Learner corpora: design, development and applications

Yukio Tono
Graduate School of Applied Linguistics
Meikai University, JAPAN

1. Introduction

Following the 1996 AILA symposium, it was in 1999 when the first international symposium on learner corpora was held in Hong Kong, organized by Sylviane Granger and Joseph Hung. I remember that, with the exception of a few members, including myself, most of the participants were members of the ICLE project, which indicated that the majority of research activities were centred around the ICLE project and that not much else was going on except for some commercial projects such as Longman's. Now, 3 years after the Hong Kong conference, it is a pleasant surprise for the organizers of this pre-conference workshop to see that more than half of the speakers are working for projects other than ICLE, which shows that there has been a growing interest and diversity in this new field among SLA researchers and foreign language teachers.

In this paper, I will give an overview of learner corpus research in terms of the theme of the workshop: its design, development and applications. I will clarify major theoretical and methodological issues involved in learner corpus research in order to facilitate discussions within this relatively new research community. I would also like to review the developments in this research area by introducing some of the major projects and their findings.

2. Design issues of learner corpora

Since people are interested in different aspects of learner language, it is quite natural that the design of learner corpora will vary from project to project, although the tremendous influence of the ICLE project cannot be underestimated here. Following the principles of Contrastive Interlanguage Analysis (CIA), we can use the ICLE corpora to compare native and non-native varieties of the same language, bringing out quantitative differences in frequency in the use of words, word categories, syntactic structures and discourse features (Granger 1998b: 47). Soon we will be able to enjoy the first harvest in the form of an ICLE CD-ROM, which will enable us to empirically verify how many of the ICLE research objectives have been achieved. To my knowledge, this is the first large collection of computerised learner data to be made available for research[1] and we all welcome this valuable addition, considering the paucity of learner corpus data in the past.

Table 1 shows some design considerations for building learner corpora. There are three major categories: (a) language-related criteria (e.g. mode, medium, genre, topic), (b) task-related criteria (e.g. longitudinal vs. cross-sectional; spontaneous vs. prepared), and (c) learner-related criteria (e.g. EFL or ESL, age, sex, mother tongue, overseas experience).

Types of feature		
language-related	task-related	learner-related
mode [written/spoken]	data collection [cross-sectional/longitudinal]	internal-cognitive [age/cognitive style]
genre [letter/diary/fiction/essay]	elicitation [spontaneous/prepared]	internal-affective [motivation/attitude]
style [narration/argumentation]	use of references [dictionary/source text]	L1 background
topic [general/leisure/ etc]	time limitation [fixed/free/homework]	L2 environment [ESL/EFL]/ [level of school]
		L2 proficiency [standard test score]

Table 1: Design considerations for building learner corpora

Whilst the features in Table 1 can be seen in other types of learner data such as generic language

[1] One exception is the Longman Learner's Corpus, which has been commercially available for research.

proficiency tests, learner corpora data will undergo further processing using the techniques of natural language processing. The types of processing available are illustrated in Table 2:

Extra-textual information	Header information (learner/ language/ task variables)
Level of transcription	Orthographic (+ phonemic/ phonetic for spoken corpora)[2]
Level of annotation	Sentence-boundary disambiguation Tokenisation POS tagging Lemmatisation Parsing (Treebanking) Semantic tagging (word senses/ semantic relationships and categories) Discourse tagging (apologies/greetings/politeness/?? moves/acts??/etc.) Error tagging Prosody annotation Anaphoric annotation

Table 2: Processing of learner data

One of the strengths of corpus-based research on learner language is that we can share corpora with other researchers so that findings can be subjected to the careful scrutiny of other researchers. It is in this very respect that most researchers who compile learner corpora for the first time have to be guided more carefully. There are quite a few projects in which not enough attention appears to have been paid to design considerations. If data is gathered in an opportunistic way without proper control and documentation of learner and task variables, the resulting corpus will be unlikely to be of much use.

Granger proposed an approach called Contrastive Interlanguage Analysis (CIA), by which she means that a comparison can be made between native and non-native speakers as well as between learners with different L1 backgrounds in order to identify the features common to all learners and the ones unique to learners with a particular L1 background. This will enable us to distinguish "universal" errors from "L1-specific" errors. This is certainly a very interesting research avenue and one worth investigating, but it raises some methodological questions as well. Selection based upon external criteria such as school year or age does not necessarily ensure that the subjects selected are comparable in terms of language proficiency. This happens to be the case for the Japanese-speaking EFL learner group. Although their learner profile fulfilled all the criteria, their proficiency levels are so markedly lower than those from other European countries that the inclusion of the Japanese data seems to skew the overall results. This is inevitable, considering the learner-external criteria (i.e. second-year university English-majors) set for the project. It would be more appropriate to set an objective internal criterion such as a standard English proficiency test score. However, imposing such a strict criterion will reduce the number of subjects that can be included. This is a dilemma that we need to sort out in order to make our data truly comparable in the future. I would like to propose one solution to this problem. It would be good if we could administer a very simple, short language test for assessing the subjects' proficiency levels. For instance, WHO?? are trying to develop a 5-minute vocabulary levels text, similar to Paul Nation's, but much shorter, in order to see whether such a short quiz can distinguish subjects' proficiency levels in a valid and reliable way. If such a test were made available, it would greatly facilitate a better sampling of the subjects.

Another methodological issue that I would like to mention here is concerned with the standardisation of corpus annotation. Within the corpus linguistics community generally, there is a growing awareness that corpus formatting and annotation should be standardised as much as possible (e.g. TEI, CES/XCES, EAGLES, ATLAS, TUSNELDA, MATE among others). It is of course useful to adopt generic annotation schemes such as above for learner data in general, but there is one area in which learner corpus researchers have yet to agree on a general scheme: "error annotation". As shown in the history of error analysis, categorizing learner errors is a laborious and oftentimes fruitless job, for there are various ways of classifying errors, depending on research interest and theories involved and it is often the case that the classification is only as valid as the theory it is based on. Also, most people have different perspectives on error types, thus leading to very low inter-rater (or classifier) reliability. A generic error tagset, however, stills seems to be very useful goal to work towards, especially if an

[2] In this paper, phonetic transcription of the corpus is largely ignored because of its extra complexity, even though I am aware of such projects such as ISLE (http://nats-www.informatik.uni-hamburg.de/~isle/).

international project such as ICLE can lead the way by producing a large set of learner corpora with standardised error annotations. It will give us an opportunity to survey the error patterns in the corpora in relation to the error analysis scheme and judge which areas we should focus future developments on. Even if it is not possible to use the generic or standard tagset wholesale for one's research purposes, one can still start with such a generic tagset and adapt it or add to it for more problem-oriented research purposes.

3. Development of learner corpora

Table 3 summarises the major learner corpus projects. As this is a rapidly growing field of research, it is increasingly difficult to be kept up-to-date and fully informed of all the various projects around the world, so if there is any project missing from this list, I would encourage you to contact me.

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 2002	- IL – IL (different L1s) - TL – IL	Granger (1993; 1994; 1996; 1998b; 2002)
LINDSEI (Louvain International Database of Spoken English Interlanguage)	- 50 interviews - University EFL 3/4 year students - 100,000 words - CH/IT/FR/JP	- orthographic	- IL – IL (different L1s)	Granger (2001)
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 UAM Corpus	Corpus of teacher and students ' production data.		IL-IL longitudinal	Garcia (in this workshop)
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 (1998)
Cambridge Learners Corpus (CLC)	- All-levels - 10 million	- POS tagged - Error-tagged (2.5 million) - In-house use only	- 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)

Table 3: Learner corpus projects around the world

Keys : IL = interlanguage ; TL = target language ; L1 = first language

Table 3 (continued)

Project	Subjects/ Tasks Size	Annotation Availability	Comparison	References
ASIA:				
JEFLL Corpus (Japan)	- All levels; EFL - Written & spoken - 1 million (expected in 2004)	- POS-tagged - Error-tagged (partial) - Will be available	- IL – IL cross-sectional - L1 – IL - TL – IL	Tono (2000a, b) Tono and Aoki (1998) Tono (2002)
Corpus of English by Japanese Learners	- All levels; EFL - Written - 1 million	- Plain text - Error tagged (partial) - Will be available	- IL – IL cross-sectional	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
Standard Speaking Test (SST) Corpus (also called the TAO Corpus)	- All levels; EFL - Spoken - 1,000,000 - 15 min interview	- Error tagged (partial) - Will be available	- IL-IL (developmental)	Tono et al. (2001)
TELEC Student Corpus	- Hong Kong learners - Univ. exam scripts - 3 million	- Plain text - Restricted availability	- TL – IL	Allan (1998)
Poly U Corpus	- Postgraduates - thesis drafts, etc. - 282,000	- Plain text	- 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	He (1998)
HKUST Corpus of Learner English (HKUST)	- 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 Tsang (1993)

The column "comparison" shows what types of comparison can be made by using the given corpus. Most corpora aim to be comparable with native corpora in order to reveal differences between NS and NNS performance. Some projects compare different stages of ILs (i.e. IL-IL) in order to identify the characteristics of different interlanguage stages. Very few learner corpora incorporate L1 data as an integral part of the design. This will become more important in future learner corpora projects as we are beginning to realise the need to identify specific features of L1-related errors or over/underuse patterns. The quality of TL and L1 corpora is also a critical issue, as over reliance on only one type of data will sometimes skew the picture. We also have to take into account what we should set as the target norm for the L2 learners that we are interested in. Should our students aim for the language proficiencies as represented in the British National Corpus, or should they aim towards something else? Since

comparing frequencies and distributions is an essential part of the corpus-based study of learner language, we should have a clear understanding about the nature of the corpus data we use and how to make valid and meaningful comparisons.

4. Learner corpus analysis

In this section, I will briefly summarise the research areas in which modern-day learner corpora have been exploited. As shown in Table 3, most learner corpus projects were launched in the last decade and the research output based on them is consequently relatively limited. There are, however, a growing number of studies based on learner corpora, and these can show us how researchers are exploiting the new resources. I will summarise four main categories of study: studies related to error analysis (4.1), those investigating quantitative differences between native and non-native language (4.2), those describing the features of the interlanguage in its entirety (4.3), and those applying learner corpora-based research to language teaching methodology and materials design (4.4).

4.1. Studies related to error analysis

In this area, two further subcategories can be made: (1) studies on the development and evaluation of automatic error detection/tagging, and (2) computer learner-corpus-based error analysis.

4.1.1. Studies on error analysis tools and error tagging

Several studies (cf. Bolt 1992; Granger and Meunier 1994) tested the effectiveness of grammar and spelling checkers, demonstrating that while spelling checkers could be used for analysing interlanguage, these automatic tools correct only a very small number of learner errors. Whilst lists of the common errors of EFL learners are available (cf. Turton and Heaton 1997), we have no information on the frequency of these errors. Nor is there any information showing that certain error patterns occur more frequently in one particular learner group compared to others. We also do not know the contexts in which these errors are likely to be made. Without such information, it is impossible to develop either rule-based or probabilistic programs for identifying errors (Milton and Chowdhury 1994).

Due to the lack of precision of currently available grammar checkers, some researchers attempt large-scale manual tagging of all lexical expressions in learner corpora. Since such error tagging is done manually, however, there is always the issue of validity and reliability. As regards validity, error taxonomy is a thorny issue. No matter how general the tagging scheme may be, it should at least include two aspects: (a) *linguistic category classification* (e.g. [grammar] - [verb] - [morpheme] - [tense]) and (b) *target modification taxonomy* (e.g. [omission/ addition/ misformation/etc.] (James 1998). As most error analysis studies in the 70s failed to provide a generic error taxonomy, we should learn a lesson from the past and make the tagging scheme purpose-oriented. Tono (2000b), for instance, replicated the morpheme studies of the 1970s and 80s. This involved manually error-tagging my learner data. I found that it was impossible not to have to develop my own tagging scheme for this particular study. Validity of error tagging should be assessed in the light of the research goals of any particular study.

Reliability is a further issue. As Milton and Chowdhury (1994) commented, accounting for the uncertainty of error type is a serious problem. There are often cases where there is insufficient evidence to assign one unambiguous interpretation of an error. Thus we have to develop tagging schemes which allow for alternative possibilities in terms of target forms. However, while one may try to annotate reasonable alternative possibilities, it is doubtful that any analysis could guarantee total coverage of every possible option (ibid: 129). Granger et al. (1994) recommend that errors should not be normalised, as this involves a high degree of subjectivity, given that many errors can be corrected in many different ways (ibid: 105). This vague status of error correction makes the development of a tagging manual extremely important for the annotator. Granger and her team in ICLE have been developing a Windows-based error editor with an error-tagging manual (Dagneaux, Denness and Granger 1998; Dagneaux, Denness, Granger and Meunier 1996). Their attempt to make manual tagging work easy and consistent is worthwhile. Tono et al. (2001) and Izumi (in this workshop) have developed a generic error tagset and an associated editor. Sharing such tools will better facilitate the standardisation of corpus annotation in the future.

There have been a few other attempts to automate parts of the error tagging process. Mason and Uzar (2000), for example, tested NLP (natural language processing) techniques for detecting zero articles in an interlanguage corpus and demonstrated the possibility of identifying missing articles in learner language. This could lead the way towards automatic error tagging based on POS information. In the same vein, Tono (2000a) also demonstrated the process of semi-automatically annotating learner

data with morpheme tags, using automatically tagged POS information.

4.1.2. Computer learner-corpus-based error analysis

Another research area involving error analysis is the investigation of L2 learners' interlanguage errors using learner corpora. So far the results seem to be still fragmentary in nature, but there is a growing body of research into specific areas of interlanguage errors: for example, collocation (Chi et al. 1994; Lorenz 1997; Granger 1998c; Chen 1998), connectors (Milton and Tsang 1993; Granger and Tyson 1996; Satoh and Fang 1998), irregular past tense (Tono and Aoki 1998), and the English article system (Mason and Uzar 2000).

4.2. Quantitative differences between native and non-native language

There is genuine interest in quantitative differences in the use of certain syntactic, lexical and discoursal features between native and non-native speakers. This is especially true of learner corpora which consist of data from advanced learners, where there is general conformity to native speaker norms in terms of the basic rules of syntax and morphology. "Their deviations from the norm usually concern rather fine points of lexico-grammar and style" (Lorenz 1998:53). Thus the main interest is naturally shifted towards whether they use particular linguistic features more frequently or less frequently than native speakers.

Comparisons are often made between NS and NNS as well as between different NNS groups. The ICLE project members have published articles extensively on this subject. The research topics include adverbial connectors (Altenberg and Tapper 1998; Lorenz 1998), multiword units (De Cock 1998, De Cock et al. 1998), direct questions (Virtanen 1998a), the progressive (Virtanen 1998b), tense morphology (Granger 1999) and phrasal verbs (Lam and Hung 1998).

Biber and Reppen (1998) also conducted an analysis of complement clauses in a large native corpus (the Longman Grammar Corpus) and a small learner corpus (an early version of the Longman Learners' Corpus). While they found quite similar patterns of use for complement clauses, they are quick to caution that since there is also considerable variation among the different tasks required for student writing, it would be necessary to compile learner corpora designed to represent the full range of student writing (and speaking) tasks (ibid: 157) to draw firmer conclusions.

4.3. Description of overall IL development

Although the number is still small, some attempts have been made to exploit learner corpora to describe overall IL characteristics at a fixed stage or at different developmental stages. Granger and Rayson (1998), for example, demonstrated the potential of automatic profiling for revealing the stylistic characteristics of EFL texts vis-à-vis NS texts. They not only produced word frequency profiles for this purpose but also used various measures such as word category profiles (using POS information), which can reveal significant patterns of the over/underuse of major word categories. They concluded that their automatic profiling techniques highlighted the speech-like nature of learner writing (ibid: 129). Leńko-Szymańska (2000a, b) traced vocabulary growth in L2 learners' production by using learner corpora.

Researchers involved in the ICLE project investigated the characteristics of learner language by examining sequences of POS tags (de Haan 1997, 1998; Aarts and Granger 1998). They investigated tag n-grams generated from POS-tagged corpora of three groups of learners (French, Dutch and Finnish). de Haan (1997) found that Finnish students tend to use the combinations involving articles least frequently, which is attributable to the fact that Finnish has no articles and that the use of articles in English is a notoriously difficult topic for Finnish learners (cf. Sajavaara 1981). This coincides with the observations made by Tono (2000a) and Mason and Uzar (2000) that the lack of articles in Japanese and Polish respectively is indeed shown to affect the use of the article system by Japanese- and Polish-speaking learners of English. Aarts and Granger (1998) found a marked similarity in tag sequence frequencies between the three categories of learners (Dutch, Finnish and French), and a striking difference in sentence beginnings: nouns are underused and pronouns overused in sentence-initial sequences in all three learner corpora (ibid: 134). Aarts and Granger also found consistent underuse of the four most common trigrams (sequences of three tags) in NS writing, all of which contain prepositions. Tono (2000b) investigated the tag sequences of developmental spoken interlanguage corpora and found that while L1 child data (taken from the CHILDES database) contain very high proportions of nouns in their top trigrams at the beginning stage, L2 learners start with verb-centred trigrams first instead of nouns. This shows that at least in the case of Japanese EFL learners, beginners produce utterances which contain the basic elements of syntax: subject, verb, and predicate, although the sentences are very short. What is striking in the subsequent developmental

stages is the lack of tense/aspect morphology and the constant underuse of modal verbs as well as prepositional phrases. The Japanese EFL learner data also showed very different tag sequence patterns from the three groups examined in Aarts and Granger (1998), which indicates that the proficiency level of Japanese learners is much lower than the subjects represented in the ICLE Corpus.

Finally, there are a few studies investigating the L2 acquisition of particular linguistic or lexical items. These studies are characterized by the use of learner corpora as a testbed for SLA theories. These studies cover such areas as the acquisition of tense/aspect morphology (Housen 1998; Tono and Aoki 1998), grammatical morphemes (Tono 1998, 2000a), and verb semantics (Oshita 1997). Focusing on L2 acquisition of verb Subcategorisation Frame (SF) patterns, Tono (2002) proposed a rigorous multiple-comparison between interlanguage, L1 and TL corpora in order to identify the relative effects of the four major variables: inherent verb semantics, similarities in SF patterns between L1 and TL, the influence of L2 input (as shown in the frequencies of SF patterns in the textbooks) and developmental effects.

5. Pedagogical applications of learner corpora

Those who are interested in pedagogical applications exploit the results of analyses of learner data to improve various aspects of foreign language teaching. At present, there are still very few studies which relate the findings from learner corpora to actual classroom practice. This is understandable if one takes into account the fact that investigation of interlanguage does not directly lead to better pedagogical practice. It may generate more research questions regarding the way formal instructionis given in a particular EFL context, but it is necessary to conduct follow-up studies to confirm the effect of such methodological changes as suggested by the corpus findings. I should say that we are still not at an advanced stage of development and further research will be needed first to accumulate solid findings from learner corpora.

There are, however, some good examples of pedagogical innovations using learner corpora. For example, Milton (1998) investigated Chinese EFL learners' problematical areas in their writing and designed a CALL program (*AutoWord*) to assist students in essay writing. His research was based upon the 10-million-word HKUST Learner Corpus. The *AutoWord* program contains components such as error recognition (i.e. editing) exercises, a hypertext online grammar, and databases of the 'underused' lexical and grammatical phrases. Milton also developed the wordlist-driven concordancer, WordPilot , in order to integrate concordancing into the L2 essay writing system for novice learners (Milton 2001). Kevin Mark (1998a, b) has developed a very unique parallel corpus database based on his students' writing. His data is firmly rooted in his classroom activities, thus his error classification is practical, which makes the results of his database extremely useful for improving the quality of the activities.

Other areas of applications involve L2 lexicography (Tono 1996, 2001; Gillard and Gadsby 1998), ELT textbook design (Kaszubski 1998), teaching methodology (Granger and Tribble 1998), and developing a learning list of grammar items (Tono and Aoki 1998).

6. Conclusion

I hope that I have been able to show the dynamic nature of this new research area. I will close by making a few statements about desiderata for future research. Firstly, learner corpus researchers should exchange ideas with SLA researchers in a more structured and systematic way. Many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them effectively, while SLA researchers typically know little about what corpora can do for them. By improving the communication lines, we will be able to learn from each other. Secondly, the compilation of a corpus takes time and effort. This means that it will take time to produce useful and useable results. Some people want quick solutions and tend to use cut corners when designing and building corpora, but this will confuse others by producing the results which are not always valid or reproducible. We should enrich the research community with the expertise we have gained from previous projects and should encourage one another not to jump on the bandwagon of corpus-based research without sufficient knowledge of corpus building.

References

Aart, J. and S. Granger, 1998 Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In Granger (ed.) 1998a, pp. 132-141.

Alderson, C. 1996 Do corpora have a role in language assessment? In Thomas, J. and M. Short (eds.) *Using Corpora for Language Research.* London: Longman, pp. 248-259.

Allan, Q. G. 1998 The TELEC Student Corpus: a resource for teacher development. In S. Granger and J. Hung (eds) 1998, pp. 4-6.
Altenberg, B. and M. Tapper 1998 The use of adverbial connectors in advanced Swedish learners' written English. In Granger (ed.) 1998a, pp. 80-93.
Asao, K. 1998 Corpus of English by Japanese learners. In S. Granger and J. Hung (eds) 1998, pp. 10-13.
Biber, D. and R. Reppen 1998 Comparing native and learner perspectives on English grammar: a study of complement clauses. In Granger (ed.) 1998a, pp. 145-158.
Bolt, P. 1992 An evaluation of grammar-checking programs as self-helping learning aids for learners of English as a foreign language. *CALL* 5: 49-91.
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.
Chi, Amy, K. Wong Pui-yui and M.W. Chau-ping (1994) Collocational problems amongst ESL learners: a corpus-based study. In Flowerdew, L. and A. K.K. Tong (eds.) *Entering Text*. Language Centre. The Hong Kong University of Science and Technology.
Dagneaux, E., S. Denness, S. Granger, and F. Meunier 1996 *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve.
Dagneaux, E., S. Denness, and S. Granger, 1998 Computer-aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics* 26(2): 163-174.
David, M. 1998 The pedagogical implication of a learner corpus. In S. Granger and J. Hung (eds) 1998, pp. 87-88.
De Cock, S. 1998 A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English, *International Journal of Corpus Linguistics* 3(1): 59-80.
De Cock, S., S. Granger, G. Leech and T. McEnery 1998 An automated approach to the phrasicon of EFL learners. In Granger (ed.) 1998a, pp. 67-79.
Dodd, B. 1997 Exploiting a corpus of written German for advanced language learning. In Wichmann et al. 1997, pp. 131-145.
Farmer, R. and K. Mead 1998 The language of citations: an analysis via computer learner corpus. In S. Granger and J. Hung (eds) 1998, pp. 34-37.
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.
Flowerdew, L. 1998 Concordancing on an expert and learner corpus in ESP. *CÆLL Journal* 8 (3): 3-7.
Foster-Cohen, S.H. 1999 SLA and First Language Acquisition. *Annual Review of Applied Linguistics* 19: 3-21.
French, F. 1949 *Common Errors in English*. London: Oxford University Press.
Gillard, P. and A. Gadsby 1998 Using a learners' corpus in compiling ELT dictionaries. In Granger (ed.) 1998a, pp. 159-171.
Granger, S. 1993 The International Corpus of Learner English. In Aarts, J., P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, pp. 57-69.
Granger, S. 1994 The learner corpus: a revolution in applied linguistics. English Today 39 (10/3): 25-9.
Granger, S. 1996 From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Aijmer, K., B. Altenberg and M. Johansson (eds.) *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*, Lund 4-5 March 1994, Lund: Lund University Press, pp. 37-51..
Granger, S. (ed.) 1998a *Learner English on Computer*. London: Addison Wesley Longman.
Granger, S. 1998b A bird's eye view of computer learner corpus research. In S. Granger and J. Hung (eds) 1998, pp. 45-48.
Granger, S. 1998c Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (ed.) *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, pp. 145-160.
Granger, S. 1999 Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus. In Hasselgard, H. and S. Oksefjell (eds) *Out of Corpora - Studies in Honour of Stig Johansson*. Amsterdam: Rodopi, pp. 191-202.
Granger, S. 2002 A Bird's-eye View of Computer Learner Corpus Research. In Granger, S., Hung, J. and Petch-Tyson, S. (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam and Philadelphia: Benjamins.

Granger, S. and F. Meunier 1994 Towards a grammar checker for learners of English. In Fries, U. and G. Tottie (eds.) *Creating and Using English Language Corpora*. Amsterdam and Atlanta: Rodopi, pp.79-89.
Granger, S. and P. Rayson 1998 Automatic Lexical Profiling of Learner Texts. In Granger (ed.) 1998a, pp. 119-131.
Granger, S. and C. Tribble 1998 Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In Granger (ed.) 1998a, pp. 199-209.
Granger, S. and S. Tyson 1996 Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15: 19-29.
Granger, S. and J. Hung (eds.) 1998 *First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 14-16 December, 1998, The Chinese University of Hong Kong: Symposium Proceedings.
de Haan, P. 1997 An experiment in English learner data analysis. In Aarts, J., de Mönnink, I. and Wekker, H. (eds) *Studies in English Language and Teaching*. Amsterdam: Rodopi, pp. 215-229.
de Haan, P. 1998 How native-like are advanced learners of English? In Renouf, A. (ed.) *Explorations in Corpus Linguistics*, Amsterdam: Rodopi, pp. 55-65.
de Haan, P. 1999 English writing by Dutch-speaking students. In Hasselgård, H. and Oksefjell, S. (eds) *Out of Corpora*, Amsterdam: Rodopi, pp. 203-212.
de Haan, P. (forthcoming) Tagging non-native English with the TOSCA-ICLE tagger. In Mair, C. (ed.) *Proceedings of the 20th ICAME Conference*, Freiburg 1999. Amsterdam: Rodopi.
He, A. 1998 A corpus-based analysis of middle school students' English spelling errors. In Garnger and Hung (eds.), pp. 54-58.
Housen, A. 1998 An analysis of grammatical form-function mapping in L2 data using the CHILDES system. In S. Granger and J. Hung (eds.) 1998, pp. 59-62.
James, C. 1980 *Contrastive Analysis*. London: Longman.
James, C. 1998 *Errors in Language Learning and Use*. London: Addison Wesley Longman.
Johns, T. 1993 Data-driven learning: an update. *TELL & CALL* 3.
Johns, T. 1997 Contexts: the background, development and trialling of a concordance-based CALL program. In Wichmann et al. 1997, pp. 100-115.
József, H. 1998 *Advanced Writing in English as a Foreign Language: A Corpus-based Study of Processes and Products*. Unpublished PhD dissertation. Janus Pannonius University, Pécs, Hungary.
Kaszubski, P. 1998 Enhancing a writing textbook: a national perspective. In Granger, S. (ed.) 1998a pp. 172-185.
Lam, P and J.Hung 1998 The use of multi-word verbs in advanced Chinese ESL learners. In S. Granger and J. Hung (eds) 1998, pp. 80-82.
Leńko-Szymańska, A. 2000a Passive and active vocabulary knowledge in advanced learners of English. In Lewandowska-Tomaszczyk, B. and J.P. Melia (eds), pp. 287-302.
Leńko-Szymańska, A. 2000b How to trace the growth in learner's active vocabulary. A corpus-based study. Paper presented at the 4th International Conference on Teaching and Language Corpora. Graz, 19-23 July 2000.
Lewandowska-Tomaszczyk, B., T. Leńko-Szymańska, and A. McEnery 2000 Lexical problem areas in the PELCRA learner corpus of English. In Lewandowska- Tomaszczyk, B. and J.P.Melia (eds.) 2000, pp. 303-312.
Lewandowska-Tomaszczyk, B. and J.P. Melia (eds.) 2000 *PALC' 99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang.
Lorenz, G. 1997 *Introducing a Learner Corpus of English Writing: The Hidden Potentials of Adjective Intensification*. Amsterdam: Rodopi.
Mark, K. 1998a A parallel learner corpus approach to English curriculum development at a Japanese university. In S. Granger and J. Hung (eds) 1998, pp. 89-90.
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. and R. Uzar 2000 NLP meets TEFL: Tracing the zero article. In Lewandowska-Tomaszczyk, B. and J.P. Melia (eds.) 2000, pp. 105-116.
Meunier, F. 1995 Tagging and parsing interlanguage. In Beheydt, L. (ed.) La Linguistique Appliquée dans les années 90. *ABLA Review* 16, 21-29.
Meunier, F. 1998 Computer tools for the analysis of learner corpora. In Granger (ed.) 1998a, pp. 19-37.
Milton, J. 1998 WORDPILOT: enabling learners to navigate lexical universes. In S. Granger and J. Hung (eds) 1998: 97-98.
Milton, J. 2001 *Describing and overcoming environmental limitations on the interlanguage of Hong*

Kong Chinese learners of English: a computational and corpus-based methodology. Unpublished PhD thesis. Lancaster University.

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

Milton, J. and N. Chowdhury 1994 Tagging the interlanguage of Chinese learners of English. In Flowerdew, L. and A. K. K. Tong (eds.) *Entering Text*. Language Centre, The Hong Kong University of Science and Technology, pp. 127-143.

Milton, J. and R. Freeman 1996 Lexical variation in the writing of Chinese learners of English. In C.E. Percy, C.F. Meyer and I. Lancashire (eds.) *Synchronic Corpus Linguistics*. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora, Amsterdam: Rodopi, pp. 121-131.

Milton, J. and K. Hyland 1999 Assertions in students' academic essays: a comparison of L1 and L2 writers. In R. Berry, B. Asker, K. Hyland and M. Lam (eds.) *Language Analysis, Description and Pedagogy*. Hong Kong: HKUST, pp. 147-161.

Oshita, H. 1997 *"The unaccusative trap": L2 acquisition of English intransitive verbs*. Unpublished PhD thesis. University of Southern California.

Sajavaara, K 1981 The nature of first language transfer: English as L2 in a foreign language setting. Paper presented at the first European-North American Workshop on Cross-Linguistics Second Language Acquisition Research, Lake Arrowhead, California.

Satoh, K. and C.A. Fang 1998 A corpus-based study of the grammar and lexis of Japanese learners' English. In S. Granger and J. Hung (eds) 1998, pp. 103-104.

Tono, Y. 1996 Using learner corpora for L2 lexicography. *LEXIKOS* 6: 116-132. Stellenbosch: Universiteit van Stellenbosch.

Tono, Y. 1998 A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *TALC (Teaching and Language Corpora) 98 – Conference Proceedings*, Keble College Oxford, 24-27 July 1998, pp. 183-187.

Tono, Y. 2000a A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In Burnard, L. and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, pp. 123-132.

Tono, Y. 2000b A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In Lewandowska-Tomaszczyk, B. and J.P.Melia 2000, pp. 323-343.

Tono, Y. 2001 *Research on Dictionary Use in the Context of Foreign Language Learning*. Tübingen: Max Niemeyer Verlag.

Tono, Y. 2002 *The Role of Learner Corpora in SLA Research and Foreign Language Learning: The Multiple Comparison Approach*. Unpublished PhD thesis. Lancaster University.

Tono, Y. and K. Kanatani 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. and M. Aoki 1998 Developing the optimal learning list of irregular verbs based on the native and learner corpora. In S. Granger and J. Hung (eds) 1998, pp. 113-118.

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

Tribble, C. 1997 Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In Lewandowska-Tomaszczyk, B. & P. J. Melia, (eds.) *PALC '97* (Practical Applications in Language Corpora), pp. 106-117. Frankfurt: Peter Lang.

Turton, N.D. and J.B. Heaton 1997 *Longman Dictionary of Common Errors*. Harlow: Longman.

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

Virtanen, T. 1998a Direct questions in argumentative student writing. In Granger (ed.) 1998a, pp.94-118.

Virtanen, T. 1998b Argumentative uses of the progressive in NS and NNS student compositions: notes on clause status and grounding. In S. Granger and J. Hung (eds) 1998, pp. 119-120.