Corpus-Based Research and Its Implications for Second Language Acquisition and English Language Teaching

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

Abstract

The use of corpora has been increasingly common in various fields of applied linguistics. This new trend has also made a significant impact on the nature of Second Language Acquisition (SLA) research and English language teaching. In this paper, I will discuss three major issues. First, I will argue that the new methods of exploiting corpora of texts produced by L2 learners (called *learner corpora*) shed invaluable light on how L2 learners will acquire and use language in a specific educational context. It reveals complex patterns of use, focusing on the frequencies and distributions of different error patterns as well as underuse/overuse phenomena across proficiency. I will show some results of my research using the two learner corpora for Japanese-speaking learners of English, the JEFLL Corpus and the NICT JLE Corpus. Second, the pedagogical implications of corpus-based research will be discussed. In Japan, corpus-based language teaching materials and resources have become increasingly popular. I will show how corpus resources help developing teaching materials, including my NHK TV English conversation program called "*100 Go de Start Eikaiwa* (Let's start English with 100 Keywords)!", which is the first corpus-based TV English conversation program ever made. Finally, I will argue that corpus linguistics is a methodology, and thus it is important for teachers and researchers to know how to "use the tool." I hope to suggest several ways of bridging the gap between corpus linguists and English teachers in order to improve English language teaching by making it more "data-oriented."

Key Words: corpus, learner corpus, SLA, corpus-based approach, corpus applications

Introduction

In order to tackle the problem of explaining how people acquire a language and how differently they do so for the second or subsequent languages, we have been inventing a series of theories and accompanying language observation and elicitation devices. Traditionally, the overall patterns seem to fall into three major paradigms: (a) rationalist view, (b) empiricist view, and (c) socio-culturalist view. Rationalists see a language on the assumption that human beings are endowed with language-specific faculty, which makes it possible for a child to acquire a language with a remarkable consistency and high proficiency in spite of the so-called "poverty of stimulus." Chomsky and his followers, as well as UG-based SLA researchers are in this camp (cf. White, 2003). Empiricists, on the other hand, see things quite differently. They believe that language acquisition is shaped by environment and experience, and the human cognitive faculty provides a general cognitive framework, which can handle both language learning and other types of learning. They do not treat language acquisition as a special case. Most structuralists before Chomsky (cf. Lado, 1957), recently flourishing cognitive linguists (cf. Langacker, 1987) and so-called "emergentists" (cf. MacWhinney, 1999) belong to this group, although the latter two heavily shift their attention to the internal cognitive nature of a human being. The third camp sees a language and language acquisition as a cultural phenomenon. They see that language learning takes place in a particular social and cultural context, without which it is difficult to explain the transition that learners experience as they encounter the second language and gradually shift from monolingual to bilingual capacities. John Schumann's acculturation model (Schumann, 1986) and Vygotskian perspectives (cf. Lantolf & Appel, 1994) are some good examples of this view.

Methodologically, these different camps used to use different techniques for data elicitation. Rationalists are interested in the innate capacity of a person learning a language, and thus they focus on how to describe the speaker/hearer's knowledge of a language, which makes it possible for a person to produce only well-formed sentences and not ungrammatical ones. Their primary tool, therefore, is a metalingual judgment, asking for a judgment on a set of grammatical vs. ungrammatical sentences, using a grammaticality judgment test. Empiricists are more interested in the learner's environment and experience, which leads them to utilize a technique of collecting samples of the actual input and output of the learner in a more direct way, such as observing the classrooms, administering elicitation tasks, collecting protocols of student vs. teacher interaction, among others. Socio-culturalists tend to describe the situation of language learning in its entirety. Thus they prefer to gather data on the individual's perceptions about the language and society, including the classroom settings, peer relationship, learner's introspections regarding the goals and values of learning a foreign language.

In the last decade, however, this theoretical and methodological paradigm seems to have undergone a major restructuring, mainly because of technological breakthroughs that allow researchers to handle language resources and linguistic data in a way previously unimaginable. The advent of computer and the Internet technologies will now make it possible to access a huge amount of texts on the web. Chomsky once criticized the use of corpus data for its small size and skewedness (Chomsky, 1962), but more and more linguists have begun to understand that with this vast amount of data, meaningful patterns of use will emerge from the language produced in actual contexts and that it would be significant to integrate such observations into a theory of language (McEnery, Xiao, & Tono, 2006). Theoretically, more and more linguists and psycholinguists are beginning to see a language as more probabilistic in nature (cf. Bybee & Hopper, 2001). There is a growing awareness among people working in Natural Language Processing (NLP) that rule-based NLP does not seem to work so nicely as stochastic (i.e. probabilistic) NLP in many areas such as machine translation, information retrieval, pattern recognition, etc (cf. Dale, Herman & Somers, 2000). Since Chomsky proposed the Minimalist Program, people consider that a main job of language learning is to master a lexicon of an individual language, which leads to the realization that in formulating the model of the mental lexicon, probabilistic information is considered as crucial in order to explain the process of accessing the lexicon in auditory or visual input.

Learner Corpus Research

Nowadays, 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 (Granger, 1998). As is mentioned in the previous section, there is a theoretical interest in the sense that a vast amount of learner language data can produce various probabilistic information about their use of lexis and structures, sometimes overt (the combinations of words) and sometimes covert (the combination of parts of speech, for example). This probabilistic information is useful for constructing a computational model of SLA as well as confirming existing SLA findings with attested language use data (Tono, 2009). I will provide the details of the following three major areas of research: (a) compilation of learner corpora, (b) analysis of learner corpora and SLA theory construction based on the analysis, and (c) applications of learner corpus data for pedagogies and practice.

Compilation of Learner Corpora

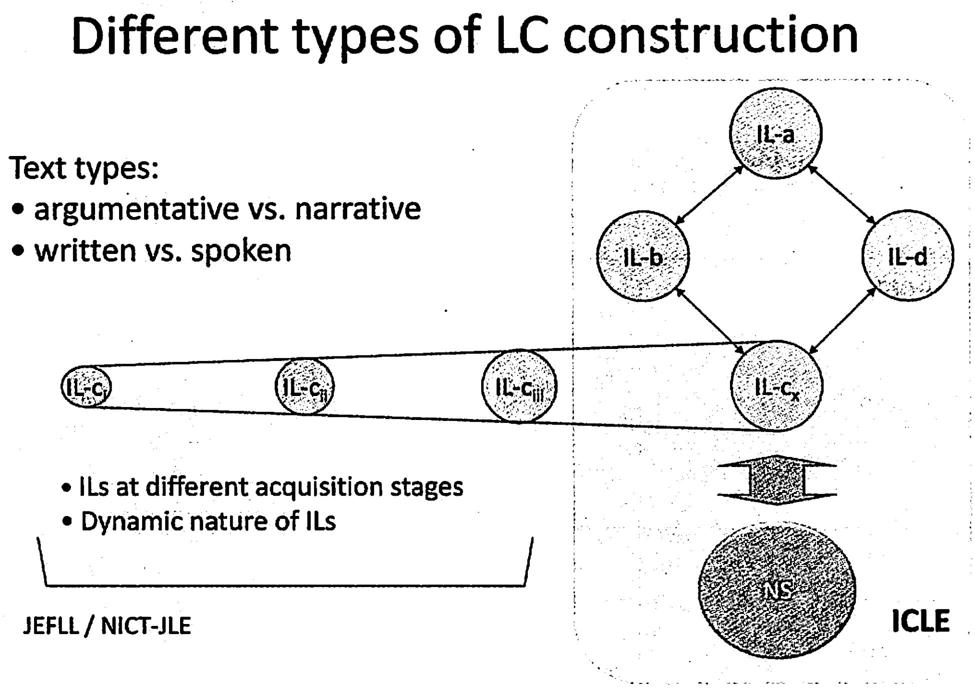
In learner corpus research, like other corpus linguistic studies, compiling a learner corpus is a very important project in itself. I have been involved in two major corpus building projects for Japanese-speaking learners of English. One is called the NICT JLE Corpus, which is a 2-million word spoken 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 school year 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://scn02.corpora.jp/~jefll04dev/), where you can access the corpus via the web-based query tool.

Analysis of Learner Corpora

We have been working on the analysis of learner language using the two corpora mentioned above. Some of the major objectives are (a) the description of Interlanguage development in terms of overuse vs. underuse as well as correct use vs. misuse of certain linguistic features, (b) the identification of criterial features which distinguish one proficiency level from another, and (c) 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 backgrounds, 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 (Granger, 1998).



Different types of LC construction

Text types:

- argumentative vs. narrative
- written vs. spoken

- ILs at different acquisition stages
- Dynamic nature of ILs

JEFLL / NICT-JLE

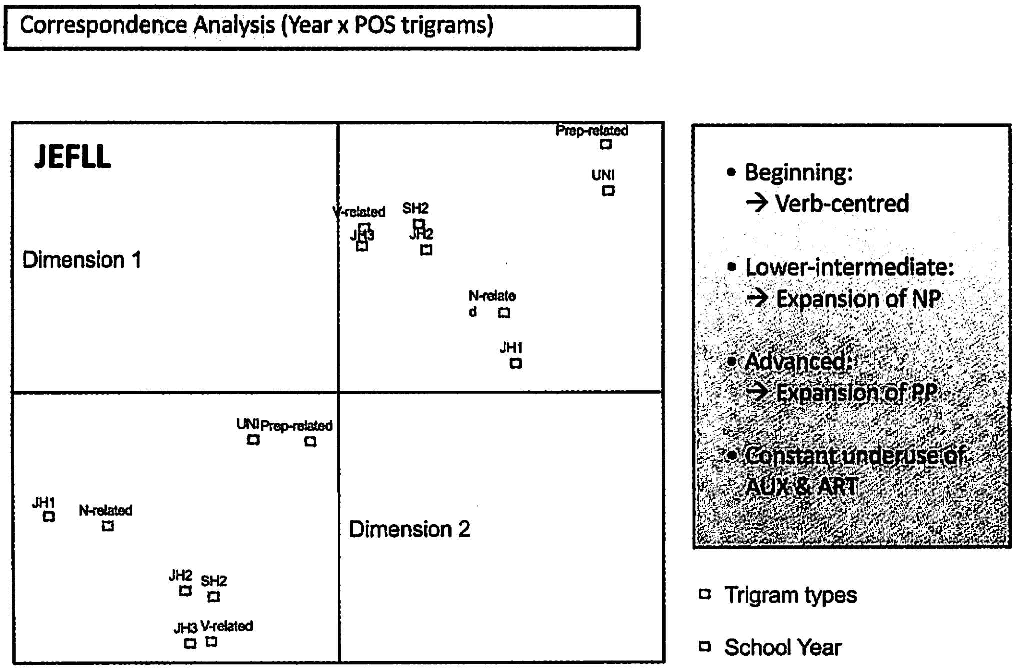
ICLE

Figure 1. Different types of learner corpus construction.

Our research group is 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 (Tono, 2007).

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 (e.g. sequences of part-of-speech tags or parsed units) in order to identify criterial features 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.

Identification of criterial features. 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.



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

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

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.

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*).

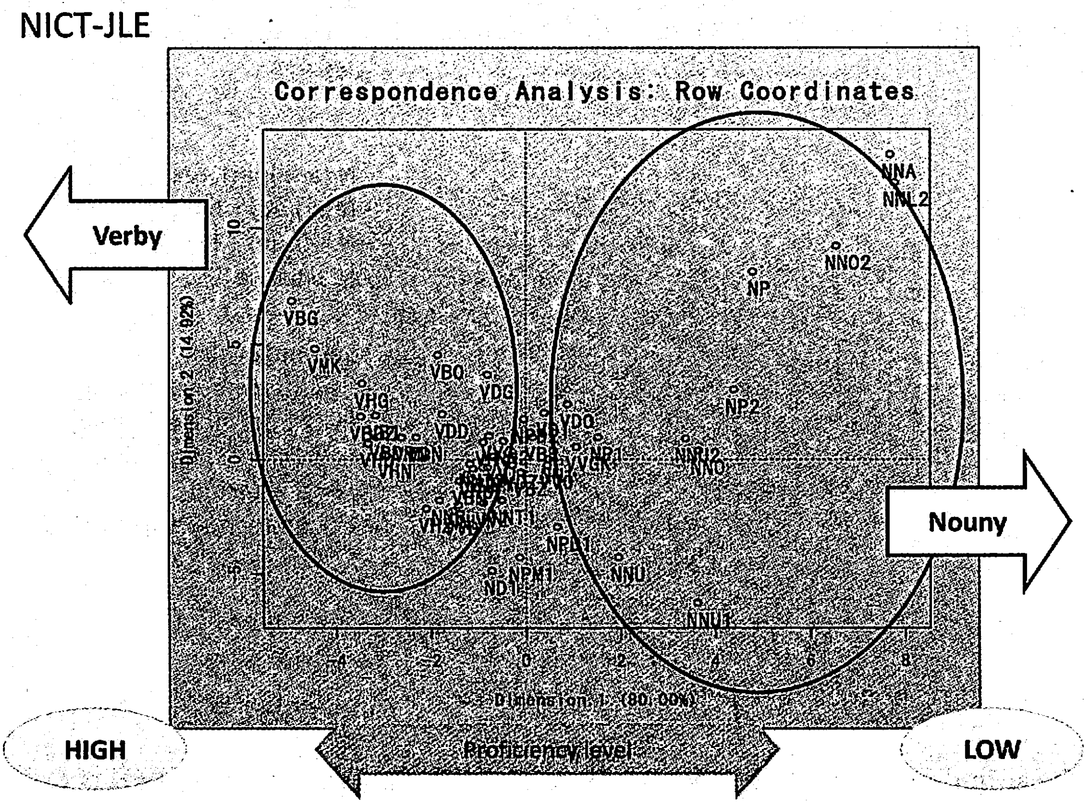


Figure 3. Correspondence Analysis over single POS tags in NICT JLE (Kobayashi, 2006).

It is noteworthy that this is a single tag distribution and not the trigram patterns shown in Tono (2000a). 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 (2000a) 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).

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.

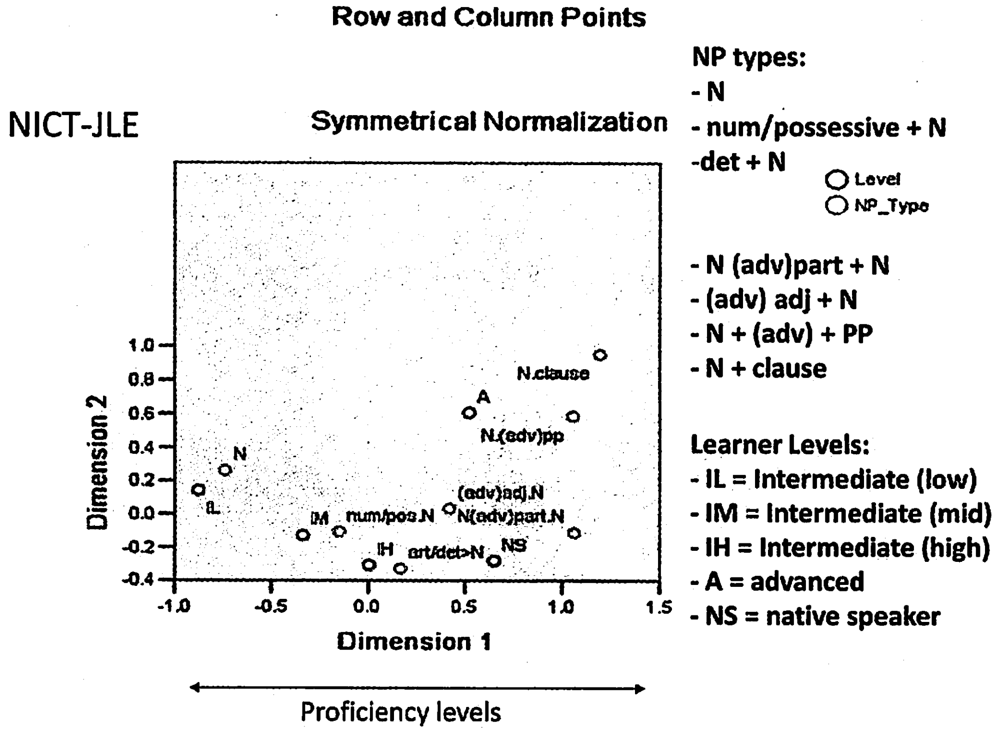


Figure 4. Correspondence Analysis over NP structures in NICT JLE (Kaneko, 2006).

These findings seem to be rather obvious to some people, but the point is that this kind of detailed descriptions of learner language for different lexico-grammatical 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.

Error frequencies across proficiency. Let me move on to the second area of studies, that is, the analysis of error frequencies across proficiency. Tono (2000b) 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 a 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 acquisition 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, a possessive maker *–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).

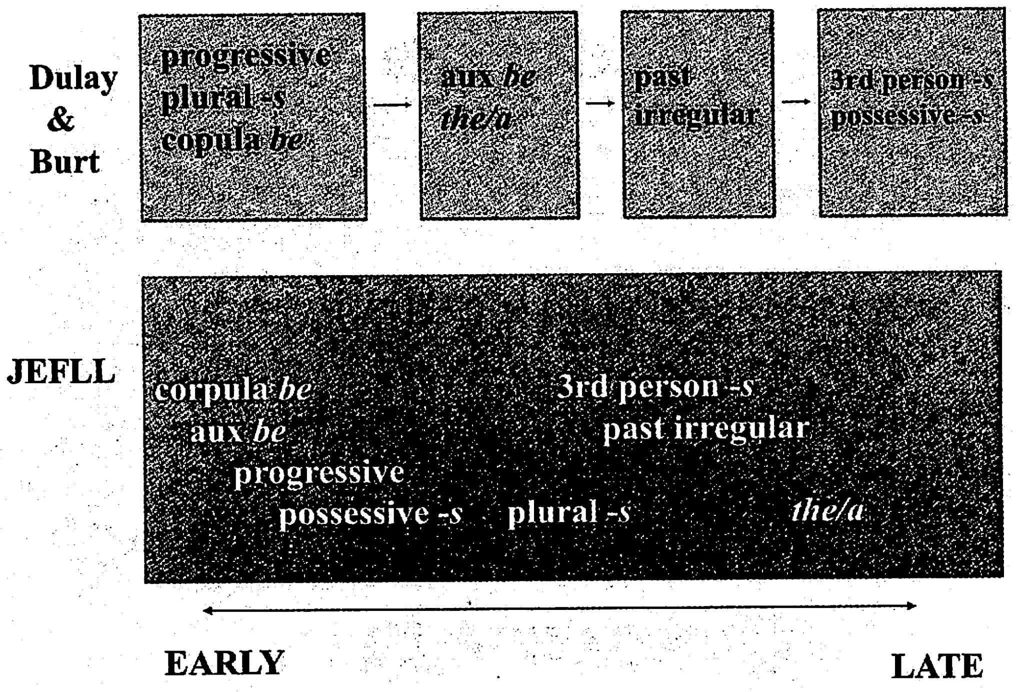


Figure 5. Morpheme order study replicated using learner corpora (Tono, 2000b).

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 across 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.

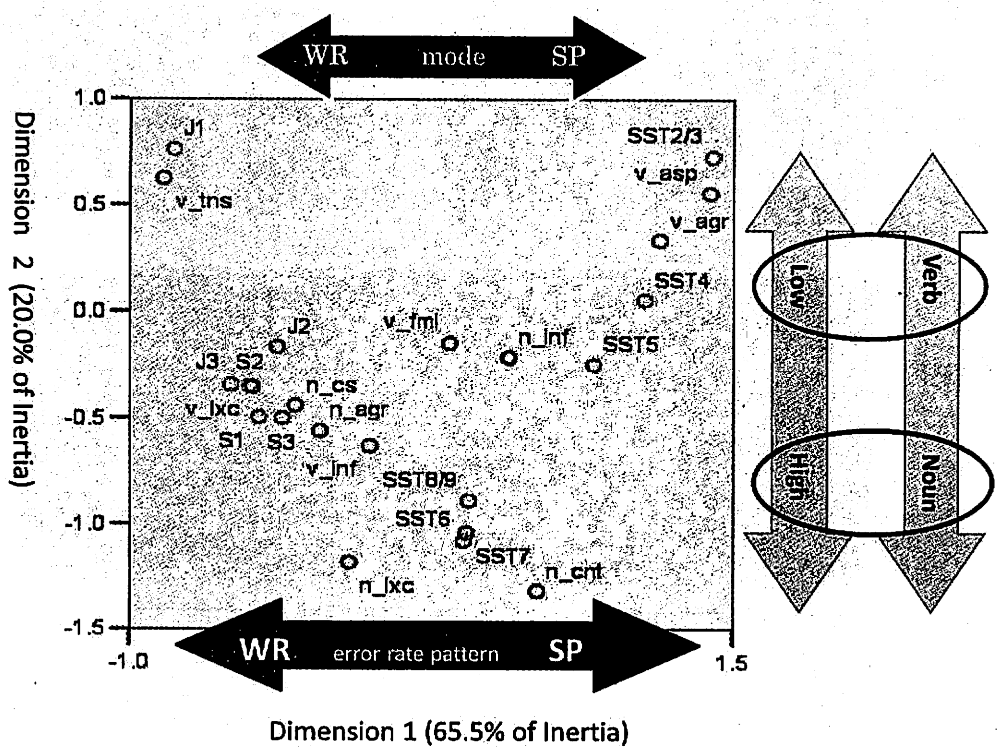


Figure 6. Error distributions across proficiency in JEFLL and NICT JLE (Abe & Tono, 2005).

Automatic Analysis of Learner Language

In the last few years, not only corpus linguists but also a group of researchers working on CALL have shown interest in the analysis of learner language on computer. In March, 2008, they organized a special pre-conference workshop for CALICO, titled "automatic analysis of learner language" (AALL'08; https://www.calico.org/p-364-CALICO%2008%20 Workshop.html).

There are several areas where automatization of learner language analysis would be possible:

1. automatic identification of learner errors
2. automatic correction of learner errors
3. automatic identification of learner proficiency levels
4. automatic scoring/evaluation of learner performance

As one can see, there is a close link between the four areas above and the e-learning system, where L2 learners' performance should be judged against certain criteria and receive proper feedback and evaluation. Good CALL systems should have this kind of function. At the moment, the accuracy and robustness of automatically identifying and correcting errors depends on the linguistic items. Agreement errors (e.g. third person singular *-s*, subject-verb agreements in singular/plural forms, plural noun marker *-s*, etc.) are relatively easy. Other grammatical morphemes such as genitive markers (*-'s*) depends on the complexity of noun phrases. Tense/aspect markers (*-ed*; *-ing*) are more difficult to judge because they are often context-dependent. Errors in verb argument structures are almost hopeless unless the corpus data is syntactically parsed properly in advance. Many researchers work on a particular area of learner errors, such as the article system, to improve the heuristics. Since we are more interested in capturing the overall picture of L2 acquisition and identifying criterial features that distinguish one proficiency level from another, we decided to explore the possibility of automatically extracting all the differences between the original learner data and its corrected version. For this, we asked one experienced native speaker instructor to correct every composition in the JEFLL Corpus. He was carefully instructed to correct errors as minimally and locally as possible so that we can compare the original and corrected versions of the sentences in more detail. In this way, we have prepared the parallel sets of original vs. corrected versions of the JEFLL Corpus.

My colleague, Hajime Mochizuki, wrote a perl script to compare the original texts against the corrected counterpart line by line to identify what and how pairs of sentences differ with each other. The algorithm used here is similar to the one called "Edit-Distance", a metric for measuring the amount of difference between two sequences (Levenshtein, 1966). The patterns of deviation will be classified into the following three surface modification error categories: (a) addition, (b) omission, and (c) misformation (James, 1998). It basically looks at the first n-word sequences in the original compositions as long as they match with the strings in the corrected version and if they do not match, it will pick up any words that are inserted in between the correct sequence of words as addition errors, and any words that are supposed to be there but missing as omission errors. It will also detect any difference that is neither an omission or addition error but differs in terms of forms from the corrected sentence in the supposedly same position. The following diagram shows a sample output for the sentence "mother is cook very well," which is corrected as "My mother cooks very well":

```
T: My mother cooks very well                                   ← corrected sentence
O: mother is cook very well                                      ← original sentence
A: <oms>My</oms> mother <add>is</add> cook[*]:msf very well              ← output
Correspondence ratio: Word level: 3/5     Character level: 3.80/5(76%)
```

Notes: T = target; O = original; A = analysis; <oms> = omission; <add> = addition; msf = misformation

It also provides the correspondence ratio statistics. The sample sentence above, for example, shows the correspondence ratio of 3 out of 5 words (60%). It also gives the ratio at character level because the original script was developed in order to detect dictation errors.

For the moment, this program works very well for simple sentences, but it has been found that the script cannot handle some of the cases properly. For example, if a pair of sentences contains a particular word in different positions (e.g. the word "breakfast" was transposed in a different position), it mistakenly counted "breakfast" as two independent errors, which has to be classified as a transposition (or word order) error:

T: I eat bread and milk for breakfast
O: I eat the breakfast bread and milk
A: I eat <add>the</add> <add>breakfast</add> bread and milk <oms>for</oms> <oms>breakfast</oms>

Another problem is a very low accuracy in detection of misformation errors. For instance, the following pair of sentences was analyzed correctly:

T: I like bread and butter too
O: I like bread and better too
A: I like bread and b[e]tter:msf too

However, the following case is problematical, where the output is supposed to identify the phrase "*and ocha* (=*tea*)" should be replaced by "*or tea*" in the target sentence, but it fails. If more than one word is different, the analysis based on n-grams suddenly become unstable due to the lack of shared words to fall back on and become more susceptible to errors:

T: But I don't like rice or tea
O: But I don't like rice and ocha
A: But I don't like rice <oms>or</oms> [**]a[nd]:msf <add>ocha</add>

Despite this kind of shortcomings, we still find the program very useful in automatically processing all the learner data at the same time and see the general patterns. Especially omission and addition errors are found to be fairly accurately detected, compared to misformation errors. Figure 7 shows the overall distribution patterns of omission and addition errors:

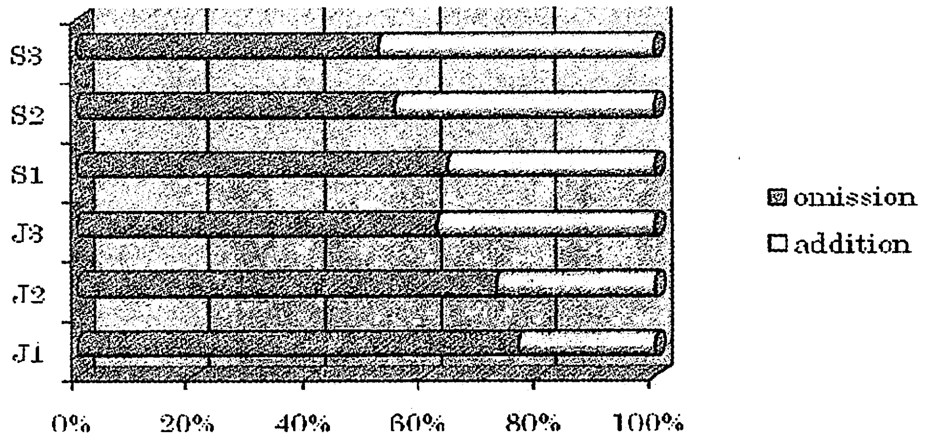


Figure 7. Distribution of omission vs. addition errors.
(Note: J = junior high school; S = senior high school)

As shown in Figure 7, novice-beginner level learners (J1 and J2) have approximately 70% of omission errors while addition errors are less than 30%. This ratio will be about 60% to 40% for the lower-intermediate level learners (J3-S1) and gradually approach toward 50-50 for more advanced levels (S1 and S2). There is a strong tendency that Japanese-speaking learners of English tend to avoid using items at earlier stages, thus producing more omission errors and as their level goes up and start using language, they produce more addition errors.

Article errors. Article errors are very common among Japanese EFL learners. Figure 8 shows the omission and addition errors for articles (*the* and *a*). We have observed in Figure 7 that the proportion of omission errors will decrease as more addition errors occur as the school level goes up. In the case of article errors, however, omission errors occur highly frequently and consistently throughout the different school levels, which indicates that article errors show quite a different pattern from the general error type ration found in Figure 7.

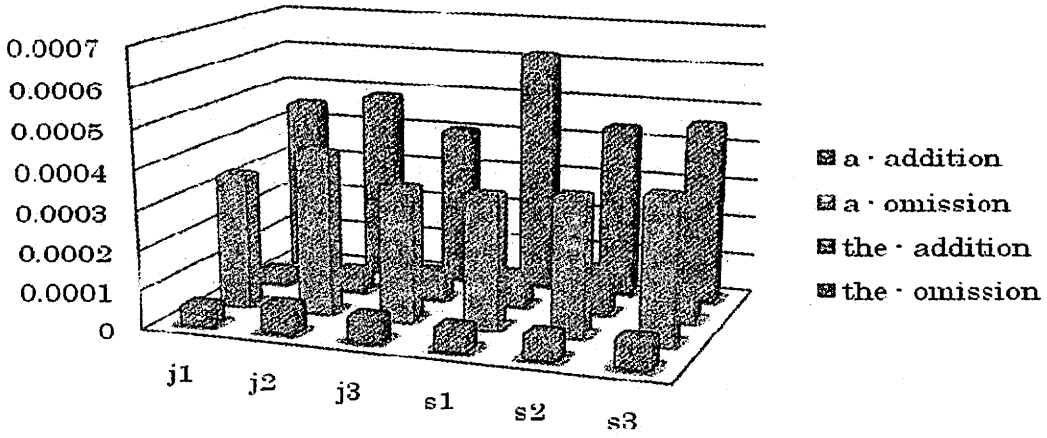


Figure 8. Article errors.
(Note: The addition errors for the indefinite article "*a*" come in the front row of the figure.)

Preposition errors. Figure 9 compares addition and omission errors for two prepositions, *of* and *to*.

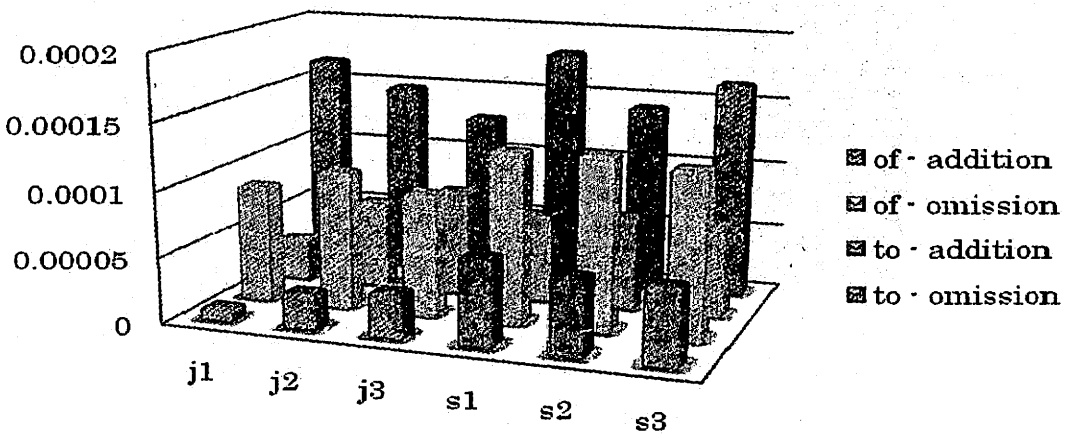


Figure 9. Preposition errors (*of* and *to*).

Overall, learners produce more errors with a preposition *to* than *of*. *To* also works for infinitive markers and this figure does not distinguish the use of infinitive markers from normal prepositional use, thus possibly boosting the figure even more. The preposition *of*, on the other hand, is a very popular marker for typical written texts to show elaboration. By using the preposition *of*, we explicitly describe nouns with further modifications, thus producing more complex noun phrases. Nominalization of verbs is typically realized by the preposition *of* (e.g. *production of* Interlanguage). It is interesting to see in Figure 9 that errors related to the preposition *of* are increasing throughout the school years as they move to higher grades. This is a good sign of L2 learners attempting to make a longer, thus more complex noun phrases.

Modal verbs. Modal auxiliaries (*can, could, will, would, may, might, should, must*) are also constantly underused items for Japanese-speaking learners of English. Figure 10 shows the summary of the errors for all the major modal verbs in terms of addition and omission errors. As is illustrated in Figure 10, the ratio of omission errors is generally much higher than that of addition errors. Omission errors also show some interesting pictures. At the very beginning stage (J1), omission errors are very high, which indicates that modal verbs are not properly introduced at this stage. In J1, most grammar items are introduced in present form. In J2, however, the errors decreased dramatically, partly due to the fact that most of the modal usage introduced at this stage is a simple formula such as "*I can* ..." or "*Can you* ...?" and the types of modals are still very limited in number and usage. At later stages, as they learn more modal verbs, addition errors gradually increase, although the total number of errors is still much lower than that of omission errors.

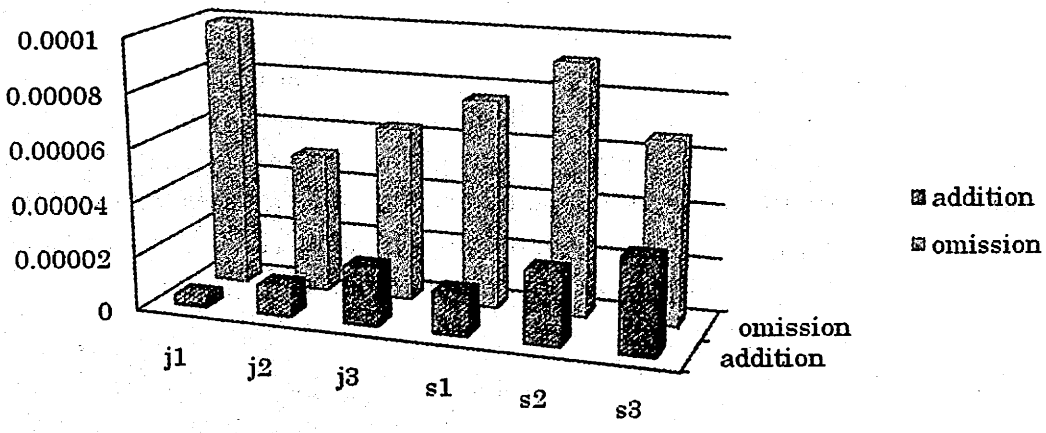


Figure 10. Modal verb errors.

Automatic analysis of learner errors is still in its infancy, but the present study on the JEFLL Corpus shows that it has a great potential. Our goal is to extract the frequencies of omission/addition/misformation errors for all the relevant linguistic features across proficiency and perform multivariate analysis over those features in order to determine what features will work best as criterial features characterizing particular proficiency levels of L2 learners. This kind of statistics then can be compared to the frequencies of linguistic features in ELT textbooks used at school to see the relationship between textbook input and learner production. Another possibility is to compare such data with L1 mother tongue corpora in order to see the effects of L1 influence on the error-prone linguistic items. Tono (2004) did a similar type of multifactorial studies focusing on the use/misuse of verb subcategorisation, but this time we have a possibility of conducting same types of research in a much larger and more extensive scale.

Pedagogical Application of Corpora

Corpus-based research also has practical value. Finding from corpora how native speakers use language will become a very useful resource for developing teaching materials. It is a well-known fact that most pedagogical monolingual dictionaries of English such as the *COBUILD English Dictionary*, the *Longman Dictionary of Contemporary English*, the *Macmillan English Dictionary* among others, are now fully corpus-based.

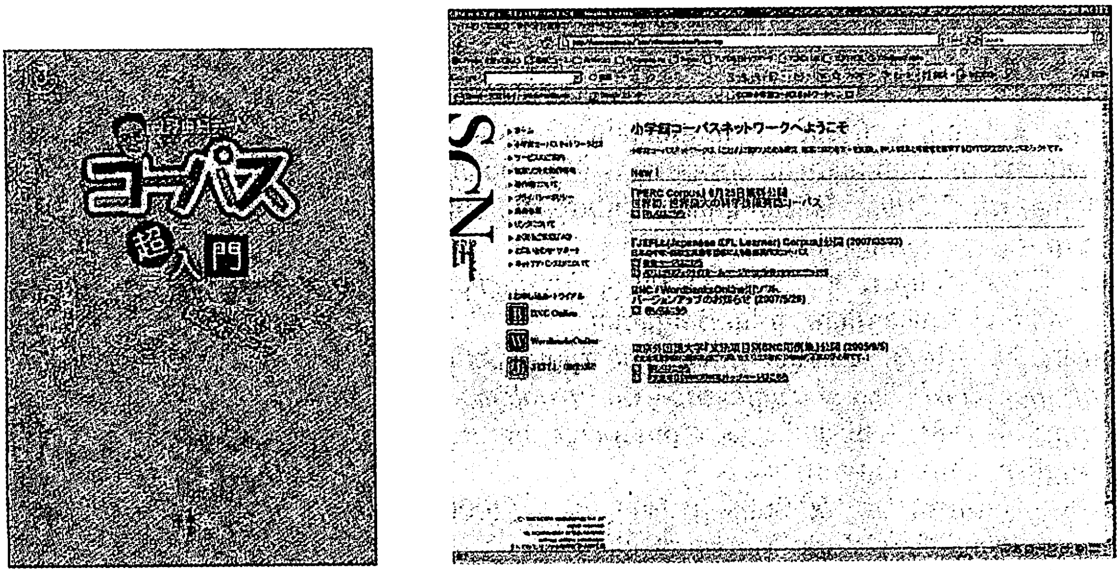
I was one of the first among corpus linguists, who had brought the notion of "corpus" to English classrooms and English teaching/learning communities in Japan. One such effort was to work for NHK (Nihon Hosou Kyokai, i.e. the Japan Broadcasting Center) to produce a corpus-based TV English conversation program. It consists of one hundred units, each of which lasts for 10 minutes, featuring one English basic keyword (mainly basic verbs, prepositions, auxiliaries, conjunctions, and *wh*-pronouns) and show how the keyword is used in context. All the keywords and their collocates for practice were selected from the spoken component of the British National Corpus. 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 11) 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 11. Corpus-based English conversation textbooks and DVD series written by the author.

Bridging the Gap

Another pedagogical practice that is worth noting in Japan is to bridge the gap between corpus linguists and ordinary classroom teachers. Until several years ago, the term “corpus” was not very familiar to English teachers. After my TV program, however, English teachers in Japan became aware of the importance of corpora as language resources and they started to look for opportunities to learn more about corpora and how to use them for their study and classroom teaching. I found it very useful to discuss the use of corpora with language teachers in a seminar or workshop, and thus decided to provide a better environment for language teachers to access corpus data.

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 WordBanks Online (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 using electronic dictionaries or Google. This site is especially welcomed by English teachers who wish to access corpus data for their preparation for teaching. Figure 12 shows the website of the Shogakukan Corpus Network and my recently published book introducing 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 (2007/03/23)

BNC / WordbanksOnline
バージョンアップのお知らせ (2007/5/28)

東京外国語大学「文法項目別BNC用例集」公開 (2005/9/5)

Figure 12. An easy guide to Corpus and the Shogakukan Corpus Network.

Conclusion

While it takes time and efforts to educate language teachers about the value of using corpora in the classroom, this is a very important step ahead because life will be changed if they know how to use corpora. It is time for non-native teachers of English with limited intuition about the target language to feel confident. They will be able to confirm various aspects of English usage by accessing corpora directly for themselves. Many of them will be free from the negative impressions that they always rely on native speakers of English. The drastic change in perspectives in understanding the target language by observing real language use data will be so tremendous that it will change the whole notion of materials selection, ordering, presentation and practice by teachers themselves, with a proper emphasis on more frequent and fundamental items. Learners will then have an opportunity to access corpus data for themselves in the future, which will open up a new possibility of Data Driven Learning (DDL) and we can move ahead to empirically test such an approach in the real classroom settings. In the meantime, corpus linguists like myself need to bridge the gap by providing the results of the corpus analysis of how native speakers use the target language as well as how L2 learners perform in speech and writing. All sorts of new paradigm will emerge from this exciting empirical knowledge base called corpora. I sincerely hope that further research into the nature of native speaker vs. learner corpora, together with various other sources, will greatly improve pedagogy and practice of English language teaching around the world.

References

Abe, M., & Tono, Y. (2005). Variations in L2 spoken and written English: Investigating patterns of grammatical errors across proficiency levels. *Proceedings of Corpus Linguistics 2005*. Birmingham University, July 14–17, 2005. Retrieved from http://www.corpus.bham.ac.uk/PCLC

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Chomsky, N. (1962). A Transformational Approach to Syntax. In Hill (Ed.), *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English* (pp.124–158). Third Texas Conference on Problems of Linguistic Analysis in English, University of Texas.

Dale, R., Herman, M., & Somers, H. (Eds.). (2000). *Handbook of natural language processing*. New York: Marcel Dekker.

Dulay, H. C., & Burt, M. K. (1972). Goofing: An indicator of children's second language learning strategies. *Language Learning, 22*(2), 235–252.

Dulay, H. C., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning, 24*(1), 37–53.

Granger, S. (Ed.). (1998). *Learner English on computer*. London: Longman.

Izumi, E., Uchimoto, K., & Isahara, H. (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, October 21, 2006.

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

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites*. Stanford: Stanford University Press.

Lantolf, J. P., & Appel, G. (Eds.). (1994). *Vygotskian approaches to second language research*. Norwood, NJ: Ablex Publishing Company.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

MacWhinney, B. (1999). *The emergence of language*. Mahwah, NJ: Erlbaum.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Schumann, J. H. (1986). Research on the acculturation model for second language acquisition. *Journal of Multilingual and Multicultural Development, 7*(5), 379–392.

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

Tono, Y. (2000a). A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. In B. Lewandowska-

Tomaszczyk & J. P. Melia (Eds.), *PALC '99: Practical applications in language corpora* (pp. 323–340). Frankfurt am Main: Peter Lang.

Tono, Y. (2000b). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora* (pp.123–132). Frankfurt am Main: Peter Lang.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorisation patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 45–66). Amsterdam: John Benjamins.

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.

Tono, Y. (2009). Integrating learner corpus analysis into a probabilistic model of second language acquisition. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 185–203). London: Continuum.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.