Corpus-Based SLA Research:
State of the Art of Learner Corpus Studies

YUKIO TONO
Meikai University

Studies in Language Sciences (4)
Papers from the Fourth Annual Conference of
the Japanese Society for Language Sciences

Edited by
MASAHIKO MINAMI
HARUMI KOBAYASHI
MINEHARU NAKAYAMA
AND
HIDETOSI SIRAI

Kurosio Publishers, Tokyo.
2005

Corpus-Based SLA Research: State of the Art of Learner Corpus Studies

YUKIO TONO, Meikai University

Abstract

This paper discusses the potential of learner corpora in SLA research. I will first present methodological issues of interlanguage (IL) studies from a historical perspective to show how learner production data was used in different research paradigms. Second, I will briefly summarize the features of present-day learner corpora and the current projects. Third, I will present the results of my study as an example of exploiting learner corpora, focusing especially on the effectiveness of a multiple comparison of interlanguage (IL), first language (L1) and target language (TL) corpora.

1. Introduction

The word "corpus" has both positive and negative connotations. It often reminds us of the remark made by Chomsky (1962) to the extent that corpora, by their very nature, are incomplete and "skewed" (p. 159). Chomsky's comments about corpora made many linguists, especially those who work in the field of theoretical linguistics, stay away from using corpus data as a primary source of evidence and rely heavily on introspection instead. Whilst his comments show some significant facts about corpora which we should take seriously, we should also note that these criticisms did not stop all corpus-based work. In the field of phonetics, for example, naturally observed data remained the dominant source of evidence. In the field of language acquisition also, the observation of naturally occurring evidence remained dominant. Chomsky (1964) himself cautioned that his rejection of performance data as a source of evidence was inappropriate for language acquisition studies.

Recently corpora have been in the spotlight in the field of applied as well as theoretical linguistics. In applied linguistics, there is a growing interest in the use of corpora for the study of language use. The primary reason for this move is the dramatically improved availability and accessibility of corpora. Such mega-corpora as British National Corpus (BNC)[1] or Bank of English[2] have improved the description of English by providing statistics about language use which could not otherwise have been obtained. Major pedagogical dictionaries are now based on large corpora.

---

[1] http://www.hcu.ox.ac.uk/BNC/

[2] http://titania.cobuild.collins.co.uk/

Longman Grammar of Spoken and Written English by Biber, et al. (1999) shows that a corpus has some intriguing possibilities of describing an English grammar from the viewpoint of how a language is really used. The detailed bibliographical and demographic descriptions of the corpus data in the BNC have become extremely useful research tools for those who work in sociolinguistics and language variation studies. The spoken parts of the BNC or International Corpus of English – Great Britain (ICE-GB)[3] provide researchers analyzing a spoken language with the opportunity to compare spoken data with written data. The on-going corpus compilation projects such as the American National Corpus[4] will facilitate the collection of major varieties of English with up to 100 million words.

The use of corpora has been regaining ground in the field of theoretical linguistics as well. The most notable area is a shift towards a greater preoccupation with the lexicon. Many aspects of language that earlier Chomskyan models dealt with as 'syntax' are now handled as idiosyncrasies of lexical items. The syntax itself is considerably simplified by the omission of many rules, at the cost of greatly increased lexical information. This leads to a revealing insight into the usefulness of corpus data for the description of the lexicon. There is another perspective called a 'usage-based' approach (cf. Barlow & Kemmer, 2000). Psycholinguistic and cognitive linguistic theories of language acquisition hold that all linguistic units are abstracted from language use. In these usage-based perspectives, the acquisition of grammar is the piecemeal learning of many thousands of constructions and the frequency-biased abstraction of regularities within them. Language learning is the associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is thus a key determinant of acquisition. Frequency underpins regularity effects in the acquisition of orthographic, phonological and morphological form, and learning accords to the power law of practice (Ellis, 2002). In order to construct such a theory, it is essential that one can obtain frequency information of given linguistic features, and the principal source of such data comes from properly sampled corpora.

Despite the popularity of corpus data in the field of language sciences in the last decade, there seems to be little discussion about how corpora should be best compiled or fully exploited, one of the main issues in corpus linguistics. In this paper, I will first present some fundamental concepts about a corpus, and then review interlanguage (IL) studies from a historical viewpoint in order to show the need of good learner corpora for research in SLA and TEFL. Secondly, I will summarize on-going projects on learner

[3] http://www.ucl.ac.uk/english-usage/ice-gb/
[4] http://americannationalcorpus.org/

corpora around the world. Finally, I will demonstrate how learner corpus data can be exploited by comparing an IL corpus with L1 and TL corpora.

2. Fundamental Concepts

The word "corpus" is a cover term for all sorts of collections of text, but it has been increasingly used to refer to only those which are specially assembled for the purpose of linguistic analysis and/or natural language processing. The new generation of corpus linguistics is characterized by the systematic compilation of representative samples of a particular variety of language for analysis by computer (Leech, 1991). Leech (1992) describes five key characteristics of 'the scientific method' and evaluates the extent to which the corpus-based methodology conforms to these scientific norms: (a) falsifiability, (b) completeness, (c) simplicity, (d) strength, and (e) objectivity. I will not go into the detail of evaluation of his claim here, but I argue that the first norm "falsifiability" as well as verifiability will be greatly enhanced as more researchers share corpus data with each other.

Another important definition of corpus is the one made by McEnery and Wilson (2001). Corpus linguistics is not an aspect of language requiring explanation or description such as syntax, semantics, pragmatics and so on, but a methodology. Corpus linguistics is a methodology that may be used in almost any area of linguistics, but it does not truly delimit an area of linguistics itself" (ibid, p. 2). This understanding is crucial as we see more and more people use corpora as testbeds for verifying their theoretical claims.

3. Use of Corpora for IL Studies: Historical Perspectives

3.1. Contrastive Analysis and Error Analysis

The idea of using learner data in SLA research is not new. In fact it has been around for more than 30 years. The treatment of learner data in most studies to date has been, however, rather haphazard. In the late 1950s and early 1960s, when a major concern was to determine L1 influence on L2 learning and use, the predominant method used was Contrastive Analysis (henceforth CA). CA made no use of learner language data. Instead, analysts compared the target language with the first language to identify the similarities and differences, which they believed would predict the relative difficulty of learning (Fries, 1957; Lado, 1957). Their comparisons between the L1 and the TL, however, were largely based upon expert knowledge and not corpus-based.

Following on from a focus on CA in the 1960s, Error Analysis (EA) became an important paradigm in the 1970s. Researchers began to examine L2 learners' errors, not as something unwanted, but as evidence of the development of the IL system (Corder, 1967; Richards, 1971). The difference in approach towards the language data between CA and EA is

noteworthy. While in CA only the first language (L1) and the target language (TL) were compared, resulting in total neglect of learner language data in EA, a comparison was made between different stages of the IL system only and no attention was paid to the L1 or the TL. Most EA studies concur that the majority of errors are intralingual (Ellis, 1994: 69) and thus there was no use for L1 source language data. There were many studies in which learner performance data were actually obtained but were only accessed through audiotapes and never transcribed as text (see, for example, Selinker, Swain, & Dumas, 1975; Tarone, Frauenfelder, & Selinker, 1976; Hendrickson, 1976). There is considerable evidence that the idea of analyzing learner performance data started to be accepted in the late 1960s, when people began to collect learner language samples in order to better understand IL development. However, learner performance data were not fully exploited in their own right because researchers of the day were only interested in the errors themselves. Thus, after error patterns were extracted from the data, the transcription was either discarded or did not undergo further processing to be used as a corpus. Frequency information concerning the errors was often missing or lacked precision, which gave rise to concern among SLA researchers with regard to the empirical value of such error data.

3.2. Performance Analysis

Performance Analysis (PA) emerged in the early 1970s. The difference between EA and PA is that the former attempts to reconstruct learners' acquisition processes on the basis of errors alone while the latter makes use of the whole of their performance, both correct and erroneous. PA was considered to be superior to EA because EA depended only on the analysis of learners' errors and did not take into account what learners had already acquired. This notion of studying the entire performance of language learners, both errors and correct forms, facilitated a more careful treatment of learner production data. Although they did not call it a "learner corpus" and did not pay attention to the format or nature of the data itself, researchers in the 70s surely had a collection of learner language, sometimes quite copious amounts, thus showing that a primitive form of learner corpora was available at that time. For example, Dulay and Burt (1973, 1975) claimed that they examined more than 800 learners with the instrument called the Bilingual Syntax Measure, an elicitation device using picture retelling tasks. If this is true, they had a database of 800 learners' spontaneous speech although it was not very clear whether those data were actually transcribed. Hakuta's (1976) study on child L2 acquisition collected 30 sessions of 2 hour spontaneous speech data with a 5-year-old Japanese girl, once a fortnight for over a period of 60 weeks. The data were recorded and later orthographically transcribed. He referred to the data as

"30 bulky loose-leaf notebooks filled with transcriptions." This corpus, however, was never converted into an electronic form.

3.3. Classroom Process Research

In the 1980s, interest in investigating the classroom interaction between a foreign language teacher and learners grew. This was partly due to the recognition that systematic observation was necessary in order to fully understand how instruction and learning take place. There was also a growing interest in classroom SLA. After the debate over the effect of formal instruction (cf. Long, 1983), people came to realize that formal instruction is indeed valuable for IL development and that its effect needs to be investigated more systematically. Since the L2 acquisition process was found to be closely related to the classroom input and interaction, L2 researchers felt the need for more careful research in order to identify the effect of classroom activities on the IL development. Compared with EA and PA, classroom process research focuses more on the interaction between teachers and learners in the classroom context. In early classroom process research, most of the data consisted of frequencies of events in pre-defined categories, such as how many times the teacher asked questions/ accepted feelings or how often pupils responded (Moskowitz, 1967; Fanselow, 1977; Allwright, 1980). Therefore, from the viewpoint of learner language data, very little was available for further processing as a corpus. In the later classroom process research, however, there seemed to be more and more data available, at least in audio- or video-recording format. For instance, Sinclair and Coulthard (1975) tried to take account of the findings of more theoretical analyses of classroom discourse and the observation schedule they used attempted to preserve the discourse structure of a lesson. Later schedules were designed for use with recorded or transcribed data, but, with only a few exceptions such as the written protocol printed in Sinclair and Coulthard (1975) or the sample transcriptions in Fanselow (1977), Allwright (1980), and Van Lier (1982), there were no published databases or transcriptions of the classroom observations. The reason for this lack of published transcriptions or recordings is that the researchers had no plan to make their data public and usually received consent from teachers to make observations for private research purposes only. Unfortunately, therefore, very little data seems to have been actually transcribed and exploited as a corpus.

3.4. The Prototype of Learner Corpora

The idea of ensuring that the data maximally represents the target group has been often mentioned among L2 researchers, but in reality it has always been difficult to gather attested language use data that meet this demand without sufficient time and money. Thus most studies in the 1970s and 80s

failed to achieve this goal. Two projects, the ZISA Project in Germany and the ESF Database in European nations, however, are worth mentioning because they are among the few projects that generated corpora which approximate closely to modern-day learner corpora.

3.4.1. ZISA Project

The ZISA Project has been reported in a series of papers by Meisel, Clahsen, and Pienemann (for example, Meisel, Clahsen, & Pienemann, 1981; Clahsen, 1980, 1984; Clahsen, Meisel, & Pienemann, 1983; Pienemann, 1980). They found that there was a clear development pattern in the acquisition of German word order rules by L2 German learners. There were two phases of data collection by the research team ZISA: first, a cross-sectional study with 45 adult workers from Italy, Spain and Portugal, was conducted from 1977 to 1978 via interviews conducted in the manner of unguided conversations (Clahsen, 1980: 59). The results of this study were then tested in a longitudinal study with 12 adult learners of the same origin. It was scheduled for three years (2 years of observation) (Meisel, 1980: 27). In both forms of investigation all of the interviews were audio-taped and transcribed afterwards (Clahsen, 1980: 59). Thus, there was a corpus of L2 German learners in a naturalistic acquisition context, sampled both cross-sectionally and longitudinally, which is a well-planned overall design even by current standards. However, available reports show that the use of the corpus data was rather fragmented in nature. Since their primary focus was on the development of word order rules, they omitted repetitions and fillers from transcripts. Also they did not analyze the whole corpus, using only a selective transcription consisting, for each learner, of at least 50 utterances from each session (ibid.). The corpus has not been made available to us, which makes it difficult for us to verify their findings.

3.4.2. ESF Database

The other project I would like to comment on is the European Science Foundation Second Language Database (the ESF Database). This project was initiated by Clive Perdue and her team, supported by the European Science Foundation. It is a text database collected by research groups within the ESF-project in five European countries: France, Germany, Great Britain, The Netherlands and Sweden. Immigrants of five different source languages (Arabic, Finish, Italian, Punjabi, Turkish) were observed during a period of 3 years acquiring a target language to which they were exposed (Dutch, English, French, German, Swedish). For each target language, two source languages were selected so that a cross-linguistic analysis could be made between two source languages and one target language. The project concentrated on spontaneous second language acquisition by 40 adult immigrant workers living in Western Europe, and their communication with

native speakers in the respective host countries. The database consists of transcribed recordings of those migrant workers learning the language of their resident country, which includes several types of language use gathered in three data cycles over 2.5 years. The design of the database shows that approximately four to eight subjects were selected from each L1 group. Since the corpus is longitudinal in nature, the database is an invaluable resource despite the small sample size. The ESF database has been available in the CHAT format since 1993 and consequently tools designed for the CHILDES Project can also be used on this database (for CHAT and CHILDES, see MacWhinney (1995)). See Perdue (1984, 1993) for more details.

3.5. Summary

This section examined how L2 researchers treated learner performance data in the past research paradigm. The review reveals that the idea of using learner data is not new. In fact it has been around for more than 30 years. The treatment of learner data in most studies to date has been, however, rather haphazard. Often only a fragment of the data is analyzed and the rest simply ignored. Although there had been a growing awareness that learner data should be investigated in its entirety, it was not until about a decade ago that a systematic collection of learner production data in the framework of corpus linguistics actually started.

It is also the case that in most past approaches to IL studies, very little attention was paid to the IL development in its entirety, especially in relation to learners' L1 knowledge and the status of the target language. Having said that, I would like to move on to the description of present-day learner corpora to show how different they are from their predecessors.

4. The JEFLL Corpus and Multiple Comparison Approach

Before introducing major projects of learner corpora, let me briefly describe my project. This will, I hope, be able to show how present-day learner corpora are designed and compiled. The project is called the Japanese EFL Learner (JEFLL) Corpus. It aims to compile a corpus of Japanese EFL learners from Year 7 to university levels. The strength of the JEFLL Corpus is that it contains L1 and TL corpora as an integral part of its design. As was shown in the last section, very few studies have made use of both attested L2 learner data and L1/TL data to identify features of interlanguage development, let alone a corpus-based analysis of these data. Most learner corpus studies to date have made use of NS corpora because the studies are typically focused on learning English, and many native English corpora are readily available as a standard reference, whereas very

few studies (except for PELCRA[5] and JEFLL) collect L1 source corpora for comparison. Table 1 shows the overall structure of the JEFLL Corpus. The total size of the L2 corpus is approximately 400,000 running words of written texts and 50,000 words of the orthographically transcribed spoken data. The L1 corpus consists of a corpus of Japanese newspaper texts (approximately 11 million words) plus a corpus of student compositions written in Japanese. These L2 essays were written on the same topics as the ones given to the L1 English writing.

Table 1. The JEFLL Corpus Project: The Overall Structure[6]

Part 1: L2 learner corpora
- Written corpus (composition): c. 400,000 words
- Spoken corpus (picture description): c. 50,000 words
Part 2: L1 corpora
- Japanese written corpus (composition): 50,000 word texts on the same tasks as English
- Japanese newspaper corpus: c. 11,000,000 words
Part 3: TL corpus
- EFL textbook corpus: 650,000 running words (Y7-9: 150,000; Y10-12; 500,000)

The third part of the JEFLL Corpus comprises the TL corpus. It is a corpus of EFL textbooks. It covers both junior and senior high school textbooks. As regards the junior high school textbooks, they are used officially at every junior high school in Japan. There are seven competing publishers producing such textbooks. Irrespective of which publisher one chooses, each publishes three books corresponding to the three recognized proficiency grades for years 7-9. Senior high school textbooks are more diversified and more than 50 titles have been published. This corpus contains mainly the textbooks for English I and II (general English).

I argue that textbook English is a useful target corpus to use in the study of learner language. As this claim runs counter to that of other researchers (e.g., Ljung, YEAR; Mindt, YEAR; Granger, YEAR), it is important to examine the basis of this claim in some detail. Firstly, the target language which learners are measured by should reflect the learning environment of learners. It is not always appropriate to use a general corpus such as BNC or the Bank of English to make comparisons with non-native-speaker corpora. The difference you will find between L2 corpora and those general corpora

[5] See Table 2 for details.
[6] As of July 2002.

will be the one between the learner English and the English produced by professional native-speaker writers. This comparison could be meaningful for highly advanced learners of English or professional non-native translators. The output of such highly advanced learners, however, is something which the vast majority of L2 learners in Japan never aspire to. We have to consider very seriously what the target norm should be for the learners we have in mind. In the present case, it is certainly not the language of BNC that the Japanese learners of English are aiming at, but, rather, a modified English which represents what they are more exposed to in EFL settings in Japan. I am fully aware of the fact that the languages used in ELT textbooks themselves are unnatural in comparison to the native speaker usage (see, for instance, Ljung, 1990; 1991). Pedagogically, however, beginning- or intermediate-level texts should contain such modified features of English in order to promote students' learning. Given that textbooks, with all of their peculiarities with comparison to L1 corpora, represent the primary source of input for L2 learners in Japan, their use in explaining and assessing L2 attainment is surely crucial.

The textbook is the primary source of input in Japan. Inside the classroom, some teachers will use classroom English, and others do not use English at all as a medium of instructions. Even if they do use English in the classroom, they usually limit their expressions to the structures and vocabulary that previously appeared in the textbook. Outside the classroom, those who go to cram schools will receive extra input, but still that input is comprised of questions borrowed from past entrance exams, or questions based on the contents of the textbooks (Rohlen, 1983). Hence, it is fair to say that the English used in the textbooks is the target for most learners of English in Japan. If we exclude textbooks from our investigation explaining the differences between TL and IL usage may be impossible. However, where textbooks are included in an exploitation of L2 learning, they can explain differences between NS and NNS usage (McEnery & Kifle, 2001).

While the above argument presents the basis for my inclusion of textbooks in my model of the study of learner language, more evidence is required to substantiate this claim. Later in the description of some of my studies, the textbook corpus will be called upon to provide explanation of differences between IL and TL, substantiating my claim further. For the moment I will take the argument presented so far as sufficient evidence to warrant the inclusion of textbook material in my learner corpus exploitation model. My proposal, therefore, is that standard reference (e.g. BNC), textbook and learner corpora all have roles to play in its exploration of learner language.

Figure 1 illustrates this point diagrammatically.

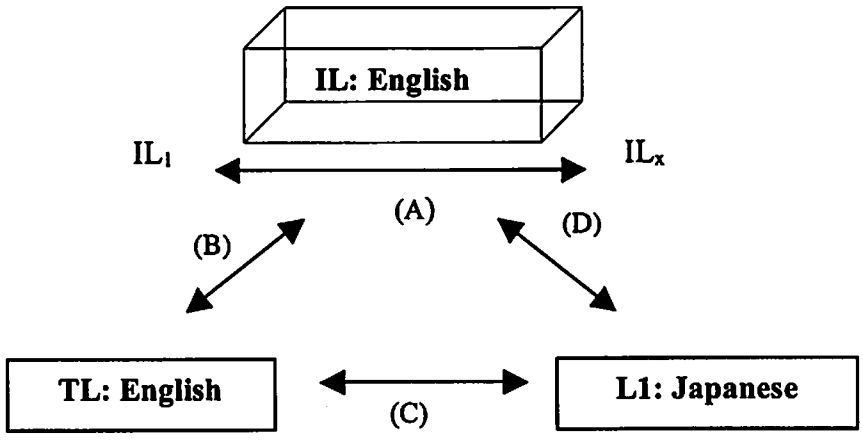

Figure 1. Multiple comparison of L1, TL and IL corpora

"IL_1... IL_x" in Figure 1shows that the L2 learner texts may be divided into subcorpora according to the subjects' academic year. Let me call studies based on such learner subcorpora *IL-IL comparison*. IL-IL comparison can be of several different types, depending on the learner variables. For instance, if the independent variable (i.e. the variable that you manipulate) is age or the academic year of the learners, with all other variables constant, one can make a comparison of different IL corpora from different age groups. In ICLE[7], on the other hand, the age (or proficiency level) factor is held constant, and research using ICLE centers around the IL characteristics of different L1 groups. Although there are some variations, I will call this type of comparison IL-IL comparison.

A comparison between L2 corpora and TL corpora can also be made (see (B) in Figure 1). One can use either a general standard corpus such as BNC to look at differences in, for example, lexicogrammar between native speakers and L2 learners, or use a more comparable corpus of native-speaker texts, e.g. LOCNESS in ICLE, to compare like with like. Let me call this type of comparison *IL-TL comparison*.

If TL corpora are compared with L1 corpora, it is called *TL-L1 comparison* (see (C) in Figure 1). This kind of comparison can be used for describing the target adult grammar system and identifying potential causes of L1 transfer. This analysis should be combined with L2 corpus analysis. TL-L1 comparison could provide significant information on the influence of the source language on the acquisition of the target language.

The final type of comparison which can be made is that between IL corpora and L1 mother tongue corpora (see (D) in Figure 1). Let me call this *L1-IL comparison*. L1 corpora can provide features of the L2 learners'

[7] International Corpus of Learner English. See Table 3.

native language, which can help us understand potential sources of L1-related errors or overuse/underuse phenomena. Despite the sophistication of recent error taxonomies, it is rather difficult to distinguish interlingual errors from intralingual ones unless some empirical data are available on the pattern of a particular linguistic feature in both languages. L1-IL comparisons will provide fundamental data for interlanguage studies.

Table 2 summarizes each comparison type and its description:

Table 2. Multiple Comparison Approach

Comparison	Description
IL-IL comparison	Comparisons between different stages of ILs or ILs by learners with different L1 backgrounds.
IL-TL comparison	Comparisons between learner corpora and target language corpora (i.e. ELT textbook corpora in the present study or general native corpora).
TL-L1 comparison	Comparisons between target language corpora and L1 mother tongue corpora to identify potential causes of L1 transfer.
L1-IL comparison	Comparisons between L1 corpora and learner corpora to identify L1-related errors or overuse/underuse phenomena.
IL-L1-TL comparison	Combination of the above comparisons to identify the complex relationship between IL, L1 and TL corpora on L2 learners' error patterns or overuse/underuse phenomena.

5. Current Projects of Learner Corpora

Table 3 summarizes the learner corpus projects currently underway around the world. The table focuses on the design criteria, size, annotation, and availability, and lists the relevant references. It also identifies whether each project involves the comparison of different IL corpora (developmental/ different L1s) with L1/TL corpora from the viewpoints of multiple comparison.

Table 3. Learner Corpus Projects Around the World

Project	Subjects/ Tasks Size	Annotation Availability	Comparison	References
EUROPE:				
International Corpus of Learner English (ICLE)	- University EFL 3/4 year students - 15 nationalities - Written essays - 3 million	- Error tagged - POS tagged - Available in 2001	- IL – IL (different L1s) - TL – IL	Granger (1993, 1994, 1996, 1998)
Longman Learners' Corpus (LLC)	- All-levels - Written essays - 10 million	- POS tagged - Available for commercial purposes	- IL – IL	Gillard and Gadsby (1998)
Polish-English Language Corpus Research and Applications (PELCRA)	- All-levels - Written/spoken essays - Polish learners	- POS tagged - Not available	- IL – IL (developmental) - L1 – IL - TL – IL	Uzar (1997) Mason & Uzar (2000)
The ISLE Corpus of non-native spoken English	- 20 minute speech - German & Italian intermediate learners of English	- Orthographic - Phone-stress - Available from ELRA	- TL - IL	http://nats-www.informatik.uni-hamburg.de/~isle/speech.html
JPU (Janus Pannonius University) Corpus	- University EFL - Written - c.400,000	- Plain text - Will be available	- IL – IL (developmental)	József (1999)
Cambridge Learners Corpus (CLC)	- All-levels - 16 million	- POS tagged - Error-tagged (2.5 million) - Not available	- IL – IL	http://uk.cambridge.org/elt/reference/clc.htm
Indianapolis Business Learner Corpus (IBLC)	- US univ. business students - business writing - plain text	- Plain text - Not available	- IL – IL (different L1s)	Connor & Precht (1998)
ASIA:				
JEFLL Corpus (Japan)	- All levels; EFL - Written & spoken - 350,000	- POS-tagged - Error-tagged (partial) - available in 2003	- IL – IL (developmental) - L1 – IL - TL – IL	Tono (2002) Tono (2000a, b) Tono and Aoki (1998)

Table 3. (continued)

ASIA:				
Corpus of English by Japanese Learners	- All levels; EFL - Written - 1 million	- Plain text - Error tagged (partial) - Will be available	- IL – IL (develop mental)	Asao (1998)
Japanese/ English Translation corpus	- junior & senior high EFL students - L1/L2 translation	- Plain text - Available via the web	- TL – IL	http://home.hiroshima-u.ac.jp/d052121/eigo1.html
TELEC Student Corpus	- Hong Kong learners - Univ. exam scripts - 3 million	- Plain text - Restricted availability	- TL – IL	Allan (1998)
PolyU Corpus	- Postgraduates - thesis drafts, etc. - 282,000	- Plain text - not available	- TL – IL	Farmer and Mead (1998)
NTOU Corpus	- EFL - 53,000	- Plain text	- TL – IL - IL – IL	Chen (1998)
A parallel corpus of Japanese learners of English	- Short English compositions - Paired with Japanese translations & NS's rewritings	- Database format	- TL – IL - IL – L1	Mark (1998a, b)
MET Corpus	- Chinese middle school students - Written - c. 150000	- Plain text	- TL – IL	Anping (1998)
HKUST Corpus of Learner English	- University EFL Chinese students - 10 million - Written essays & exam scripts	- POS tagged (1M) - Error tagged (100,000 words)	- IL – IL	Flowerdew (1996) Flowerdew (1997) Milton (1998) Milton and Tsung (1993)
Standard Speaking Test (SST) Corpus	- Oral proficiency interview test corpus - 100M - Japanese EFL learners	- Error tagged - Back translation - Parallel (with Japanese)	- IL – IL - IL – TL - IL –L1	Tono et al. (2001)

6. An Example of the Multiple Comparison Approach: L2 Acquisition of Argument Structure

6.1. Overview

Let me demonstrate how the multiple comparison approach works. Here I will report on the study of the patterns of misuse of verb subcategorization frames (henceforth SF) by Japanese learners of English. The acquisition of SF patterns is often associated with the broader issue of the acquisition of argument structure (Pinker, 1984, 1987, 1989). The development of argument structure can possibly be influenced by several factors. Four main factors (verb semantics, learning stage, L1 knowledge, and L2 input) were selected and the relationship of these factors on the use/misuse of argument structure was investigated. An L1 corpus was used to define the influence of verb SF patterns in L1 while ELT textbook corpora were used for determining the degree of exposure to certain SF patterns in the classroom. Based on the data from these corpora, I compared the SF patterns of a group of high-frequency verbs in the JEFLL corpus.

6.2. Factors Affecting the Acquisition of SF Patterns

6.2.1. Views From L1 Acquisition Research

There are competing theories seeking to explain the acquisition of argument structure in L1 acquisition. The major issue is how to explain the children's initial acquisition of argument structure. Do they learn the argument structure patterns from the meaning of verbs they initially acquire or do they acquire the structure first, then move on to the acquisition of verb meanings? The two bootstrapping hypotheses, semantic and syntactic, claim that the acquisition of argument structure is bootstrapped by first acquiring either semantic or syntactic properties of the verbs. Pinker (1987) is keen to identify what happens at the very first stage of syntax acquisition while Gleitman (1990) states the hypothesis in such a way that it applies not only to the initial stage but the entire process of acquisition. As Grimshaw (1994) argues, however, those two hypotheses could complement each other, once the initial state issue is solved.

Despite the difference in the view of how the acquisition of argument structure starts, Pinker and Gleitman have both agreed that knowledge of the relationship between a verb's semantics and its morpho-syntax is guided in part by UG because adult grammars go beyond the input available. Goldberg (1999), on the other hand, proposes the theory which claims that it is a construction itself which carries the meaning. Although verbs and associated argument structures are initially learned on an item-by-item basis, increased vocabulary leads to categorization and generalization. Light verbs, due to the fact that they are introduced at a very early stage and are highly frequent, act as a centre of gravity, forming the prototype of the semantic category associated with the formal pattern.

The perspective which Goldberg and other construction grammarians have taken on children's grammar learning is fundamentally that of 'general' nativism. They reject the claim of 'special' nativism in its particular guise of Universal Grammar, but they still assume other, innate, aspects of human cognitive functioning accounting for language acquisition. As a matter of fact, this position is increasingly widely supported nowadays by those who take more general cognitive approaches, including so-called emergentism (Elman, et al., 1996; MacWhinney, 1998), cognitive linguistics (Langacker, 1987, 1991; Ungerer & Schmid, 1996) and constructivist child language research (Slobin, 1997; Tomasello, 1992).

One of the purposes of my study presented here is to determine the relative effect of L1 knowledge, classroom input, developmental factors and inherent verb semantics on the use/misuse and overuse/underuse of SF patterns by Japanese learners of English. It should be noted that the study does not need to rely on a specific acquisition theory at this stage. Rather, this corpus-based study will reveal the nature of IL development by weighting the factors which are possibly relevant to the acquisition of argument structure. This will help to evaluate the validity and plausibility of the claims made in L1 acquisition research in the light of SLA theory construction. For instance, if the study shows the strong effect of frequencies of verbs used in the ELT textbooks on the use of particular SF patterns, then the results may indicate that L2 acquisition can be better explained by the theory that attaches more importance to the frequency of the items to be acquired in the input. From this viewpoint, Goldberg's theory is more plausible. On the other hand, if the effect of verb semantics is highly significant, one may be inclined to agree with the theory that emphasizes the semantic properties of verbs as the driving force for the acquisition of argument structure. Hence one would be more likely to adopt the theoretical framework of semantic bootstrapping theory prepared by Pinker.

This study has the potential, therefore, to tease out possible factors affecting L2 acquisition in the light of L1 acquisition theories, making observations on L1, TL, and IL corpus data while controlling all those selected factors, and finally each factor given a weighting according to the results of the corpus analysis. This weighting of the factors relevant to L2 acquisition will then contribute to the decision-making about which L1 acquisition theory is more plausible.

6.2.2. Views From L2 Acquisition Research

Whilst a vast literature exists on the L1 acquisition of semantics-syntax correspondences, second language acquisition of verb semantics and morpho-syntax only really attracted detailed attention in the 1990s. The major issues in L2 acquisition of argument structure are: (1) whether or not

L1 effects are strong in this area, (2) whether there is any evidence of universal patterns of development, and (3) the role of input in the acquisition of argument structure.

From the previous SLA studies, L1 effects appear strong in the acquisition of argument structure. Especially SF frames are a case in point. Recently, there has been much investigation of the proposal that the SF requirements of a lexical item might be predictable from its meaning (Levin, 1993: 12). The issue here is whether such lexical knowledge in L1 or in UG will affect L2 acquisition. This is usually investigated through the study of the acquisition of argument structure alternations - alternations in the expressions of arguments, sometimes accompanied by changes of meaning.

In the case of dative alternations (White, 1987, 1991; Bley-Vroman & Yoshinaga, 1992; Sawyer, 1996; Inagaki, 1997; Montrul, 1998), most evidence seems to indicate that the initial hypothesis for knowledge of syntactic frames is the L1. The studies on the locative alternations (Juffs, 1996; Thepsura, 1998) indicate that there is a difference in the way a hypothesis is formed by learners at different proficiency levels. While beginning learners start off with a wider grammar for non-alternating locative verbs, the very advanced learners end up with a narrower grammar (Juffs 1996). There are several studies (Zobl, 1989; Hirakawa, 1995; Oshita, 1997) that indicate the L1 transfer effect on transitivity alternations and the unergative/unaccusative distinction. To recapitulate, L1 effects appear strong in this area of grammar. Based on their L1, learners transfer and overgeneralize in dative movement and the locative alternation. They also show a preference for morphology for inchoatives. Consequently, learners are helped if their L1 has certain features which are also in the L2. Advanced learners, however, seem able to recover from overgeneralization errors in some instances by acquiring narrow conflation classes which are not in their L1. Thus there seems to be an interaction effect between L1 influence and proficiency levels.

In spite of studies showing the L1 effects, there is some evidence of universal patterns of development. Learners from a variety of backgrounds seem to use passive morphology for NP movement in English L2 with pure unaccusatives (Yip, 1994; Oshita, 1997). English-speaking learners of Spanish seem to use *se* selectively for the same purpose even when it is not required with unaccusative verbs (Toth, 1997). Montrul (1998) found evidence which indicates that L2 learners have an initial hypothesis that all verbs can have a default transitive template, allowing an SVO structure in English even with pure unaccusatives and unergatives. Hence, learners seem to overgeneralize causativity in root morphemes more than children acquiring their first language do.

There are not many studies on the role of input in the acquisition of verb

meaning and the way such knowledge relates to syntax. Inagaki (1997) argues that the fact that the English native speakers showed a strong distinction between the *tell/whisper* verbs than between the *throw/push* verbs is also consistent with the hypothesis that the double-object datives containing the *tell* verbs were more frequent in the input than those containing the *throw* verbs (ibid, p.660). Unfortunately, measuring the frequency in L2 input is difficult since so few analyses of input corpora for L2 learners exist (Juffs, 2000: 202).

6.3. The Relationship Between Factors and Corpora Used

Table 4 shows the relationship between the factors to be examined in this study and how corpus data can supply the relevant information. The multiple comparisons of L1, TL, and IL corpora only make this design possible. Note that the primary purpose of this study is not to identify the role of specific UG constraints in L2 acquisition. Rather, the study aims to capture the cause-effect relationship among those variables and to identify their relative effects on the acquisition of argument structure in L2 English.

Table 4. The Relationship Between the Factors in This Study and the Information From the Corpora Used

Factors	Corpus data
The L1 effects	Frequency of similar/different argument Structure properties in L1 corpus
The L2 input	Frequency of subcategorization patterns in ELT textbook corpus
Developmental stages	Frequency of use/misuse of subcategorization patterns from the developmental IL corpus
The L2 internal effects	Frequency of different verb classes and alternations from the IL corpus

6.4. Research Design

6.4.1. Research Questions

This study has the following research questions:

I. Which of the following variables affect L2 acquisition of argument structure most?
   - The L1 effects
   - The L2 input effects
   - The L2 internal effects
   - The developmental effects

II. Are there any interaction effects between the variables? If so, how?

The clarification of the relationship between the above questions will contribute to the current SLA research especially in terms of the possible role of L1 knowledge, L2 classroom input, and verb semantics-syntax correspondences in the acquisition of argument structure.

6.4.2. Variables and Operational Definitions

Each variable is operationally defined as follows:

I. L1 effects:

L1 effects were examined with respect to the following two aspects: the degree of similarities in SF patterns between English and Japanese in terms of (a) the degree of SF matching and (b) frequencies of the similar SF patterns in the L1 Japanese corpus and the COMLEX Lexicon (TL).

II. L2 input effects:

L2 input effects were defined as "the frequencies of the given SF patterns in the L2 textbook corpus".

III. L2 internal effects:

These characteristics pertain to the English verb system. They were defined as "the difference in verb classes and alternation types based on Levin's (1993) classification".

IV. Developmental effects:

Developmental effects were simply defined as the three groups of the subjects based on their school years (Year 7-8; 9-10; 11-12).

6.4.3. Extraction of SF Patterns

In this study, I parsed the learner and textbook corpora using the Apple Pie Parser (APP), a statistical parser developed by Satoshi Sekine at New York University (see Sekine, 1998, for details). The accuracy rate of the APP is approximately 70%, hence it was not very efficient to extract SF patterns automatically using the APP alone. Consequently, after running the parser over the corpus, I exported concordance lines of verbs with syntactic information into Excel and categorized them into SF patterns using pattern matching. This proved to be an efficient means of studying verb SFs.

The Comlex Lexicon (Macleod et al., 1996; Grishman et al., 1994) was also referred to for frequency information relating to each subcategorization frame in the TL corpus. The Comlex Lexicon itself does not provide complete frequency data for SF patterns. However, it has frequency information for the subcategorization frames of the first 100 words appearing in the Brown Corpus. I calculated the percentages of each SF patterns in the Comlex database and used the information to supplement the data from the textbook corpora.

For the L1 corpus, a Japanese morphological analyzer, *ChaSen* (Matsumoto, et al., 2000), was used for tokenization and morphological analysis and the frequencies of SF patterns were detected by using pattern matching. SF extraction was done after extracting all the instances of a particular verb under study, and thus manual postediting was also possible.

6.4.4. Categorisation of Verb Classes

The verb classification of Levin (1993) was used to categorize verbs into groups with similar meanings. Levin divided verb classes into two major categories: (a) a list of diathesis alternations and (b) a list of semantically coherent verb classes. While Levin's classification is very important for the study of lexical knowledge in the human mind, it should also be noted that the actual use of those verb classes is only limited to certain verb classes only. For instance, out of 49 verb classes Levin created, only 22 classes were found in the top 40 most frequent verbs in the BNC. Note that a small number of categories, which meet essential communication needs (e.g. 'communication', 'motion', and 'change of possession'), dominate the verb usage. This shows that the input consists of only a handful of highly frequent verb classes and the rest of the classes are quite infrequent.

The information on Japanese SFs was provided in the IPAL Electronic Dictionary Project. After making the matching database of corresponding verbs in English and Japanese, the frequency information of SFs was extracted from the Comlex Lexicon. SFs were extracted from the ELT textbook corpus for TL (English) and from the Japanese corpus I made for L1 Japanese.

6.4.5. Log-Linear Analysis

The objective of log-linear analysis is to find the model that gives the most parsimonious description of the data. For each of the different models, the expected cell frequencies are compared to the observed frequencies. A Chi-square test can then be used to determine whether the difference between expected and observed cell frequencies is acceptable. The least economical model, the one that contains the maximal number of effects, is the *saturated* model; it will by definition yield a 'perfect' fit between the expected and observed frequencies. The associated χ^2 [MM: Something is missing here.] is zero. In this study, the procedure called *backward deletion* or *elimination* was employed (Christensen, 1997). This begins with the saturated model and then effects are successively left out of the model and it is checked whether the value of χ^2 [MM: Something is missing here.] of the more parsimonious model passes the critical level. When this happens, the effect that was left out last is deemed essential to the model and should be included.

Several statistical packages contain procedures to carry out a log-linear

analysis on contingency tables. In SPSS (a statistical program), the HILOG-LINEAR procedure is available in order to find the best model, which, by means of the option BACKWARD, helps in selecting the best-fitting model through backward deletion. Other statistical programs such as STATISTICA and SAS also have the same types of procedures although the presentation formats of the statistical results are different. In this study, STATISTICA was mainly used to do the model testing.

6.4.6. Subcategorization Frame Database

For each high-frequency verb, the following information was gathered and put into the database format:

- Parsed example sentences containing the target verb
- School year categories (year 7-8; 9-10; 11-12)
- Verb name
- Verb class
- Verb meaning
- Alternation type
- SF for each example
- Frequency of SF in COMLEX Lexicon
- TL frequency of the given SF (i.e. textbook corpora)
- Learner errors
- Parsing errors
- Japanese verb equivalents
- L1 frequency of the equivalent SF (i.e. Japanese corpus)

The database was made for each of the high-frequency verbs and the data was exported to the statistical software used for further analysis. In order to process the data by log-linear analysis, the frequencies of TL and L1 were converted into categorical data ([HIGH]/ [MID]/ [LOW]). In order to study the acquisition of argument structure, ten verbs were selected for the analysis (*bring, buy, eat, get, go, like, make, take, think*, and *want*). While it would be desirable to cover as many verbs as possible from different verb classes for the study, due to the fact that the frequencies of SF patterns become extremely small if I had chosen low frequency verbs, I had to reduce the number of verbs under investigation to the ten most frequent ones in my data. These verbs, allowed a sufficient number of observations to be made for each verb. Even though they are frequent, *be* and *have* were excluded from the analysis because their status as lexical verbs is very different from other verbs. Due to the limitation of the space, I cannot describe into detail the SF patterns of those verbs selected for the study (see Tono, 2002, for further details).

6.5. The Results of Log-linear Analysis for Individual Verbs

Log-linear analysis tested the model by the combination of the following six factors (Factors 1 – 6):

- L2 learners developmental factor (Factor 1):
  - 3 levels: Year 7-8/ Year 9-10/ Year 11-12
- Subcategorization matching between L1 and L2 (Factor 2): [by comparing similarities in SFs based on IPAL and Comlex]
  - 2 levels: Matched/ Unmatched
- Subcategorization frequencies of each SF pattern in COMLEX (Factor 3):
  - 3 levels: High/ Mid/ Low
- Subcategorization frequencies of each SF pattern based on IPAL in L1 Japanese Corpus (Factor 4):
  - 3 levels: High/ Mid/ Low
- Subcategorization frequencies of each SF pattern in Textbook Corpus (Factor 5):
  - 3 levels: High/ Mid/ Low
- L2 learner errors (Factor 6):
  - 2 levels: Error/ Non-error

The results of log-linear analysis for each individual verb revealed quite an interesting picture of the relationship between learner errors and a number of relevant factors. Here let me summarize the results by putting all the best fitting models together in a table and examining which factor exerts most influence on learner performance across the ten verbs. Table 5 shows the summary of log-linear analysis for the ten verbs used for the study. The numbers in each cell show the main and interaction effects of the six variables (Factors 1 – 6). Thus, in the case of the verb bring, there is no main effect but only one two-way interaction (factors 1 x 5, denoted as 51). All the other interactions are three-way (643, 543, 532, 432). This shows that the best fitting model for the verb bring is the combination of the following interaction effects [51, 643, 543, 532, 432]. In the case of the verb buy, on the other hand, the main effect of Factor 1 (YEAR) was observed (see the number 1 in Table 5) together with the combinations of the following three-way interactions [642, 632, 543, 542, 532].

Table 5. The Summary of Log-linear Analysis

Verbs	Factor 1 YEAR	Factor 2 SUBMATCH	Factor 3 COMLEX	Factor 4 L1FRQ	Factor 5 TEXTFRQ	Factor 6 LERR
bring	**51**	532, 432 532, 432	643, 543 432	643, 543 **51**	543, 532	643
buy	**1**	642, 632 542, 532	532, 543 632	642, 543 542	543, 542 532	642, 632
eat		642, 632 432, 521	632, 531 432	642, 432	531, 521	642, 632
get	**1, 61**	432, 532	643, 543 432	643, 543 432	543, 532	**61**, 643
go	**1**	632, 542 432, 532	632, 543 432, 532	543, 542 432	543, 542, 532	632
like	**51**	652, 542 532	643, 543 532	643, 543 542	**51**, 652, 543 542, 532	652, 643
make	**1**	642, 632 542, 532	632, 543 532	642, 543 542	543, 542, 532	642, 632
take	**51**	632, 632 532	632, 543 532	642, 543	**51**, 543, 532	642, 632
think	**1**	642, 632 542, 532	632, 543 532	642, 543 542	543,542, 532	642, 632
want	**31**	642, 632 542	632, 543 **31**	642, 543 542	543, 542	642, 632

In order to analyse the interactions, graphical interpretations of higher dimensional log-linear models are sometimes used (e.g. McEnery, 1995; Kennedy, 1992). However, as I am dealing with six dimensional models here, attempting to interpret them using graphical models would be extremely complicated. Also, my primary aim is not to interpret individual cases but to capture the overall picture of how factors are related across different verbs. Consequently I will not interpret the models visually. Rather I will provide a brief narrative outlining the major results.

6.5.1. Distinctive Effects of the School Year

Table 5 shows that the school year factor (YEAR) has a very strong effect across all of the verbs. For five out of the ten verbs (*buy, get, go,* make, and *think*), the main effect of YEAR was observed. The YEAR effect also has two-way interactions with the factor of text frequency

(TEXTFRQ) for three verbs (*bring, like, take*), the SF frequencies in English (COMLEX) for the verb *want* and with the learner error/non-error factor (LERR) for the verb *get*. This shows that the school year influences the way L2 learners use these verbs. It involves both the use/misuse and the overuse/underuse of verbs.

6.5.2. Strong Effects of the SF Frequencies in the Textbook Corpus

We can also see from the summary table that there are strong two-way effects between YEAR and TEXTFRQ. Note that there is only one case (652 for the verb *like*) of the interaction of the textbook frequency factor (Factor 5) with the learner error factor (Factor 6). This implies that the factor of SF frequencies in the textbooks mainly affect the overuse of the verbs, not the use/misuse.

6.5.3. SF Similarities and Frequencies in L1 and TL

The factors such as the degree of similarity in SF patterns between English and Japanese (SUBMATCH: Factor 2), the frequency from COMLEX (Factor 3), and the frequency of SF patterns in L1 Japanese (L1FRQ: Factor 4) appear many times with the learner error factor (LERR: Factor 6). These factors are different from the factors of school year and textbook frequency, as they represent more inherent linguistic features of the verbs and L1 effects. Each of the effects, however, is not very strong because none of them survived backward deletion for the one-way or two-way effects. It seems that only the interactions of these factors affect learners' use/misuse of the verbs.

6.6. The Effects of Verb Classes and Alternation Types

In order to analyze the relationship between verb classes/alternation types and the results of log-linear analysis, I used correspondence analysis. Instead of looking at each verb, I labeled each verb with its verb semantic classes and alternation types. I then gave scores to each factor according to the significance of its effects as shown in Table 5, for instance, if a certain factor has a one-way interaction, which is the strongest, I gave 10 points; if it has a two-way interaction, I gave 5 points for each of the factors involved. Only 1 point was given for each of the three-way effects. In this way, I quantified each of the effects in the best model for each verb in Table 5 and used correspondence analysis to see the relationship between the six factors and verb classes and alternation types.

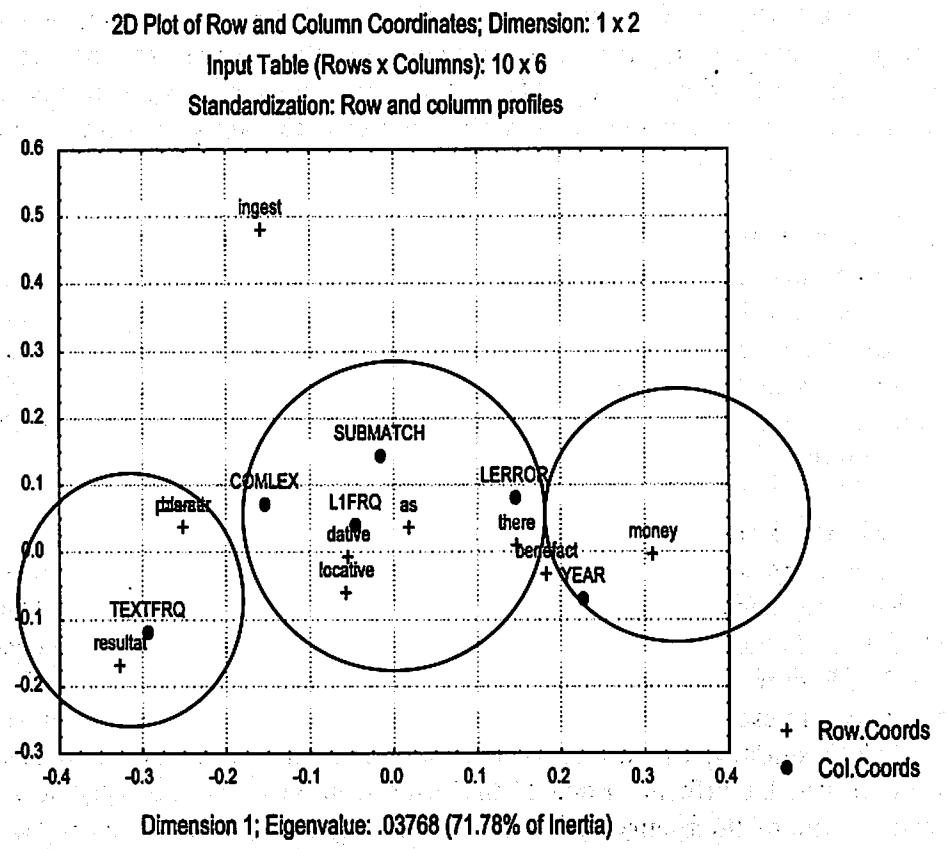

Figure 2. Correspondence analysis (alternations x effects)

Figure 2 shows the results of re-classification of the effects found by log-linear analysis for each verb according to verb alternation types. Correspondence analysis plots the variables based on the total Chi-square values (i.e. inertia) and the more the variables cluster together, the stronger the relationship is. Dimension 1 explains 71% of inertia, so we should mainly consider Dimension 1 as a primary source of interpretation. The figure shows clearly that there are three major groups of effects: the factor of SF patterns in the textbook corpus (TEXTFRQ) in the left corner, three effects (SF frequencies in L1 corpus, the degree of matching between English and Japanese SFs, the SF frequencies in COMLEX) in the centre, and the learner error effect and the school year effect toward the right side. As was discussed above, the school year represents the developmental aspect of verb learning while the three factors in the middle represent linguistic features in each verb, and the textbook frequency represents L2 input effects.

There is a tendency for verbs involving benefactive alternations (*buy, get, make,* and *take*), sum of money alternations (*buy, get,* and *make*), and

there insertions (*go*) to cluster around the school year factor and the error factor. Thus these verb alternation classes seem to be sensitive to the developmental factor of acquisition.

Dative (*bring, make, take, think*, and *want*), locative (*take, go*) and as alternations (*make, take* and *think*) cluster around inherent linguistic factors such as the degree of SF matching and SF frequencies in L1 and TL.

The verbs involving resultative alternations (*bring* and *take*) cluster around the SF frequencies factor in the textbook corpus. Post-attributive and *blame* alternations are both features of the verbs *like* and *want*. These two alternation types also cluster together close to the textbook frequency effect. These are the verbs showing a strong relationship with L2 input effects.

There is only one alternation type that did not cluster with any other groups; ingestion (*eat*). The verb eat was very frequent in learner data and was thus included in the analysis, but it turned out that there were neither very many errors nor many varieties of alternations for this verb. The results look very different from the other nine verbs.

7. Conclusion

The study shows some interesting findings about the developmental effect on learner errors, L2 input effects on the overuse of SF patterns, L1 effects on some SF errors and L2 internal effects (i.e. verb classes and alternations) on the overall use of verbs. I hope that this paper has demonstrated the effectiveness of the multiple comparison approach of IL, L1 and TL corpora. A large body of L2 learner corpora will become an indispensable resource for SLA researchers in the near future. Also, as we work together with researchers in Artificial Intelligence or Natural Language Processing, there will be very intriguing possibilities of developing a computational model of L2 acquisition. Machine learning techniques will facilitate the testing of prototypical acquisition models or a collection of probabilistic information of IL based on corpora. Computational analyses of IL data will shed light on the process of IL development in the way we never thought possible. For this to happen, there will be a genuine need for well-balanced representative corpora of L2 learners.

Acknowledgements

I would like to thank the editors of this volume for their valuable comments. This research was supported by the Miyata Research Grant by Meikai University and by the Grant-in-Aid for Scientific Research (Category B) entitled "Building linguistic resources for L2 lexicography" from the Japanese Ministry of Education, Science, Sport and Culture. Correspondence should be sent to Yukio Tono, Department of Foreign

Languages and Cultures, Meikai University, 8 Akemi Urayasu, Chiba 279-8550, Japan.

References

Allan, Q. G. (1998). The TELEC student corpus: A resource for teacher development. In S. Granger & J. Huang (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 4-6). Chinese University of Hong Kong.

Allwright, R. (1980). Topics, turns and tasks: Patterns of participation in language learning. In D. Larsen-Freeman (Ed.), *Discourse analysis in second language acquisition research.* Rowley, Mass: Newbury House.

Anping H. (1998). A Corpus-based analysis of Chinese middle-school-students' English spelling errors. In S. Granger & J. Huang (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 6-9). Chinese University of Hong Kong.

Asao, K. (1998). Corpus of English by Japanese learners. In S. Granger & J. Huang (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp.10-13). Chinese University of Hong Kong.

Barlow, M., & Kemmer, S. (Eds.) (2000). *Usage-based models of language.* CSLI publications. Stanford University.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Longman.

Bley-Vroman, R., & Yoshinaga, N. (1992). Broad and narrow constraints on the English dative alternation: some fundamental differences between native speakers and foreign language learners. *University of Hawai'i working papers in ESL* (vol. 11, pp. 157-199). University of Hawai'i at Manoa.

Chen, H. -J. H. (1998). Underuse, overuse and misuse in Taiwanese EFL learner corpus. In S. Granger and J. Hung (Eds.), (1998), pp. 25-28.

Chomsky, N. (1962). Paper given at the University of Texas 1958, 3rd Texas Conference on Problems of Linguistic Analysis in English. Austin: University of Texas.

Chomsky, N. (1964). Formal discussion. In U. Bellugi, & R. Brown (Eds.), *The acquisition of language.* Monograph of the Society for Research in Child Development, 29.

Christensen, R. (1997). *Log-linear models and logistic regression.* Springer-Verlag.

Clahsen, H. (1980). Psycholinguistic aspects of L2 acquisition. In S. Felix (Ed.), *Second language development: Trends and issues.* Tübingen:

Gunter Narr.

Clahsen, H. (1984). The acquisition of German word order: a test case for cognitive approaches to L2 development. In Andersen, R. W. (Ed.), *Second languages: A cross-linguistic perspectives* (pp. 219-242). Rowley, MA: Newbury House.

Clahsen, H., Meisel, J., & Pienemann, M. (1983). *Deutsch als Zweitsprache: der Spracherwerb ausländischer Arbeiter.* Tübingen: Gunter Narr.

Connor, U., & Precht, K. (1998). Business English: learner data from Belgium and the U.S. In S. Granger & J. Hung (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 29-33). Chinese University of Hong Kong.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*, 161-9.

Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning, 23*, 37-53.

Dulay, H., & Burt, M. (1975). Creative construction in second language learning and teaching. In M. Burt & H. Dulay (Eds.), *On TESOL '75: New directions in second language learning, teaching and bilingual education.* Washington, D.C.: TESOL.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24 (2)*, 143-188.

Ellis, R. (1994). *The study of second language acquisition.* Oxford University Press.

Elman, J. L., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspectives on development.* Cambridge, MA: A Bradford Book.

Fanselow, J. (1977). Beyond rashomon: Conceptualising and describing the teaching act. *TESOL Quarterly, 11*, 17-39.

Farmer, R., & Mead, K. (1998). The language of citations: an analysis via computer learner corpus. In S. Granger & J. Hung (Eds.) (1998), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 34-37). Chinese University of Hong Kong.

Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Eds.), *The power of CALL* (pp. 97-113). Houston, TX: Athelstan.

Flowerdew, L. (1997). Interpersonal strategies: investigating interlanguage corpora. *RELC Journal, 28(1)*, 72-88.

Fries, C. C. (1957). A foreword to Lado (1957).

Gillard, P., & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger, (Ed.), *Learner English on computer* (pp. 159-171). London and New York: Addison Wesley Longman.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition, 1*, 3-55.

Goldberg, A. (1999). The emergence of the semantics of argument structure constructions. In B. MacWhinney (Ed.), *The emergence of language.* Mahwah, NZ: Lawrence Erlbaum Associates.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: design, analysis and exploitation* (pp. 57-69). Amsterdam: Rodopi.

Granger, S. (1994). The learner corpus: a revolution in applied linguistics. *English Today, 39(10/3)*, 25-29.

Granger, S. (1996). From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994* (pp. 37-51). Lund: Lund University Press.

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

Granger, S., & Huang, J. (Eds.) (1998). *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching.* Chinese University of Hong Kong.

Grimshaw, J. (1994). Lexical reconciliation. In L. Gleitman & B. Landau (Eds.), *The acquisition of the lexicon.* Cambridge, MA: MIT Press.

Grishman, R., Macleod, C., & Meyers, A. (1994). Comlex syntax: Building a computational lexicon. *Proceedings of 15th International Conference in Computational Linguistics (COLING 94).* Kyoto, Japan, August 1994.

Hakuta, K. (1976). A case study of a Japanese child learning English as a second language. *Language Learning, 26*, 321-51.

Hendrickson, J. M. (1976). *Error analysis and error correction for adult intermediate ESL learners: an experiment.* Unpublished doctoral dissertation, Ohio State University.

Hirakawa, M. (1995). L2 acquisition of English unaccusative constructions. In D MacClaughlin & S. McEwen (Eds.), *Proceedings of the 19th Boston university conference on language development 1* (pp. 291-302). Somerville, MA: Cascadilla Press.

Inagaki, S. (1997). Japanese and Chinese learners' acquisition of the narrow-range rules for the dative alternation in English. *Language Learning, 47*, 637-669.

József, H. (1998). *Advanced writing in English as a foreign language: A corpus-based study of processes and products*. Unpublished doctoral dissertation. Janus Pannonius University, Pécs, Hungary.

Juffs, A. (1996). *Learnability and the lexicon: Theories and second language acquisition research*. Amsterdam: John Benjamins.

Juffs, A. (2000). An overview of the second language acquisition of links between verb semantics and morpho-syntax. In J. Archibald (Ed.), *Second language acquisition and linguistic theory* (pp. 187-227). Oxford: Blackwell.

Kennedy, J. (1992). *Analyzing qualitative data. Log-linear analysis for behavioural research*. New York: Praeger.

Lado, R. (1957). *Linguistics across cultures*. Ann Arbor, MI: University of Michigan Press.

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

Langacker, R. W. (1991). *Foundations of cognitive grammar. Vol.2: Descriptive application*. Stanford, CA: Stanford University Press.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: In honour of Jan Svartvik*. London: Longman.

Leech, G. (1992). Corpus linguistics and theories of linguistic performance. In J. Svartvik (Ed.), *New directions in corpus linguistics: Proceedings of nobel symposium 82* (pp. 125-148). Berlin: Mouton de Gruyter.

Levin, B. (1993). *English verb classes and alternations*. Chicago: The University of Chicago Press.

Lewandowska-Tomaszczyk, B., & Melia, J. P. (Eds.) (2000). *PALC'99: Practical applications in language corpora*. Peter Lang GmbH, Frankfurt am Main.

Ljung, M. (1990). *A study of TEFL vocabulary* (Stockholm Studies in English 78). Stockholm: Almqvist & Wiksell.

Ljung, M. (1991). Swedish TEFL meets reality. In S. Johansson & A. -B. Stenström (Eds.), *English computer corpora*, pp. 245-256. Berlin: Mouton de Gruyter.

Long, M. (1983). Does second language instruction make a difference? A review of the research. *TESOL Quarterly, 17*, 359-382.

Macleod, C. Meyers, A., & Grishman, R. (1996). The Influence of Tagging on the Classification of Lexical Complements. *Proceedings of the 16th International Conference in Computational Linguistics (COLING 96)*. Copenhagen, Denmark, August.

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mark, K. (1998a). A parallel learner corpus approach to English curriculum development at a Japanese university. In S. Granger & J. Hung (Eds.),

Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching, (pp. 89-90). Chinese University of Hong Kong.

Mark, K. (1998b). The significance of learner Corpus Data in Relation to the Problems of Language Teaching. *Bulletin of General Education, 312*, 77-90. Meiji University.

Mason, O., & Uzar, R. (2000). NLP meets TEFL: Tracing the zero article. In B. Lewandowska-Tomaszczyk & J. P. Melia (Eds.), *PALC'99: Practical applications in language corpora*, pp. 105-116. Peter Lang GmbH, Frankfurt am Main.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., & Asahara, M. (2000). *Morphological analysis system ChaSen version 2.2.1 manual.* Nara Institute of Advanced Technology.

McEnery, T. (1995). *Computational pragmatics: Probability, deeming and uncertain references.* Unpublished doctoral dissertation. Lancaster University.

McEnery, T. & Kifle (2001). Non-native speaker and native speaker argumentative compositions - a corpus-based study. In J. Flowerdew (Ed.), *Academic discourse* (pp.182-195). Harlow: Longman.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

Meisel, J. (1980). Linguistic simplification. In S. Felix (Ed.), *Second language development: Trends and issues.* Tübingen: Gunter Narr.

Meisel, J., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition, 3*, 109-135.

Milton, J. (1998). WORDPILOT: enabling learners to navigate lexical universes. In S. Granger & J. Huang (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 97-98). Chinese University of Hong Kong.

Milton, J., & Tsang, E. (1993). A corpus-based study of logical connectors in EFL students' writing. In R. Pemberton & E. Tsang (Eds.), *Studies in lexis* (pp. 215-246). Language Centre, The Hong Kong University of Science and Technology.

Montrul, S.A. (1998). The L2 acquisition of dative experiencer subjects. *Second Language Research, 14 (1)*, 27-61.

Montrul, S.A. (2001). Introduction to the special issue on "representational and developmental issues in the lexico-syntactic interface: acquiring verb meaning in a second language". *Studies in Second Language Acquisition, 23 (2)*, 145-153.

Moskowitz, G. (1967). The flint system: an observational tool for the foreign language classroom. In A. Simon & E. Boyer (Eds.), *Mirrors*

for behavior: An anthology of classroom observation instruments. Philadelphia: Center for the Study of Teaching at Temple University.

Oshita, H. (1997). *"The Unaccusative Trap": L2 acquisition of English intransitive verbs.* Unpublished doctoral dissertation. University of Southern California.

Perdue, C. (Ed.) (1984). *Second language acquisition by adult immigrants. A field manual.* Rowley, Mass.: Newbury House.

Perdue, C. (Ed.) (1993). *Adult language acquisition: Cross-linguistic perspectives. 2 vols.* Cambridge: Cambridge University Press.

Pienemann, M. (1980). The second language acquisition of immigrant children. In S. Felix (Ed.), *Second language development: Trends and issues.* Tübingen: Gunter Narr.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Erlbaum.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Richards, J. (1971). Error analysis and second language strategies. *Language Sciences, 17,* 12-22.

Rohlen, T.P. (1983). *Japan's high School.* Berkeley: University of California Press.

Sawyer, M. (1996). L1 and L2 sensitivity to semantic constraints on argument structure. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (Eds.), *Proceedings of the 20th Annual Boston University Conference on Language Development, 2,* 646-657.

Sekine, S. (1998). *Corpus based parsing and sublanguage studies.* Unpublished doctoral dissertation. New York University.

Selinker, L., Swain, M., & Dumas, G. (1975). The interlanguage hypothesis extended to children. *Language Learning, 25 (1),* 139-152.

Sinclair, J., & Coulthard, M. (1975). *Towards an analysis of discourse.* Oxford: Oxford University Press.

Slobin, D. (1997). *The crosslinguistic study of language acquisition. Vol.5: Expanding the contexts.* Mahwah, NJ; London: Lawrence Erlbaum.

Tarone, E., Frauenfelder, U., & Selinker, L. (1976). Systematicity/ variability and stability/ instability in interlanguage systems. In H. D. Brown (Ed.), *Papers in second language acquisition* (Language Learning Special Issue No.4), 81-92.

Thepsura, S. (1998). *The Acquisition of lexical causatives by Thai EFL learners.* MS., of MA paper, Georgetown University. (Cited in Juffs, 2000).

Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Cambridge: Cambridge University Press.

Tono, Y. (2000a). 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* (pp. 123-132). Frankfurt: Peter Lang.

Tono, Y. (2000b). A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk & J. P. Melia (Eds.), *PALC'99: Practical applications in language corpora,* (pp. 323-343). Peter Lang GmbH, Frankfurt am Main.

Tono, Y. (2002). *The role of learner corpora in SLA research and foreign language teaching: The multiple comparison approach.* Unpublished doctoral dissertation, Lancaster University.

Tono, Y., & Kanatani, K. (1995). EFL learners' proficiency and roles of feedback: towards the most appropriate feedback for EFL writing. *Annual Review of English Language Education in Japan, 6,* 1-11.

Tono, Y., & Aoki, M. (1998). Developing the optimal learning list of irregular verbs based on the native and learner corpora. In S. Granger & J. Huang (Eds.), *Proceedings of first international symposium on computer learner corpora, second language acquisition and foreign language teaching* (pp. 113-118). Chinese University of Hong Kong.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T., & Izumi, E. (2001). The Standard Speaking Test Corpus: a 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In S. Lee (Ed.), *ASIALEX 2001 Proceedings: Asian bilingualism and the dictionary. The Second Asialex International Congress, August 8-10, 2001* (pp. 257-262). Yonsei University, Korea.

Toth, P. D. (1997). *Linguistic and pedagogical perspectives on acquiring second language morpho-syntax: a look at Spanish se.* Unpublished doctoral dissertation, University of Pittsburgh.

Ungerer, F., & Schmid, H. J. (1996). *An introduction to cognitive linguistics.* Harlow Essex: Addison Wesley Longman.

Uzar, R. (1997). Was PELE a linguist? In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '97 (Practical Applications in Language Corpora).* Lódz, Poland 10-14 April 1997.

Van Lier, L. (1982). *Analysing interaction in second language classrooms.* Unpublished doctoral dissertation, Lancaster University.

White, L. (1987). Markedness and second language acquisition: the question of transfer. *Studies in Second Language Acquisition, 9,* 261-286.

White, L. (1991). Argument structure in second language acquisition. *Journal of French Language Studies, 1,* 189-207.

Yip, V. (1994). Grammatical consciousness-raising and learnability. In T. Odlin (Ed.), *Perspectives on pedagogical grammar.* Cambridge: Cambridge University Press.

Zobl, H. (1989). Canonical typological structures and ergativity in English L2 acquisition. In S. Gass and J. Schachter (Eds.), *Linguistic perspectives on second language acquisition.* Cambridge: Cambridge University Press.