems to be another unique feature for advanced learners. The sentences contain more nouns this time, producing more noun-related lexical errors, compared to verb errors at pre-intermediate stages. Figure 7 summarizes these results diagrammatically. As we look at the same phenomena from different angles, we can see something invariable or consistent here as an example of proficiency indices. For instance, the quantity of fillers can be always a good indicator for proficiency estimation. The same thing can be said about connector usages, a noun-verb ratio, and simple vs. complex sentence patterns. Therefore, it is crucial to observe not only something variable but also what is invariable.

The same kind of variability can be observed across different modes of production. In Figure 8, along the continuum of style shifting proposed by Tarone (1983), Interlanguage shifts toward simpler morphologies in the vernacular style, with frequent use of memorized chunks in speech. Toward the other end, in the careful style, learners supply grammatical morphemes more correctly and use prepositional phrases as nominal and verbal modifiers to make their utterances semantically more informative. Such style shifting is also another case of variability. Here again, however, we can look at this style shifting in terms of invariability. When looking at grammatical morphemes, especially verb morphology, it always serves as a reliable indicator of style shifting. Verb morphology plays an important role as a good estimator for style shifting. The same thing can be said about other phenomena such as memorized chunks and postmodification by prepositions. It is significant to find that there is an underlying invariability or consistency in those features which could serve as good indicators of style shifting.

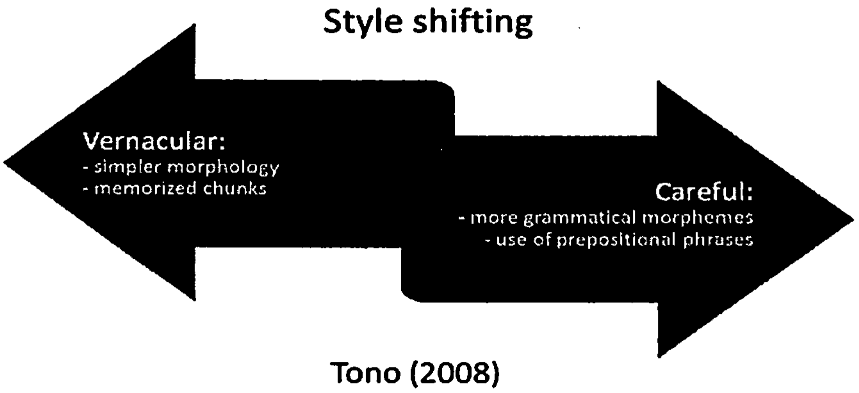


Figure 8. Style-shifting across different modes of production

Variability can be seen across different L1 backgrounds as well. Japanese, for example, is an SOV language, and synthetic. We have particle systems and have very flexible word orders because our language is agglutinative. Japanese also does not have an article system. This kind of L1 background could cause problems in acquiring English. Our research shows that the differences in the basic word order are not a very serious problem; once learners understand the difference, they usually cope with the word order pretty well from the beginning level. However, the choice of right verbs and the selection of appropriate verb complementation patterns are so difficult for Japanese learners, primarily because our language system has very different syntactic and semantic realization of verbs. Also we have difficulties in the use of English particles and prepositions to go with verbs. Auxiliaries are also constantly underused, which shows that learners have difficulties expressing modality using auxiliaries. The missing article system also causes a lot of problems in acquiring article systems in English. See Tono (2007) for further details.

What is invariable here, then? Universal error frequency comparisons show that the number of verb errors in English is universally more frequent than noun errors although noun errors are sensitive to a proficiency level. Also some errors such as article errors are always related to learners' mother tongue background. It is a well-known fact that if learners' L1 does not have an article system, it is very likely that the acquisition of articles is very slow. There are very interesting correlations between learners' L1 knowledge and L2 acquisition. Further corpus-based research will be needed here.

4. Implications

Here I will discuss the implications of learner corpus research from theoretical and pedagogical viewpoints. Learner corpus research will make a significant contribution to SLA and English language teaching in both theory and practice. Theoretically, we are interested in the possibility of explaining the acquisition processes found in our data by a probabilistic model of language. Since learner corpora can produce various probabilistic information in terms of morphosyntactic as well as lexico-grammatical features of learner language across proficiency, learner corpora could serve as very nice testbeds for a probabilistic model of SLA. We are exploring the possibility of using the Bayesian network model and the DOP model proposed by Rens Bod (1992, 1998) as heuristics. We are still wondering whether we should commit ourselves to exemplar-based models of language (Tomasello 2003; Goldberg 2006) or the hybrid of rule-based models with probability information. In any case, this probabilistic model of SLA will be the first of its kind, so we hope to extend our research in this new direction.

Secondly, I will make a brief comment on pedagogical practice. I was one of the first corpus linguists who had brought the notion of "corpus" to English classrooms and English teaching/learning communities in Japan. One such effort was to work for NHK (Nihon Hosou Kyokai, i.e. the Japan Broadcasting Center) to produce a corpus-based TV English conversation program. It consists of a set of one hundred 10-minute lessons, each of which features one English basic keyword (mainly basic verbs, prepositions, auxiliaries, conjunctions, and *wh*-markers) and show how the keyword is used in context. All the keywords and their collocates for practice were selected from the British National Corpus, a balanced corpus of 100 million-word British English. The program ran from 2003 to 2006, and turned out to be a great success. More than one million people watched the programs, and related textbooks and DVDs (see Figure 9) have been available in the market. Now English teachers become aware of the value of corpus evidence and the use of corpus data in the classroom. Indirect use of corpora for creating teaching materials like the ones I did, as well as direct use of corpus data in the classroom for Data Driven Learning (DDL) are some of the interesting possibilities to explore in the future.

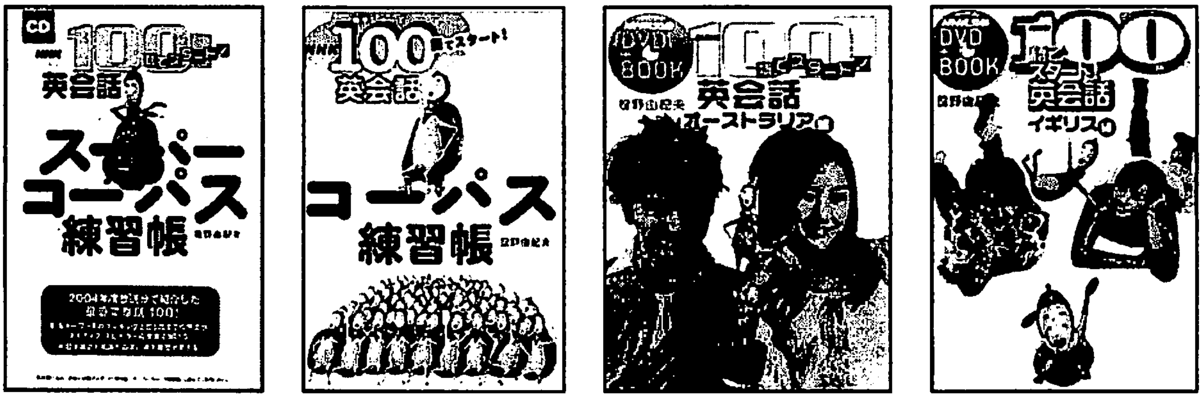
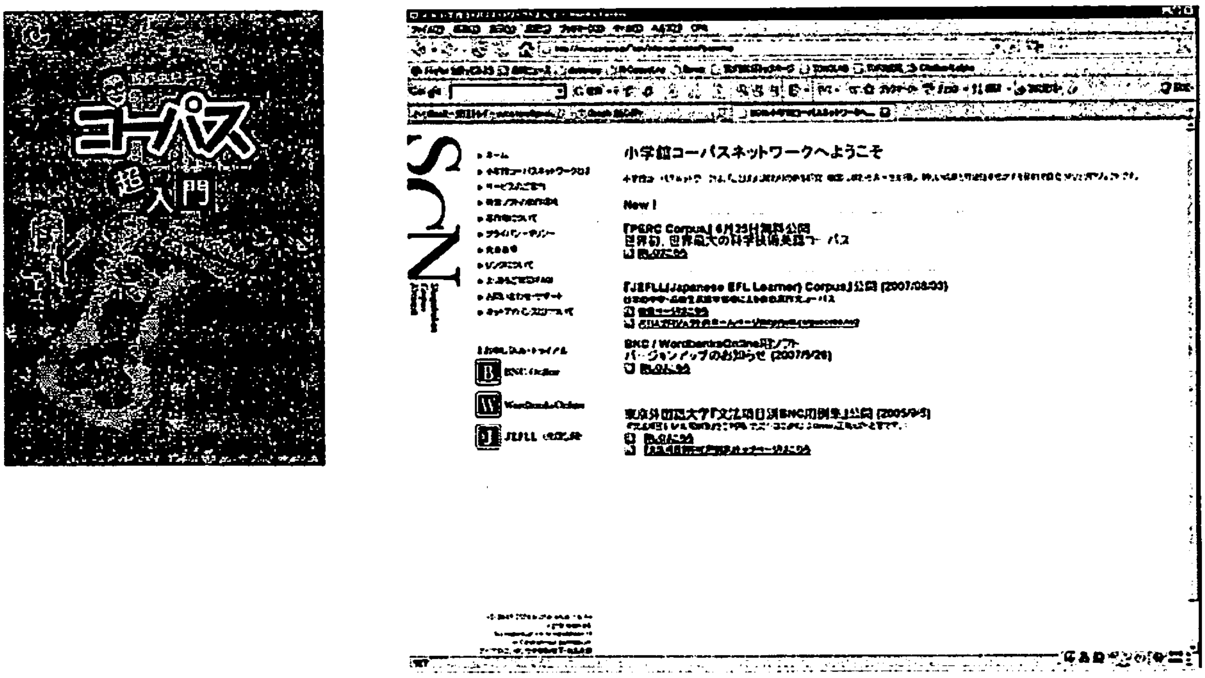


Figure 9. Corpus-based English conversation textbooks and DVD series written by the author

Another pedagogical practice that is worth noting is that I have been working with several companies to promote the use of corpora for English language teaching. Shogakukan, a major general-purpose publisher in Japan, for example, has been working with me to provide the web-interface for mega-corpora such as the BNC or the WordBanksOnline (the corpus used for the COBUILD project). It is called the Shogakukan Corpus Network (http://www.corpora.jp). I have developed the interface with the natural language processing unit inside Shogakukan for years. The motivation behind this is to provide laypersons who do not know anything about corpus linguistics can access the web corpus and do the search just like electronic dictionaries or Google. This is especially welcomed by English teachers who want to access corpus data for their preparation for teaching. Figure 10 shows the website of the Shogakukan Corpus Network and a recently published introduction to corpus linguistics for ordinary persons (Tono 2006).



小学館コーパスネットワークへようこそ

New !

「PERC Corpus」6月25日無料公開
世界初、世界最大の科学技術英語コーパス

「JEFLL (Japanese EFL Learner) Corpus」公開 (2007/08/03)

BNC / WordbanksOnline用ソフト
バージョンアップのお知らせ (2007/9/28)

東京外国語大学「文法項目別BNC用例集」公開 (2005/9/5)

Figure 10. An easy guide to Corpus and the Shogakukan Corpus Network.

5. Future directions

As a member of this new G-COE project, we have launched a new international learner corpus building project called the "International Corpus of Crosslinguistic Interlanguage" (ICCI). It is an international project of compiling a corpus of young learners of English, comparable to the JEFLL Corpus (JEFLL consists of essays written by junior and senior high school students in Japan). So far, most learner corpora available abroad are comprised of university students' data. Thus it is impossible to investigate earlier stages of acquisition using these corpora, which is the reason why we launched this new project. We will aim to compile corpora of younger learners (aged around 10 to 18) from different L1 backgrounds (Spain, Austria, Isreal, Poland, Taiwan, Hong Kong, Singapore, Korea, China, France, among others) with shared elicitation tasks in in-class free composition, comparable to the JEFLL Corpus (see Figure 11). This will provide us with great opportunities to compare Interlanguage corpora across proficiency and L1 background, which will verify the findings in JEFLL and NICT JLE and help develop more coherent theories of corpus-based SLA.

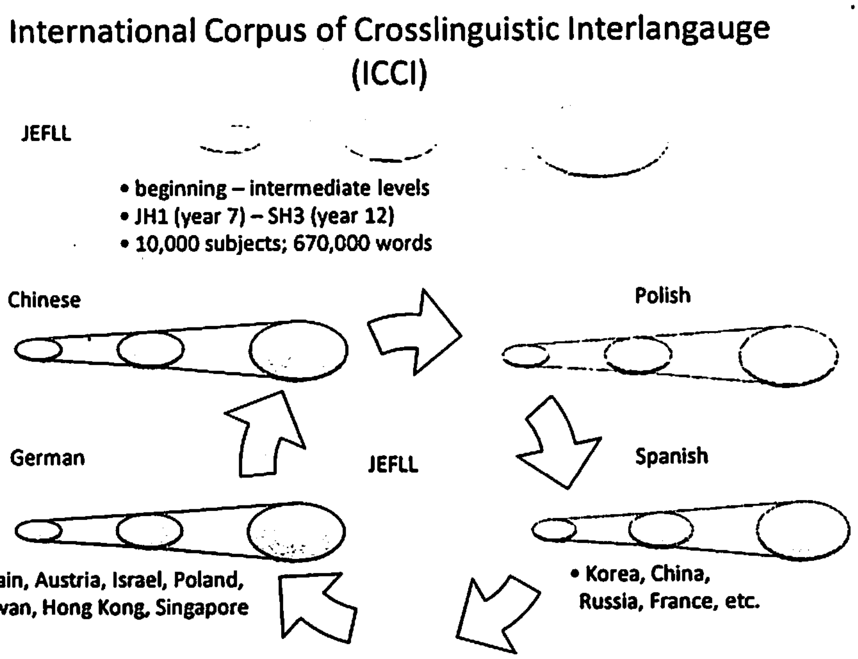

International Corpus of Crosslinguistic Interlangauge
(ICCI)
JEFLL
• beginning – intermediate levels
• JH1 (year 7) – SH3 (year 12)
• 10,000 subjects; 670,000 words
Chinese
Polish
German
JEFLL
Spanish
• Spain, Austria, Israel, Poland, Taiwan, Hong Kong, Singapore
• Korea, China, Russia, France, etc.

Figure 11. International Corpus of Crosslinguistic Interlanguage.

We had the first meeting at TUFS in February, 2008. About ten delegates from eight countries joined the symposium and had a prelimnary discussion to make a plan for the pilot study. We have decided to compile a pilot corpus by the fall of 2008 and conduct a preliminary study to examine if any modification would be necessary for task designs and data collection procedures. In 2009, we will continue to gather the data and have the second symposium to further discuss the common framework of research using the ICCI data. By 2011, we hope to make the data publicly available and organize another international symposium focusing on research using this ICCI data. We hope that this scientific endeavour will bear fruit and make English language teaching and learning even more effective on the solid empirical grounds.

References

Abe, M. and Y. Tono. 2005. "Variations in L2 spoken and written English: investigating patterns of grammatical errors across proficiency levels". *Proceedings of Corpus Lingustics 2005*. Birmingham University, 14-17 July 2005. Available online at www.corpus.bham.ac.uk/PCLC (ISSN 1747-9398).

Bod, R. 1992. "Data-oriented parsing." *Proceedings COLING'92*. Nantes. 855-859.

Bod, R. 1998. *Beyond Grammar: An Experience-Based Theory of Language.* Stanford: CSLI Publications.

Corder, S.P. 1967. "The significance of learners' errors". *International Review of Applied Linguistics* 5:4. 161-170.

Dulay, H.C. and M.K. Burt. 1972. "Goofing: an indicator of children's second language learning strategies". *Language Learning* 22:2. 235-252.

Dulay, H.C. and M.K. Burt. 1974. "Natural sequences in child second language acquisition". *Language Learning* 24:1. 37-53.

Goldberg, A. 2006. *Constructions at Work: the Nature of Generalization in Language*. Oxford: Oxford University Press.

Izumi, E., K. Uchimoto and H. Isahara. 2004. *Nihon-jin 1200-nin no Eigo Speaking Corpus. [A Spoken Corpus of 1200 Japanese Learners of English*]. Tokyo: ALC Press.

Kaneko, E. 2006. "Corpus-based research on the development of nominal modifiers in L2". Paper presented at the American Association of Applied Corpus Linguistics, 21 October 2006.

Kobayashi, Y. 2006. "Determining L2 learners' proficiency levels using POS tag information in learner corpora: the case of NICT-JLE Corpus". (Original in Japanese.) Paper given at the JAECS Eastern-Japan region chapter meeting. September 17, 2006. Daito-Bunka University.

Labov, W. 1969. "The study of language in its social context". *Studium Generale* 23. 30-87.

Larsen-Freeman, D. 1975. "The acquisition of grammatical morphemes by adult ESL students". *TESOL Quarterly* 9. 409-430.

Shirahata, T. 1988. "The learning order of English grammatical morphemes by Japanese high school students". *The JACET Bulletin* 19. 83-102.

Tarone, E. 1979. "Interlanguage as chameleon". *Language Learning* 29:2. 181-191.

Tarone, E. 1983. "On the variability of Interlanguage Systems". *Applied Linguistics* 4:2. 142-164.

Tomasello, M. 2003. *Constructing a Language*. Harvard: Harvard University Press.

Tono, Y. 2000. "A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes". *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the third international conference on Teaching and Language Corpora*, Burnard, L. and T. McEnery (eds). Frankfurt am Main: Peter Lang. 123-132.

Tono, Y. 2006. *Corpus Cho-Nyumon*. [*Corpus Linguistics for Beginners*] Tokyo: Shogakukan.

Tono, Y (ed). 2007. *Nihonjin 1-man nin no Eigo Corpus: JEFLL Corpus. [A Corpus of 10,000 Japanese Learners of English: the JEFLL Corpus*]. Tokyo: Shogakukan.

Tono, Y. 2008. "NICT JLE vs. JEFLL: n-gram wo mochiita goi/hinshi shiyou no hattatu". [NICT JLE vs. JEFLL: n-gram analyses of lexical & POS sequences across proficiency]. *English Corpus Studies* 15. 119-133.