Computer Learner Corpus Research: Current Status and Future Prospects

Sylviane Granger

University of Louvain, Belgium

Abstract

Despite a mere decade of existence, the field of computer learner corpus (CLC) research has been the focus of so much active international work that it seems worth taking a retrospective look at the research accomplished to date and considering the prospects for future research in both Second Language Acquisition (SLA) studies and Foreign Language Teaching (FLT) that emerge. One of the main distinguishing features of computer learner corpora – and indeed one of their main strengths – is that they can be used by specialists from both these fields and thus constitute a possible point of contact between them. The first three sections of this chapter are devoted to a brief overview of the main aspects of CLC research: data collection, methodological approaches, learner corpus typology, and size and representativeness. Sections 4 and 5 review the tangible results of CLC research in the fields of SLA and FLT.

1 Introduction

The relative youth of computer learner corpus (CLC) research as a field of scientific enquiry (it burgeoned as a discipline as recently as the late 1980s) renders a definitive assessment of its achievements somewhat premature. However, enough work has been done to take stock of advances made in the field and to evaluate its future prospects. The main objective of this article is to assess whether, in making Leech's (1992: 106) description of corpus linguistics our own, we would be justified in calling CLC research "a new research enterprise, a new way of thinking about ***learner*** language, which is challenging some of our most-deeply rooted ideas about ***learner*** language." After highlighting some of the main features that distinguish CLC data from other types of learner data, I will take stock of the current situation in terms of corpus collection and analysis and give an overview of the current results and future prospects in two distinct but closely related fields: Second Language Acquisition (SLA) and Foreign Language Teaching (FLT).

2 Distinguishing Features of CLC Data

There is nothing new in the idea of collecting learner data. Both FLT and SLA researchers have been collecting learner output for descriptive and/or theory-building purposes since the disciplines emerged. In view of this, it is justified to

ask what added value, if any, can be gained from using learner corpus data. Computer learner corpora typically fall into the category of natural or "open-ended" language use data, a data type which has not tended to be favoured in recent SLA research. There are many reasons why SLA researchers have tended to prefer other types of notably experimental and introspective data. The intention here however is not to expand on these (for a brief overview, see Granger 1998b: 4-6) and compare the respective values of natural and elicited data types, but instead to highlight three features which give CLC data a definite advantage over previously used natural use data, in the hope of reinstating this neglected data type.

2.1 Size

Computer learner corpora are electronic collections of spoken or written texts produced by foreign or second language learners. As the data is stored electronically, it is possible to collect a large amount of it fairly quickly. As a result, learner corpora are now counted in the millions rather than in the hundreds or thousands of words. But is big beautiful in SLA/FLT terms? The answer to this question is more of a "yes on the whole" or a "yes but" than an unqualified "yes."

Many SLA researchers have highlighted the drawback of using a very narrow empirical base. In reference to longitudinal SLA studies, which usually involve a highly limited number of subjects, Gass and Selinker (2001: 31) note that "It is difficult to know with any degree of certainty whether the results obtained are applicable only to the one or two learners studied, or whether they are indeed characteristic of a wide range of subjects." It is the same kind of dissatisfaction and mistrust that led MacWhinney (2000: 3) to build the CHILDES child language acquisition database:

> Conducting an analysis on a small and unrepresentative sample may lead to incorrect conclusions. Because child language data are so time-consuming to collect and to process, many researchers may actually avoid using empirical data to test their theoretical predictions. Or they may try to find one or two sentences that illustrate their ideas, without considering the extent to which their predictions are important for the whole of the child's language. In the case of studies of pronoun omission, early claims based on the use of a few examples were reversed when researchers took a broader look at larger quantities of transcript data.

Like child language data, L2 data is difficult to collect. While the practice of getting students to submit their homework electronically has become standard in some countries, in others this is still a very remote prospect. In any case, some types of text, for instance those produced as part of an exam or as a classroom exercise, still tend to be handwritten. The difficulty is compounded in the case of

spoken data. In the absence of reliable automatic speech recognition software, collecting and transcribing oral data remains a highly time-consuming activity. In addition, any data that has been keyed in manually or scanned needs to go through a process of careful proofreading to ensure that the original learner text is faithfully transcribed with no new errors introduced and all the original ones kept. This being said, there is no doubt that the widespread use of word processors, electronic mail and web-based learning environments will speed up learner corpus collection. Indeed some of the most recent learner corpora have been collected fully automatically (see Wible et al. 2001).

Whether collected electronically over a very short period of time or after years of painstaking work, current learner corpora tend to be rather large, which is a major asset in terms of representativeness of the data and generalizability of the results. Of course, a very large data sample is not necessary for all types of SLA research. A detailed longitudinal study of one single learner is of great value if the focus is on individual interlanguage development. Likewise in FLT, as pointed out by Ragan (1996: 211), small corpora compiled by teachers of their own pupils' work are of considerable value: "the size of the sample is less important than the preparation and tailoring of the language product and its subsequent corpus application to draw attention to an individual or group profile of learner language use." In addition, as we will see in the following section, size is only really useful if the corpus has been collected on the basis of strict design criteria.

2.2 Variability

Learner language is highly variable. It is influenced by a wide variety of linguistic, situational and psycholinguistic factors, and failure to control these factors greatly limits the reliability of findings in learner language research. The strict design criteria which should govern all corpus building make corpora a potentially very attractive type of resource for SLA research. As rightly pointed out by Cobb (2003: 396), "It is a common misconception that corpus building means collecting lots of texts from the Internet and pasting them all together." Atkins et al. (1992) list 29 variables to be considered by corpus builders. While many of these variables are also relevant for learner corpus building, the specific nature of learner language calls for the incorporation of L2-specific variables. Figure 1 represents all the variables that are controlled for and recorded in one particular CLC, the *International Corpus of Learner English (ICLE)* database. In addition to some general dialectal and diatypic variables, which are also used in native corpus building, the *ICLE* database contains a series of L2-specific variables, pertaining to the learner or the task. A search interface enables researchers to select data on the basis of these criteria (for more information on the *ICLE*, see Granger 2003a; Granger et al. 2002). This degree of control distinguishes CLC data from the samples of language use that are commonly used in SLA research. In his critique of EA (Error Analysis) studies, Ellis (1994: 49)

lists some of the factors that can bring about variation in learner output and notes that "unfortunately, many EA studies have not paid sufficient attention to these factors, with the result that they are difficult to interpret and almost impossible to replicate." Gass and Selinker (2001: 33) make a similar comment in relation to cross-sectional SLA studies: "there is often no detailed information about the learners themselves and the linguistic environment in which production was elicited."

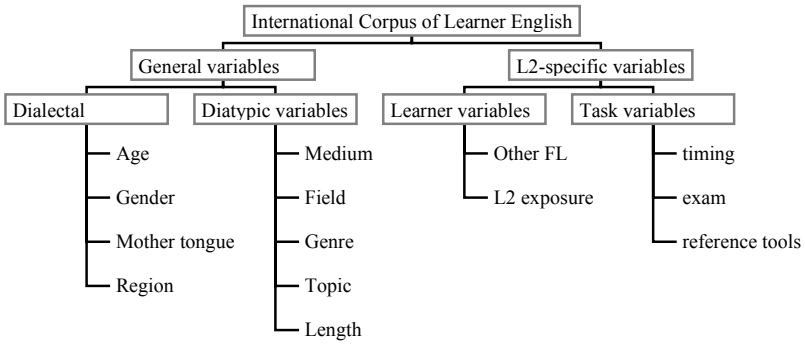

Figure 1*: ICLE* general and L2-specific variables

It would be wrong, however, to paint too rosy a picture of current CLC. In all fairness, one must admit that (a) there are not many tightly-designed learner corpora in the public domain, and (b) there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control. Ideally, as stated by Biber (1993: 256), "theoretical research should always precede the initial design and general compilation of texts." This preliminary theoretical analysis is the only way to ensure that the corpus will contain all the relevant design parameters.

2.3 Automation

So far, research on learner language has been largely manual. The ground covered in SLA and FLT research over the last decades shows that major advances can be made in the field without having recourse to computers. However, the benefit that researchers can derive from automating some of their work is so great that it would seem a pity to do without the invaluable help it can provide. While with small language samples the gain in terms of time and effort may not seem large enough to compensate for the investment necessary to become familiar with automated methods and tools, using big corpora makes it absolutely essential to use automated approaches. In the following, I will focus on four functions –

COUNT, SORT, COMPARE and ANNOTATE – which lend themselves particularly well to automation, and highlight their relevance for SLA/FLT research.

2.3.1 COUNT

This function involves a series of options, from the crude to the highly sophisticated, all of which are potentially very useful for interlanguage studies. The crudest function of all, counting the number of words in a text, is essential if one is to compare the frequency of linguistic items in various texts. To effect this type of comparison, researchers working on the basis of non-electronic texts have no other option but to count the average number of words per page and multiply the resulting figure by the number of pages in the text to obtain a rough estimate. If the data is computerised, the researcher can obtain the precise figure using the word count option on his/her word processor. More sophisticated options, provided by text handling packages, such as *WordSmith Tools* (Scott 1996), provide researchers with word frequency lists sorted in alphabetical or frequency order, type/token ratios and a series of other statistical measures (number of paragraphs, average number of words per sentence, etc.). Frequency lists of two or more word combinations are of great value to the growing number of SLA/FLT researchers interested in phraseological/routine aspects of interlanguage. In addition, all annotations inserted in the corpus (e.g., errors, grammatical categories, lemmas) can be counted and the frequencies compared across individual learners or learner populations.

2.3.2 SORT

One of the simplest but at the same time most rewarding benefits of electronic data is the multitude of possibilities offered in terms of sorting facilities. Concordancing programs give SLA/FLT researchers an unparalleled view of learners' lexico-grammatical patterning of words (i.e. their use or misuse, or over-/underuse) of collocations, colligations and other (semi-)prefabricated phrases. In addition, more sophisticated programs such as *WordSmith Tools* combine the COUNT and SORT facilities and provide a collocate display, which provides the exact frequency of all words occurring within a particular window on either side of the headword.

2.3.3 COMPARE

Interlanguage is a variety in its own right, which can be studied as such without comparing it to any other variety. However, for many purposes, both theoretical and applied, it is useful to compare it to other language varieties to bring out its specificities. This contrastive approach, which is usually referred to in CLC-based research as *Contrastive Interlanguage Analysis*, may involve two types of

comparison: a comparison of native language and learner language (L1 vs. L2) and a comparison of different varieties of interlanguage (L2 vs. L2). The "compare list" facility in *WordSmith Tools* makes it possible to automate these comparisons: it compares frequency lists from two corpora and brings out the words or phrases that are significantly over- or underused in either corpus (for illustrations, see section 4).

2.3.4 ANNOTATE

Garside et al. (1997: 2) define corpus annotation as "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written data." While a raw learner corpus is in itself a highly useful resource, it does not take long for the SLA/FLT researcher to realise that it would be even more useful if it contained an extra layer of information, which could also be counted, sorted and compared. To this end, researchers can either use off-the-shelf annotating tools or develop their own. For obvious reasons, researchers tend to prefer ready-made tools. A number are available, some free of charge (for a survey, see Meunier 1998). However, it is important to bear in mind that all these programs – whether lemmatizers, part-of-speech (POS) taggers, or parsers – have been trained on the basis of native speaker corpora, and there is no guarantee that they will perform as accurately when confronted with learner data. While the success rate of POS-taggers has been found to be quite good with advanced learner data (Meunier 1998: 21), it has proved to be very sensitive to morpho-syntactic and orthographic errors (Van Rooy and Schäfer 2003) and success rate will therefore tend to decrease as the number of these errors increases. Pilot studies aimed at testing the reliability of the annotation, and recommended whatever the type of corpus used, are therefore a must with learner corpora[1]. Similarly, while lemmatizers are potentially very useful for lexical analyses of interlanguage, researchers have to be aware that only the standard realisations of the lemma will be retrieved (i.e. for the lemma LOSE) the standard forms *lose/loses/losing/lost*, but not the (sometimes equally frequent!) non-standard forms *loose/looses/loosing/loosed*. If proved reliable, a POS-tagged learner corpus is a very powerful resource, allowing for detailed studies of the use of grammatical categories, such as prepositions, phrasal verbs, modals, passives, etc. Note, however, that the search and retrieval possibilities depend on the granularity of the tagset, which is extremely variable (from 50 up to 250 tags).

POS-taggers and lemmatizers have undeniable advantages, not least of which is the fact that they are fully automatic, but there are other types of annotation that SLA/FLT researchers may want to add to the text for which no ready-made program exists. This type of tagging, which de Haan (1984) calls "problem-oriented tagging," can be inserted with the help of editing tools to speed up the process. Any type of annotation is potentially useful (discourse annotation, semantic annotation, refined syntactic annotation, etc.), but one type, error annotation, is particularly relevant for interlanguage studies and is enjoying

growing popularity among CLC researchers. While I would not go as far as Wible et al. (2001: 311) who consider that unannotated learner corpora are "in themselves (...) worth little to teachers and researchers," I fully agree that error annotation is a major added value, especially if the corpus is compiled for FLT purposes. Several systems of annotation have been developed (Milton and Chowdhury 1994; Dagneaux et al. 1998; Nicholls 2003) and have been exploited in a series of innovative FLT applications.

These three main distinguishing features clearly differentiate computer learner corpora from the language use data types that have traditionally been used in SLA and FLT research. It should be borne in mind, however, that each type of investigation calls for its own data collection methods and, as a result, learner corpora should not be seen as a panacea, but rather as one highly versatile resource which SLA/FLT researchers can usefully add to their battery of data types.

3 Learner Corpus Collection and Analysis

This section aims to assess the current state of CLC research in terms of (1) corpus collection: What learner corpora have been compiled to date? What are their main characteristics? Are there gaps that would need to be filled? And (2) corpus analysis: What types of analysis have been carried out? What methodological approaches have been adopted? I will focus exclusively on English not only for reasons of space but also because this is where a majority of the research has been carried out to date. It should be noted, however, that the CLC movement has recently gained new momentum and CLC projects on languages other than English are mushrooming in all parts of the world. The recent launch of a "multilingual learner corpus" project, which will contain data in several L2s[2] (Tagnin 2003), is but one significant example of this trend.

3.1 Corpus Collection

Rather than duplicating Pravec's (2002) excellent survey, which gives a wealth of information (size, availability, learner background information, etc.) on the best-known written learner corpora, I will adopt a more general outlook. By situating current CLC along a series of dimensions, I hope to be able to bring out some of the main characteristics of current CLC and hence to make suggestions for future data collection.

Computer learner corpora fall into two major categories: commercial CLCs, which are initiated by major publishing companies, and academic CLCs, which are compiled in educational settings.[3] The two major commercial learner corpora, the *Longman Learners' Corpus* and the *Cambridge Learner Corpus,* are both very big (10 million words for the Longman corpus and 16 million for the Cambridge corpus). The academic corpora, far more numerous, are extremely variable in size (the *Hong Kong University of Science and Technology Learner*

Corpus contains 25 million words while the *Montclair Electronic Language Database* only contains 100,000 words). In addition to the 8 academic corpora listed by Pravec (2002), a myriad of other corpora have been or are being collected and exploited by individual researchers and/or teachers. The paradox we face is that while there is an abundance of learner corpora, hardly any of it is available for academic research. It is to be hoped that the recently published first version of the *International Corpus of Learner English* (Granger et al. 2002), comprising 2.5 million words of EFL writing, will be the first of many CLCs to become publicly available.

Current CLC can be classified along two major dimensions relating to characteristics of the learners who have produced the data and characteristics of the tasks they were requested to perform.

3.1.1 Learners

The learners represented in current CLC corpora are overwhelmingly learners of English as a Foreign Language (EFL) rather than as a Second Language (ESL). The line between the two categories is undoubtedly a fine one, but if ESL is broadly defined as taking place "with considerable access to speakers of the language being learned, whereas learning in a foreign language environment does not" (Gass and Selinker 2001: 5), it is quite clear that the latter dominates the current CLC scene. Regarding L1 background, there is a clear difference between commercial corpora, which tend to have multi-L1 coverage, and academic corpora which tend to cover learners from only one mother tongue background, the *ICLE* database being a notable exception in this respect. The learners' proficiency predominantly falls in the intermediate-advanced range. This somewhat vague description reflects the well-known fact that "one researcher's advanced category may correspond to another's intermediate category" (Gass and Selinker ibid: 37). The fuzziness is compounded by the fact that compilers, following established corpus design practices (see Atkins et al. 1992: 5), have tended to use external criteria to compile their corpus. As regards proficiency, this comes down to favouring the criterion of "institutional status" (for instance, third year English undergraduates) over other criteria such as impressionistic judgements, specific research-designed test or standardised tests (Thomas 1994).

3.1.2 Task

As regards medium, the number of written learner corpora by far exceeds the number of spoken learner corpora. Far from being restricted to learner corpora, the difficulty of collecting and transcribing spoken data also affects native corpus building, as evidenced by the limited proportion of speech in recent mega-corpora of English (the BNC has 10% spoken vs. 90% written data). However, in the case of spoken learner language, the difficulty is multiplied by a factor of 10 and the time involved in collecting and transcribing data is so prohibitive that

collaborative projects such as the LINDSEI[4] project, would seem to be the only realistic course to take. As regards the field of discourse, the language covered by learner corpora is predominantly English for General Purposes (EGP) rather than English for Specific Purposes (ESP). For writing, English for Academic Purposes (EAP), which can be seen as situated between EGP and ESP, gets the lion's share because of its importance in the EFL context.

Another dimension along which CLC can be classified is the longitudinal vs. cross-sectional dimension. The overwhelming majority of CLC covering more than one type of interlanguage data are cross-sectional (i.e. they contain data gathered from different categories of learners at a single point in time). Genuine longitudinal corpora, where data from the same learners are collected over time, are very few and far between. For this reason, researchers interested in interlanguage development tend to collect quasi-longitudinal corpora (i.e. corpora gathered at a single point in time but from learners of different proficiency levels). Though easier to collect than "real" longitudinal corpora, this type of corpus is nevertheless still relatively infrequent.

Learner corpora also differ in their degree of processing. While most current learner corpora consist of raw data (i.e. they contain the learner texts with no added annotation), there are several projects based on POS-tagged corpora. At the same time, the number of error-tagged learner corpora is clearly on the increase.

This very brief overview shows that the language data contained in current CLC falls short of covering the wide diversity that characterises learner language. A lot of work remains to be done, not only to compile CLC representing hitherto neglected data types, but also to make the numerous CLC that have been compiled – either commercially or academically – available to the scientific community. One new promising development gives cause for optimism. Synchronous corpus building projects, in which corpora are collected online while the students carry out a pedagogical task (see section 5 below), solve many of the difficulties that beset standard asynchronous CLC building and will hopefully contribute to faster corpus building and dissemination.

3.2 Corpus Analysis

For a field that is little over ten years old, CLC has already generated a very rich and diversified body of research. The learner corpus bibliography stored on the Louvain website[5] contains over 150 publications and is a good starting point for any researcher wishing to embark on learner corpus analysis. In this section, I will restrict myself to highlighting some of the areas in which research has been particularly active, distinguishing between the following three broad categories: methodological and analytic framework, contrastive interlanguage analysis (CIA) and computer-aided error analysis (CEA).

3.2.1 Methodological and Analytical Framework

Like any new discipline, computer learner corpus research has had to avail itself of a sound framework of analysis. To this end, it has been able to rely to some extent on the methodological and analytic apparatus developed in the field of corpus linguistics (CL). There are however special considerations with learner corpora, given the type of language data involved, and the reasons for collecting them differ from other corpus endeavours, specifically because of their relevance to language learning theory and practice. The CL apparatus has therefore had to be tailored for the specific needs of CLC research and several publications have contributed to this. Leech (1998) and Granger (1998, 2002) contain wide-ranging discussions of particular methodological and analytical considerations relating to CLC, including methods of analysis such as CIA and CEA. Meunier (1998) deals more specifically with the software tools that can be used in CLC research, Van Rooy and Schäfer (2003) look into the reliability of POS-tagging of CLC data and de Mönnink (2000) examines the feasibility of parsing CLC. Other descriptions of the CIA methodology can be found in Granger (1996) and Gilquin (2001), while the principles of CEA are presented in Milton and Chowdhury (1994), Dagneaux et al. (1998), de Haan (2000) and Nicholls (2003). In addition, highly valuable methodological guidelines and warnings are contained in the many CLC case studies that have appeared to date.

3.2.2 CIA studies

The bulk of CLC research so far has been of the CIA type. There has been a wide range of topics, but some fields have received a great deal of attention, in particular high frequency vocabulary (Ringbom 1998, 1999; Källkvist 1999; Altenberg 2002), modals (Aijmer 2002; McEnery and Kifle 2002; Neff et al. in press), connectors (Milton and Tsang 1993; Field 1993; Granger and Tyson 1996; Altenberg and Tapper 1998; L. Flowerdew 1998b), collocations and prefabs (Chi Man-Lai et al. 1994; De Cock 1998, 2000; De Cock et al. 1998; Howarth 1996; Granger 1998; Nesselhauf 2003). Most of the CIA studies are based on unannotated learner corpora. A few, however, make use of POS-tagged corpora and compare the frequency of grammatical categories or sequences of grammatical categories in native and learner corpora (Aarts and Granger 1998; Granger and Rayson 1998; de Haan 1999; Tono 2000). All these studies bring out the words, phrases, grammatical items or syntactic structures that are either over- or underused by learners and therefore contribute to the foreign-soundingness of advanced interlanguage even in the absence of downright errors. It is important to understand at this point that this CIA approach would draw fire from some SLA theorists for its failure to study interlanguage (IL) in its own right but rather as an incomplete version of the target language (TL). This practice, which Bley-Vroman (1983) refers to as the "comparative fallacy," is discussed as follows by Larsen-Freeman and Long (1991: 66): "researchers should not adopt a normative

TL perspective, but rather seek to discover how an IL structure which appears to be non-standard is being used meaningfully by a learner." In her recent excellent book on *Corpora in Applied Linguistics*, Hunston (2002: 211-2) expresses a similar view when she writes that one of the drawbacks of the CIA approach is that "it assumes that learners have native speaker norms as a target." However, she adds that the CLC approach also has two advantages: first, the standard is clearly identified and if felt to be inappropriate can be changed and replaced by another standard; and second, the standard is realistic: it is "what native/expert speakers actually do rather than what reference books say they do." In addition, it is important to bear in mind that most CLC research so far has involved advanced EFL learners (i.e. learners who are getting close to the end point of the interlanguage continuum and who are keen to get even closer to the NS norm). For this category of learners more than any other, it makes sense to try and identify the areas in which learners still differ from native speakers and which therefore necessitate further teaching.

3.2.3 CEA studies

CEA has led to a much more limited number of publications than CIA. Apart from articles describing error tagging systems (see above), there are a few articles focusing on certain specific error categories (lexical errors: Chi Man-lai et al. 1994; Källkvist 1995; Lenko-Szymanska 2003; tense errors: Granger 1999). In view of the investment of time necessary to error tag corpora and analyse the results, it is not surprising that CEA studies should to some extent be lagging behind. However, it should be borne in mind that in CLC research, errors are not isolated from the texts in which they originated, as was the case in traditional EA studies, but rather are studied in context alongside cases of correct use and over- and underuse. Discussions of errors can therefore be found in a large number of CLC case studies.

This brief overview gives a glimpse of the buzz of activity in the CLC field, but at the same time it leaves a certain impression of patchiness. This may well be due to the corpus linguistic bottom-up approach which, as stated by Swales (2002: 152) "involves working from small-stretch surface forms and then trying to fit them into some larger contextual frame," a method which produces a "huge amount of trial-and-error." It is important to bear in mind, however, that what can be presented as a down side of the corpus linguistic approach is also its major strength: it is the required passage to gain *new* insights into language. This being said, one must acknowledge that the wider perspective is often difficult to discern from current CLC studies. In the coming sections, I will therefore try to highlight the wider SLA (section 4) and FLT (section 5) implications of CLC research.

4 Computer Learner Corpora and SLA

To what extent can CLC contribute to SLA research? Second Language Acquisition is the study of how second languages are learned. It involves questions such as "Are the rules like those of the native language? Are they like the rules of the language being learned? Are there patterns that are common to all learners regardless of the native language and regardless of the language being learned? Do the rules created by second language learners vary according to the context of use?" (Gass and Selinker 2001: 1). CLC data can contribute to answering these questions. The use of bilingual corpora in addition to learner corpora can help answer the first question. Researchers can only say for sure if the learner's rules "are like those of the native language" if they have detailed descriptions of the learner's native language compared with the target language. This integrated contrastive perspective, which combines classic CA (Contrastive Analysis) and CIA, is a very reliable empirical platform from which to conduct interlanguage research (for illustrations of the method, see Gilquin 2001; Altenberg 2002). The following questions involve the two types of comparison that are at the heart of the CIA methodology: comparisons of native and learner data and comparisons of different interlanguages to each other. As to the last question, recourse to strictly controlled learner corpora is a good way of identifying the impact of different "contexts of use." In fact, richly documented corpora such as the *ICLE* allow researchers to carry out cross-sectional research without having to cope with the major disadvantage that is usually presented as being part and parcel of this type of study: "The disadvantage [of cross-sectional studies] is that, at least in the second language acquisition literature, there is often no detailed information about the learners themselves and the linguistic environment in which production was elicited" (Gass and Selinker 2001: 33).

On the whole, the contribution of CLC research to SLA so far has been much more substantial in description than interpretation of SLA data. In my view, there are two main reasons for this. First, as rightly pointed out by Hasselgård (1999), learner corpus research has mainly been conducted by corpus linguists rather than SLA specialists: "A question that remains unanswered is whether corpus linguistics and SLA have really met in learner corpus research. While learner language corpus research does not seem to be very controversial in relation to traditional corpus linguistics, some potential conflicts are not resolved, nor commented on by anyone from 'the other side'." It is undeniable that the term "learner corpus" – or "corpus" for that matter – is rarely found in SLA books and articles. However, there are signs that this is beginning to change. Two recent studies (Housen 2002; Wible and Ping-Yu Huang 2003) show the advantage of using CLC to test SLA hypotheses, in this case the Aspect Hypothesis. In particular, Housen (2002: 78) remarks that "computer-aided language learner corpus research provides a much needed quantificational basis" for current SLA hypotheses and makes it possible to "empirically validate previous research findings obtained from smaller transcripts, as well as to test explanatory hypotheses about pace-setting factors in second language acquisition" (ibid: 108).

The second reason for the emphasis on description has perhaps been that the type of interlanguage CLC researchers have been most interested in (i.e. the interlanguage of intermediate to advanced EFL learners) was so poorly described in the literature that they felt the need to establish the facts first before launching into theoretical generalisations. According to McLaughlin (1987: 80), this focus on description is typical of the interlanguage paradigm: "The emphasis in Interlanguage theory on description stems from a conviction that it is important to know well what one is describing before attempting to move into the explanatory realm. There is a sense that as descriptions of learners' interlanguages accumulate, answers will emerge to the larger questions about second-language acquisition."

Already now, even if it is still in the early stages, a much more accurate picture of advanced EFL interlanguage is beginning to emerge. This appears clearly from a recent excellent study by Cobb (2003) who replicated three European CLC studies with Canadian data and found a high degree of similarity. The three studies highlighted the following characteristics of advanced interlanguage: overuse of high frequency vocabulary (Ringbom 1998), high frequency of use of a limited number of prefabs (De Cock et al. 1998) and a much higher degree of involvement (Petch-Tyson 1998). Several other studies point to the stylistic deficiency of advanced learner writing, which is often characterised by an overly spoken style or a somewhat puzzling mixture of formal and informal markers. All in all, CLC studies suggest that "advanced learners are not defective native speakers cleaning up a smattering of random errors, but rather learners working through identifiable acquisition sequences. The sequences are not the *–ing* endings and third person *–s* we are familiar with, but involve more the areas of lexical expansion, genre diversification, and others yet to be identified" (Cobb 2003: 419).

Advanced interlanguage is the result of a very complex interplay of factors: developmental, teaching-induced and transfer-related, some shared by several learner populations, others more specific. An ongoing study of linkwords (Granger 2003b) in 5 different subcorpora of the *ICLE* (French, Dutch, Spanish, Italian and German learners) brings convincing evidence of this interplay of features. Some features, like the overuse of the coordinator *but* or the tendency to favour initial position for adverbial connectors, are probably partly developmental: they give evidence of a more simplified linking system. On the other hand, there are quite a few transfer-related uses. French learners' overuse of *indeed* is not shared by the other learner groups. It is clearly due to a faulty one-to-one equivalence between *indeed* and *en effet*, a tendency which is reinforced by teaching and reference books[6]. Some other phenomena, like the overuse of *nevertheless* or *on the one hand.....on the other hand* are clearly teaching-induced. They are the direct consequence of the long lists of connectors found in most ELT textbooks, which classify connectors in broad semantic categories (contrast, addition, result, etc.) but fail to provide guidelines on their precise semantic, syntactic and stylistic properties, thereby giving learners the erroneous impression that they are interchangeable. When combined, these factors can

reinforce each other. For instance, the overuse of *on the contrary*, which was attested in all five subcorpora of the *ICLE* and is probably teaching-induced, was found to be much more marked in the case of French- and Italian-speaking learners, due to the presence in the learners' mother tongue of a formally equivalent connector (*au contraire* and *al contrario*). Likewise, there is evidence that the tendency to place connectors in initial position may be reinforced by teaching (J. Flowerdew 2001: 81).

5 Computer Learner Corpora and FLT

The usefulness of computer corpora for FLT is now widely acknowledged and many would agree with Aston (1995: 261) that "corpora constitute resources which, placed in the hands of teachers and learners who are aware of their potential and limits, can significantly enrich the pedagogic environment". The main fields of application of corpus data are materials and syllabus design and classroom methodology.[7] In all three fields, there is very active work in progress, but, with the exception of ELT dictionaries, the number of concrete corpus-informed achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice. According to L. Flowerdew (1998a: 550), this is due to the fact that in most corpus studies "the implications for pedagogy are not developed in any great detail with the consequence that the findings have had little influence on ESP syllabus and materials design." As to classroom use of corpus data, although learners could undoubtedly benefit from exploring language to discover for themselves the underlying grammatical rules and/or typical patterns of use, teachers seem reluctant to introduce this type of "discovery learning" in their everyday teaching practices (see Mukherjee 2003). As learner corpora have developed much later than native corpora, one could expect CLC-informed pedagogical materials to be even more limited and yet activity in this field seems to be just as buoyant as in the native corpus field, already resulting in the production of new CLC-informed tools which address learners' attested difficulties. As space is limited, I will limit myself here to the description of two categories of CLC-informed ELT tools: learners' dictionaries and CALL (Computer-Assisted Language Learning) programs (for a more detailed survey of practical applications of learner corpora, see Granger forthcoming).

5.1 CLC-informed reference tools

Only a few years after the production of the first CLC-informed dictionary, the *Longman Essential Activator* (1997), learner corpus data have made their entry into general advanced learners' dictionaries. The latest editions of the *Longman Dictionary of Contemporary English (LDOCE)* (2003) and the *Cambridge Advanced Learner's Dictionary (CALD)* (2003) both contain language notes based on their respective learner corpora, notes intended to help learners to avoid

making common mistakes. The language notes in *LDOCE* are based on careful analysis of a raw (i.e. unannotated) corpus, while *CALD* has made use of an extensive error-tagged corpus (for a description of the error tagging system, see Nicholls 2003). The language notes are a clear added value for dictionary users as they draw their attention to very frequent errors, which in the case of advanced learners have often become fossilised (*accept* + infinitive, *persons* instead of *people*, *news* + plural, etc.). Most notes are useful but space is regrettably limited in paper versions of dictionaries and selecting the most useful information is a challenging task. There is no doubt, however, that in subsequent electronic versions of the dictionaries, where space is no longer so much of an issue, it will be possible to include much information derived from CLC analysis in the form of notes and crucially to provide much more L1-specific information, currently sorely lacking, but which is so important to learners who, even at an advanced stage of proficiency still have considerable difficulty with transfer-related interlanguage errors.

5.2 CLC-informed CALL programs

The pioneer of CLC-informed CALL programs is Milton (1998), who developed a writing kit called *WordPilot.* This program combines remedial exercises targeting Hong Kong learners' attested difficulties and a writing aid tool which helps learners to select appropriate wording by accessing native corpora of specific text types. Cowan et al.'s (2003) *ESL Tutor* program is an error correction courseware tool that contains units targeting persistent grammatical errors produced by Korean ESL students. The program is L1-specific, addressing errors that are clearly transfer-related. Wible et al.'s (2001) web-based writing environment is different from the other two as learner corpus building and analysis are integrated in normal pedagogical activities. The CALL environment contains a learner interface, where learners write their essays, send them to their teacher over the Internet and revise them when they have been corrected by the teacher, as well as a teacher interface, where teachers correct the essays using their favourite comments (comma splice, article use, etc.) stored in a personal Comment Bank. This environment is extremely attractive both for learners, who get immediate feedback on their writing and have access to lists of errors they are prone to produce, and for teachers, who progressively and painlessly build a large database of learner data from which they can draw to develop targeted exercises.

6 Conclusion

In learner corpus research, like in any corpus endeavour, "a great deal of spadework has to be done before the research results can be harvested" (Leech 1998: xvii). As I hope to have shown in this survey, researchers have spared no pains to build and analyse learner corpora and their efforts have been rewarded as the harvest has already begun. However, it is not yet time to rest on our laurels.

We need a wider range of learner corpora (in particular, ESP, speech and longitudinal data) with more elaborate processing (POS-tagging and error-tagging). Results need to be interpreted in the light of current SLA theory and incorporated in syllabus and materials design. Computer learner corpora have the potential of bridging the gap between SLA and ELT, but one must acknowledge that the ELT community has joined the learner corpus "revolution" (Granger 1994) more quickly and enthusiastically than the SLA community. There are signs that this is changing, as SLA specialists begin to recognise the value of CLC data which, by virtue of their size and representativeness, can help them validate their hypotheses and indeed formulate new ones. There are clearly exciting times ahead. Let's roll up our sleeves and get to work!

Notes

1. For an illustration of such a pilot study to test the reliability of automatic extraction of passives, see Granger 1997.
2. The USP (University of Sao Paulo) Multilingual Learner Corpus will contain German, English and Spanish L2 written data from Brazilian learners.
3. Note, however, that commercial corpora have been used for academic research and academic corpora for commercial purposes.
4. *LINDSEI* stands for *Louvain International Database of Spoken English Interlanguage*. Like its sister project, *ICLE*, it covers data from advanced EFL learners from various mother tongue backgrounds. More information on the project can be found on the following website: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm.
5. The learner corpus bibliography can be consulted on the following website: http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/publications.html. Suggestions for additions to the bibliography can be sent to granger@lige.ucl.ac.be.
6. The Robert-Collins English-French dictionary gives *en effet* as the first translation of *indeed*.
7. For an excellent overview of the usefulness of corpus data for materials development and classroom use, see Tomlinson (1998), Part A: Data collection and materials development, pp. 25-89.

References

Aarts, J. and S. Granger (1998), Tag sequences in learner corpora: A key to interlanguage grammar and discourse, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 132-141.

Aarts, J. and W. Meijs (eds) (1984), *Corpus linguistics: Recent developments in the use of computer corpora*, Amsterdam: Rodopi.

Aijmer, K. (2002), Modality in advanced Swedish learners' written interlanguage, *in* S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, pp. 55-76.

Aijmer, K., B. Altenberg, and M. Johansson (eds) (1996), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*, Lund, Sweden: Lund University Press.

Altenberg, B. (2002), Using bilingual corpus evidence in learner corpus research, *in* S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, pp. 37-54.

Altenberg, B. and M. Tapper (1998), The use of adverbial connectors in advanced Swedish learners' written English, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 80-93.

Archer, D., P. Rayson, A. Wilson, and T. McEnery (eds) (2003), *Proceedings of the Corpus Linguistics 2003 Conference*, Technical Papers 16, Lancaster University: University Centre for Computer Corpus Research on Language.

Aston, G. (1995), Corpus evidence for norms of lexical collocation, *in* G. Cook and B. Seidlhofer (eds), *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, pp. 257-270.

Atkins, S., J. Clear, and N. Ostler (1992), Corpus design criteria, *Literary and Linguistic Computing*, 7: 1-16.

Biber, D. (1993), Representativeness in corpus design, *Literary and Linguistic Computing*, 8 (4): 243-257.

Bley-Vroman, R. (1983), The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33: 1-17.

Chi Man-Lai, A., K. Wong Pui-Yiu, and M. Wong Chau-ping (1994), Collocational problems amongst ESL learners: A corpus-based study, *in* L. Flowerdew and A.K.K. Tong, *Entering text*, Hong Kong: Language Centre, Hong Kong University of Science and Technology, and Department of English, Guangzhou Institute of Foreign Languages, pp. 157-165.

Cambridge Advanced Learner's Dictionary (2003), Cambridge: Cambridge University Press.

Cobb, T. (2003), Analyzing late interlanguage with learner corpora: Québec replications of three European studies, *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 59 (3): 393-423.

Cook, G. and B. Seidlhofer (eds) (1995), *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press.

Cowan, R., H.E. Choi, and D.H. Kim (2003), Four questions for error diagnosis and correction in CALL, *CALICO Journal*, 20 (3): 451-463.

Dagneaux, E, S. Denness and S. Granger (1998), Computer-aided error analysis, *System: An International Journal of Educational Technology and Applied Linguistics*, 26: 163-174.

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: 59-80.

De Cock, S. (2000), Repetitive phrasal chunkiness and advanced EFL speech and writing, *in* C. Mair and M. Hundt (eds), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi, pp. 51-68.

De Cock, S., S. Granger, G. Leech, and T. McEnery (1998). An automated approach to the phrasicon of EFL learners, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 67-79.

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

Field, Y. (1993), Piling on the additives: The Hong Kong connection, *in* R. Pemberton and E. Tsang (eds), *Studies in lexis*, Hong Kong: Hong Kong University of Science and Technology, pp. 247-267.

Flowerdew, J. (2001), Concordancing as a tool in course design, *in* M. Ghadessy, A. Henry, and R.L. Roseberry (eds), *Small corpus studies and ELT*, Amsterdam: John Benjamins, pp. 71-92

Flowerdew, J. (ed.) (2002), *Academic discourse*, London: Longman.

Flowerdew, L. (1998a), Corpus-linguistic techniques applied to textlinguistics, *System*, 26: 541-552.

Flowerdew, L. (1998b), Integrating 'expert' and 'interlanguage' computer corpora findings on causality: Discoveries for teachers and students, *English for Specific Purposes*, 17: 329-345.

Flowerdew, L. and A.K.K. Tong (eds) (1994), *Entering text*, Hong Kong: Language Centre, Hong Kong University of Science and Technology, and Department of English, Guangzhou Institute of Foreign Languages.

Garside, R., G. Leech, and A. McEnery (eds) (1997), *Corpus annotation: Linguistic information from computer text corpora*, London: Longman.

Gass, S.M. and L. Selinker (2001), *Second language acquisition: An introductory course*, Mahwah, NJ: Lawrence Erlbaum.

Ghadessy, M., A. Henry, and R.L. Roseberry (2001), *Small corpus studies and ELT: Theory and practice*, Studies in Corpus Linguistics 5, Amsterdam: John Benjamins.

Gilquin, G. (2001), The integrated contrastive model: Spicing up your data, *Languages in Contrast*, 3 (1): 95-123.

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, and M. Johansson (eds), *Languages in contrast*, Lund, Sweden: Lund University Press, pp. 37-51.

Granger, S. (1998a), Prefabricated patterns in advanced EFL writing: Collocations and formulae, *in* A.P. Cowie (ed.), *Phraseology: Theory, analysis and applications*, Oxford: Oxford University Press, pp. 145-160.

Granger, S. (1998b), The computer learner corpus: A versatile new source of data for SLA research, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 3-18.

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

Granger, S. (1999), Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus, *in* H. Hasselgård and S. Oksefjell (eds), *Out of corpora*, Amsterdam: Rodopi, pp. 191-202.

Granger, S. (2002), A bird's-eye view of learner corpus research, *in* S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, pp. 3-33.

Granger, S. (2003a), *The International Corpus of Learner English*: A new resource for foreign language learning and teaching and second language acquisition research, to appear in *TESOL Quarterly*, special issue on corpus linguistics (Autumn 2003).

Granger, S. (2003b), A multi-contrastive approach to the use of linkwords by advanced learners of English: Evidence from the *International Corpus of Learner English*, Paper presented at the 'Pragmatic markers in contrast' workshop organized by the Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, 22-23 May 2003.

Granger, S. (forthcoming), Practical applications of learner corpora, *in* B. Lewandowska-Tomaszczyk (ed.), *Language, corpora, e-learning*, Peter Lang: Frankfurt.

Granger, S., E. Dagneaux, and F. Meunier (2002), *The International Corpus of Learner English: Handbook and CD-ROM*, Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com

Granger, S., J. Hung, and S. Petch-Tyson (eds) (2002), *Computer learner corpora, second language acquisition and foreign language teaching*, Language Learning and Language Teaching 6, Amsterdam: John Benjamins.

Granger, S. and S. Petch-Tyson (eds) (in press), *Extending the scope of corpus-based research: New applications, new challenges*, Amsterdam: Rodopi.

Granger, S. and P. Rayson (1998), Automatic profiling of learner texts, *in* S. Granger (ed.), *Learner English on computer*, pp. 119-131.

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.

de Haan, P. (1984), Problem-oriented tagging of English corpus data, *in* J. Aarts and W. Meijs (eds), *Corpus linguistics: Recent developments in the use of computer corpora*, London: Addison Wesley Longman, pp. 123-139.

de Haan, P. (1999), English writing by Dutch-speaking students, *in* H. Hasselgård and S. Oksefjell (eds), *Out of corpora*, Amsterdam: Rodopi, pp. 203-212.

de Haan, P. (2000), Tagging non-native English with the TOSCA-ICLE tagger, *in* C. Mair and M. Hundt (eds), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi, pp. 69-79.

Harmer, J. (2001), *The practice of English language teaching*, Harlow, UK: Longman.

Hasselgård, H. (1999), Review of Granger (ed.), Learner English on computer. *ICAME Journal*, 23: 148-152.

Hasselgård, H. and S. Oksefjell (eds) (1999), *Out of corpora*, Amsterdam: Rodopi.

Housen, A. (2002), A corpus-based study of the L2-acquisition of the English verb system, *in* S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, pp. 77-116.

Howarth, P. (1996), *Phraseology in English academic writing: Some implications for language learning and dictionary making*, Tübingen, Germany: Max Niemeyer Verlag.

Hunston, S. (2002), *Corpora in applied linguistics*, Cambridge: Cambridge University Press.

Källkvist, M. (1995), Lexical errors among verbs: A pilot study of the vocabulary of advanced Swedish learners of English, *Working papers in English and Applied Linguistics*, 2, Research Centre for English and Applied Linguistics, University of Cambridge: 103-115.

Källkvist, M. (1999), *Form-class and task-type effects in learner English: A study of advanced Swedish learners*, Lund Studies in English 95, Lund, Sweden: Lund University Press.

Larsen-Freeman, D. and M.H. Long (1991), *An introduction to second language acquisition research*, London: Longman.

Leech, G. (1992), Corpora and theories of linguistic performance, *in* J. Svartvik (ed.), *Directions in corpus linguistics*, Berlin: Mouton de Gruyter, pp. 105-22.

Leech, G. (1998), Learner corpora: What they are and what can be done with them, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, xiv-xx.

Lenko-Szymanska, A. (2003), Lexical problems in the advanced learner corpus of written data. Paper presented at PALC 2003 (Practical Applications of Language Corpora), Lodz, Poland, 4-6 April 2003.

Lewandowska-Tomaszczyk, B. and P.J. Melia (eds) (2000), *PALC'99: Practical applications in language corpora*, Frankfurt am Mein: Peter Lang.

Longman Dictionary of Contemporary English (2003), Harlow, UK: Longman.

Longman Essential Activator (1997), Harlow, UK: Longman.

MacWhinney, B. (2000), *The CHILDES Project, Volume 1: Tools for analysing talk: Transcription format and programs*, Mahwah, NJ: Lawrence Erlbaum.

Mair, C. and M. Hundt (eds) (2000), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi.

McEnery, T. and N.A. Kifle (2002), Epistemic modality in argumentative essays of second-language writers, *in* J. Flowerdew (ed.), *Academic discourse*, London: Longman, pp. 182-215.

McLaughlin, B. (1987), *Theories of second-language learning*, London: Edward Arnold.

Meunier, F. (1998), Computer tools for the analysis of learner corpora, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 19-37.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching, *in* S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, pp. 119-141.

Milton, J. (1998), Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 186-198.

Milton, J. and N. Chowdhury. (1994), Tagging the interlanguage of Chinese learners of English, *in* L. Flowerdew and A. K. K. Tong (eds), *Entering text*, Hong Kong: Language Centre, Hong Kong University of Science and Technology, and Department of English, Guangzhou Institute of Foreign Languages, pp. 127-143.

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

de Mönnink, I. (2000), Parsing a learner corpus, *in* C. Mair and M. Hundt (eds), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi, pp. 81-90.

Mukherjee, J. (2003), Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany, *in* this volume.

Neff J., E. Dafouz, H. Herrera, F. Martinez, J. Rica, M. Diez, R. Prieto, and C. Sancho (in press), Contrasting learner corpora: The use of modal and reporting verbs in expression of writer stance, *in* S. Granger and S.

Petch-Tyson (eds), *Extending the scope of corpus-based research: New applications, new challenges.*

Nesselhauf, N. (2003), The use of collocations by advanced learners of English and some implications for teaching, *Applied Linguistics*, 24: 223-242.

Nicholls, D. (2003), The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT, *in* D. Archer, P. Rayson, A. Wilson, and T. McEnery (eds), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*: 572-581.

Pemberton, R. and E. Tsang (eds) (1993), *Studies in lexis*, Hong Kong: Hong Kong University of Science and Technology.

Petch-Tyson, S. (1998), Writer/reader visibility in EFL written discourse, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 107-118.

Pravec, N.A. (2002), Survey of learner corpora, *ICAME Journal*, 26: 81-114.

Ragan, P.H. (1996), Classroom use of a systemic functional small learner corpus, *in* M. Ghadessy, A. Henry, and R.L. Roseberry (eds), *Small corpus studies and ELT*, Amsterdam: John Benjamins, pp. 207-236.

Renouf, A. (ed.) (1999), *Explorations in corpus linguistics*, Amsterdam: Rodopi.

Ringbom, H. (1998), Vocabulary frequencies in advanced learner English: A cross-linguistic approach, *in* S. Granger (ed.), *Learner English on computer*, London: Addison Wesley Longman, pp. 41-52.

Ringbom, H. (1999), High frequency verbs in the ICLE corpus, *in* A. Renouf (ed.), *Explorations in corpus linguistics*, Amsterdam: Rodopi, pp. 191-200.

Scott, M. (1996), *WordSmith Tools*, Oxford: Oxford University Press.

Swales, J. (2002), Integrated and fragmented worlds: EAP materials and corpus linguistics, *in* J. Flowerdew (ed.), *Academic discourse*, London: Longman, pp. 150-164.

Tagnin, S. (2003), A multilingual learner corpus in Brazil, Paper presented at the Learner Corpus Workshop organized within the framework of the Corpus Linguistics 2003 Conference (CL 2003), Lancaster, 28-32 March 2003.

Thomas, M. (1994), Assessment of L2 proficiency in second language acquisition research, *Language Learning*, 44: 307-336.

Tomlinson, B. (ed.) (1998), *Materials development in language teaching*, Cambridge: Cambridge University Press.

Tono, Y. (2000), A corpus-based analysis of interlanguage development: Analysing part-of-speech sequences of EFL learner corpora, *in* B. Lewandowska-Tomaszczyk and P.J. Melia (eds), *PALC'99: Practical applications in language corpora*, Frankfurt am Mein: Peter Lang, pp. 323-340.

Van Rooy, B. and L. Schäfer (2003), Automatic POS tagging of a learner corpus: The influence of learner error on tagger accuracy, *in* D. Archer, P. Rayson, A. Wilson, and T. McEnery (eds), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*, Lancaster University:

University Centre for Computer Corpus Research on Language, pp. 835-844.

Wible, D., C-H. Kuo, F-Y. Chien, A. Liu, and N-L. Tsao (2001), A web-based EFL writing environment: Integrating information for learners, teachers, and researchers, *Computers and education*, 37: 297-315.

Wible, D. and P-Y. Huang (2003), Using learner corpora to examine L2 acquisition of tense-aspect markings, *in* D. Archer, P. Rayson, A. Wilson, and T. McEnery (eds), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*, Lancaster University: University Centre for Computer Corpus Research on Language, pp. 889-898.