THE COMPUTER LEARNER CORPUS: A VERSATILE NEW SOURCE OF DATA FOR SLA RESEARCH

Sylviane Granger

In Granger, S. (ed.) (1998). *Learner English on Computer*. Addison Wesley Longman : London & New York, 3-18

1. Corpus linguistics and English studies

Since making its first appearance in the 1960s, the computer corpus has infiltrated all fields of language-related research, from lexicography to literary criticism through artificial intelligence and language teaching. This widespread use of the computer corpus has led to the development of a new discipline which has come to be called 'corpus linguistics', a term which refers not just to a new computer-based methodology, but as Leech (1992: 106) puts it, to a 'new research enterprise', a new way of thinking about language, which is challenging some of our most deeply rooted ideas about language. With its focus on performance (rather than competence), description (rather than universals) and quantitative as well as qualitative analysis, it can be seen as contrasting sharply with the Chomskyan approach and indeed is presented as such by Leech (ibid.: 107). The two approaches are not mutually exclusive, however. Comparing the respective merits of corpus linguistics and what he ironically calls 'armchair linguistics', Fillmore (1992: 35) comes to the conclusion that 'the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.'

The computer plays a central role in corpus linguistics. A first major advantage of computerization is that it liberates language analysts 'from drudgery and empowers [them] to focus their creative energies on doing what machines cannot do' (Rundell and Stock 1992: 14). More fundamental, however, is the heuristic power of automated linguistic analysis, i.e. its power to uncover totally new facts about language. It is this aspect rather than 'the mirroring of intuitive categories of description' (Sinclair 1986: 202) that is the most novel and exciting contribution of corpus linguistics.

English is undoubtedly the language which has been analysed most from a corpus linguistics perspective. Indeed the first computer corpus to be compiled was the Brown corpus, a corpus of American English. Since then English corpora have grown and diversified. At the time, the one million words contained in the Brown and the LOB were considered to be perfectly ample for research purposes, but they now appear microscopic in comparison to the 100 million words of the British National Corpus or the 200 million words of the Bank of English. This growth in corpus size over the years has been accompanied by a huge diversification of corpus types to cover a wide range of varieties: diachronic, stylistic (spoken vs. written; general vs. technical) and regional (British, American, Australian, Indian, etc.) (for a recent survey of English corpora, see McEnery and Wilson 1996).

Until very recently however, no attempt had been made to collect corpora of learner English, a strange omission given the number of people who speak English as a foreign language throughout the world. It was not until the early nineties that academics, EFL specialists and publishing houses alike began to recognize the theoretical and practical potential of computer learner corpora and several projects were launched,

among which the following three figure prominently: the International Corpus of Learner English (ICLE), a corpus of learner English from several mother tongue backgrounds and the result of international academic collaboration, the Longman Learners' Corpus (LLC), which also contains learner English from several mother tongue backgrounds and the Hong Kong University of Science and Technology (HKUST) Learner Corpus , which is made up of the English of Chinese learners.

2. Learner corpus data and SLA research

2.1. Empirical data in SLA research

The main goal of Second Language Acquisition (SLA)[1] research is to uncover the principles that govern the process of learning a foreign/second language. As this process is mental and therefore not directly observable, it has to be accessed via the product, i.e. learner performance data. Ellis (1994: 670) distinguishes three main data types: (1) language use data, which 'reflect learners' attempts to use the L2 in either comprehension or production'; (2) metalingual judgements, which tap learners' intuitions about the L2, for instance by asking them to judge the grammaticality of sentences; and (3) self-report data, which explore learners' strategies via questionnaires or think-aloud tasks. Language use data is said to be 'natural' if no control is exerted on the learners' performance and 'elicited' if it results from a controlled experiment.

Current SLA research is mainly based on introspective data (i.e. Ellis's types 2 and 3) and language use data of the elicited type. People have preferred not to use natural language use data for a variety of reasons. One has to do with the infrequency of some language features, i.e. the fact that 'certain properties happen to occur very rarely or not at all unless specifically elicited' (Yip 1995: 9). Secondly, as variables affecting language use are not controlled, the effect of these variables cannot be investigated systematically. Finally, natural language use data fails to reveal the entire linguistic repertoire of learners because 'they [learners] will use only those aspects in which they have the most confidence. They will avoid the troublesome aspects through circumlocution or some other device' (Larsen-Freeman and Long 1991: 26).

Introspective and elicited data also have their limitations however and their validity, particularly that of elicited data, has been put into question. The artificiality of an experimental language situation may lead learners to produce language which differs widely from the type of language they would use naturally. Also because of the constraints of experimental elicitation, SLA specialists regularly rely on a very narrow empirical base, often no more than a handful of informants, something which severely restricts the generalizability of the results. There is clearly a need for more and better quality data and this is particularly acute in the case of natural language data. In this context, learner corpora which, as will be shown in the following section, answer most of the criticisms levelled at natural language use data, are a valuable addition to current SLA data sources. Undeniably, however, all types of SLA data have their strengths and weaknesses and one can but agree with Ellis (1994: 676) that 'Good research is research that makes use of multiple sources of data.'

2.2. Contribution of learner corpora to SLA research

The ancestor of the learner corpus can be traced back to the Error Analysis (EA) era. However, learner corpora in those days bore little resemblance to current ones. First, they were usually very small, sometimes no more than 2,000 words from a dozen or so learners. Some corpora, such as the one used in the Danish PIF (Project in Foreign Language Pedagogy) project (see Faerch *et al.* 1984) were much bigger, though how much bigger is difficult to know as the exact size of the early learner corpora was generally not mentioned. This was quite simply because the compilers usually had no idea themselves. As the corpora were not computerized, counting the number of words had to be done manually, an impossible task if the corpus was relatively big. At best, it would sometimes have been possible to make a rough estimate of the size on the basis of the number of informants used and the average length of their assignments.

A further limitation is the heterogeneity of the learner data. In this connection, Ellis (1994: 49) comments that, in collecting samples of learner language, EA researchers have not paid enough attention to the variety of factors that can influence learner output, with the result that 'EA studies are difficult to interpret and almost impossible to replicate.' Results of EA studies and in fact a number of SLA studies have been inconclusive, and on occasion contradictory, because these factors have not been attended to. In his book on transfer, Odlin (1989: 151) notes 'considerable variation in the number of subjects, in the backgrounds of the subjects, and in the empirical data, which come from tape-recorded samples of speech, from student writing, from various types of tests, and from other sources' and concludes that 'improvements in data gathering would be highly desirable.'

Yet another weakness of many early learner corpora is that they were not really exploited as corpora in their own right, but merely served as depositories of errors, only to be discarded after the relevant errors had been extracted from them. EA researchers focused on decontextualized errors and disregarded the rest of the learner's performance. As a result, they 'were denied access to the whole picture' (Larsen-Freeman and Long 1991: 61) and failed to capture phenomena such as avoidance which does not lead to errors, but to under-representation of words or structures in L2 use (Van Els *et al.* 1984: 63).

Current learner corpora stand in sharp contrast to what are in effect proto-corpora. For one thing, they are much bigger and therefore lend themselves to the analysis of most language features, including infrequent ones, thereby answering one of the criticisms levelled at natural language use data (see section 2.1.). Secondly, there is a tendency for compilers of the current computer learner corpora (CLC), learning by mistakes made in the past, to adopt much stricter design criteria, thus allowing for investigations of the different variables affecting learner output. Last but not least, they are computerized. As a consequence, large amounts of data can be submitted to a whole range of linguistic software tools, thus providing a quantitative approach to learner language, a hitherto largely unexplored area. Comparing the frequency of words/structures in learner and native corpora makes it possible to study phenomena such as avoidance which were never addressed in the era of EA. Unlike previous error corpora, CLCs give us access not only to errors but to learners' total interlanguage.

2.3. Learner corpus data and ELT

The fact that CLCs are a fairly recent development does not mean that there was no previous link between corpus linguistics and the ELT world. Over the last few years, native English corpora have increasingly been used in ELT materials design. It was Collins Cobuild who set this trend and their pioneering dictionary project gave rise to a whole range of EFL tools based on authentic data. Underlying the approach was the firm belief that better descriptions of authentic native English would lead to better EFL tools and indeed, studies which have compared materials based on authentic data with traditional intuition-based materials have found this to be true. In the field of vocabulary, for example, Ljung (1991) has found that traditional textbooks tend to over-represent concrete words to the detriment of abstract and societal terms and therefore fail to prepare students for a variety of tasks, such as reading quality newspapers and report-writing. The conclusion is clear: textbooks are more useful when they are based on authentic native English.

However much of an advance they were, native corpora cannot ensure fully effective EFL learning and teaching, mainly because they contain no indication of the degree of difficulty of words and structures for learners. It is paradoxical that although it is claimed that ELT materials should be based on solid, corpus-based descriptions of native English, materials designers are content with a very fuzzy, intuitive, non-corpus-based view of the needs of an archetypal learner. There is no doubt that the efficiency of EFL tools could be improved if materials designers had access not only to authentic native data but also to authentic learner data, with the NS (native speaker) data giving information about what is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners. As a result, a new generation of CLC-informed EFL tools is beginning to emerge. Milton's (this volume*) Electronic Language Learning and Production Environment* is an electronic pedagogical tool which specifically addresses errors and patterns of over- and underuse typical of Cantonese learners of English, as attested by the HKUST Learner Corpus. In the lexicographical field, the Longman Essential Activator is the first learner's dictionary to incorporate CLC data (see Gillard and Gadsby this volume). In addition, use of CLC data could also give rise to new developments in ELT methodology (see Granger and Tribble this volume) and curriculum development (Mark 1996) within the framework of data-driven learning and form-focused instruction.

3. Learner corpus compilation

3.1. Learner corpus design criteria

'A corpus is a body of text assembled according to explicit design criteria for a specific purpose' (Atkins and Clear 1992: 5). It follows from this definition that a corpus needs to be carefully compiled. As pointed out by Sinclair (1991: 9) 'the results are only as good as the corpus'; in other words the quality of the investigation is directly related to the quality of the data. It is especially important to have clear design criteria in the case of learner language which is a very heterogeneous variety: there are many different types of learners and learning situations.

Not all features singled out by Atkins and Clear are relevant to learner corpus compilation. This is because a learner corpus is a 'special corpus', a category in which Sinclair (1995: 24) also includes the language of children, geriatrics, users of extreme dialects and very specialised areas of communication. Table 1.1 lists some of the main features which are relevant to learner corpus building.

LANGUAGE	**LEARNER**
Medium	Age
Genre	Sex
Topic	Mother tongue
Technicality	Region
Task setting	Other foreign languages
	Level
	Learning context
	Practical experience

Table 1.1: Learner corpus design criteria

Following Ellis (1994: 49) I distinguish between the features that pertain to the language situation and those that characterize the learner. The language attributes are fairly similar to those used in native corpus compilation. Medium distinguishes between written and spoken corpora. Within each medium several genres can be distinguished: for example, argumentative vs. narrative writing or spontaneous conversation vs. informal interview. It is very important to record this attribute because learner output has been shown to vary according to the task type. The topic is also a relevant factor because it affects lexical choice, while the degree of technicality affects both the lexis and the grammar (frequency of the passive, complexity of noun phrases, etc.). Task setting refers to features such as the degree of preparedness (timed vs. untimed), whether the task was part of an exam or not and whether the learners had access to ELT tools when performing the task and if so, which.

Apart from age and sex, all the attributes pertaining to the learner are proper to learner corpora. Because of the influence of the mother tongue on L2 output, it is essential to separate learners with different L1s. In addition, it is useful to record the region the learner comes from, in order to distinguish between the regional varieties of one and the same language, such as the differences in the French spoken in Belgium and in France. Learners may also be influenced in their English by other foreign languages and it is useful to be aware of these other possible influences. Whilst proficiency level is obviously of primary importance, it is also a somewhat subjective notion: people use terms such as 'intermediate' to refer to very different degrees of proficiency. For this variable - as for many others - the wisest course is to resort to 'external criteria', i.e. criteria which 'are all founded upon extra-linguistic features of texts' rather than 'internal features', which are essentially linguistic (Atkins and Clear 1992: 5). In practice, this means that the level of proficiency is defined by referring to the teaching level (primary/secondary/university) and/or the number of hours/years of English the learners have had. After this initial selection, it will be the researcher's task to characterize learners' proficiency in terms of internal evidence.

The learning context distinguishes between English as a Second Language (ESL) and English as a Foreign Language (EFL) according to whether English is learnt in an English-speaking country or not, a crucial distinction which is too often disregarded in SLA studies. Other variables which affect learner output are subsumed under the umbrella term 'practical experience' which covers the number of years of English teaching, the ELT materials used and the period of time, if any, spent in an English-speaking country.

This list, by no means exhaustive, can of course be adapted according to research goals. The main thing is to have clear criteria so as to achieve 'soundly based conclusions, making it not only possible but indeed legitimate to make comparisons between different studies' (Engwall 1994: 49).

3.2. ICLE design criteria

To illustrate how these criteria can be applied in practice, I will briefly describe the principles that have governed the compilation of the International Corpus of Learner English. As appears from Table 1.2 the criteria are of two kinds: some are common to all the subcorpora of ICLE and some are variable.

SHARED FEATURES	**VARIABLE FEATURES**
Age	Sex
Learning context	Mother tongue
Level	Region
Medium	Other foreign languages
Genre	Practical experience
Technicality	Topic
	Task setting

Table 1.2: *ICLE* design criteria

The subjects who have contributed data to ICLE share the following attributes. They are young adults (c. 20 years old), who are studying English in a non-English speaking environment, i.e. they are EFL, not ESL learners. Their level of proficiency is 'advanced', a notion which is defined on the following external ground: they are university undergraduates in English Language and Literature in their third or fourth year.

ICLE consists exclusively of written productions[2], which all represent the same genre, namely essay writing[3]. These essays, which are approximately 500 words long, are unabridged and so lend themselves to analyses of cohesion and coherence, two areas which are of particular interest in advanced learner writing. Although the essays cover a variety of topics, the content is similar in so far as the topics are all non-technical and argumentative in nature (rather than narrative, for instance)[4].

Several of the variable attributes relate to the learner. The first is sex: the corpus contains writing by both male and female learners. The second is the learner's mother tongue background. With the years ICLE has grown from just one national variety to 14 (French, German, Dutch, Spanish, Swedish, Finnish, Polish, Czech, Bulgarian, Russian, Italian, Hebrew, Japanese and Chinese) and the corpus keeps expanding, with new

varieties being added regularly. Some of the national varieties are subdivided regionally. The Dutch subcorpus, for example, includes data from learners living in the Netherlands and in Belgium. Other variables concern the amount/type of practical experience. The other variable features relate to the task: the essays cover a variety of topics; they can be timed or untimed, part of an exam or not; they may or may not have involved the use of ELT tools.

These variables are recorded via a learner profile questionnaire filled in by all learners (for a copy of the questionnaire, see Granger 1993)[5].

3.3. Corpus size

'The whole point of assembling a corpus is to gather data in quantity' (Sinclair 1995: 21). However, Sinclair hastens to add that 'In practice the size of a component tends to reflect the ease or difficulty of acquiring the material'. Is learner data easy to acquire? The answer is definitely 'no'. Even in the most favourable environment, i.e. technically advanced countries like Hong Kong where learners are expected to word-process their assignments, learner corpus compilation is a painstaking process (see Milton 1996: 235 for a description of these difficulties and useful advice on how to address them). As a result, one can hardly expect learner corpora to reach the gigantic sizes of native corpora.

One factor which has a direct influence on the size of learner corpora is the degree of control exerted on the variables described in the preceding section and this in turn depends on the analyst's objectives. If the researcher is an SLA specialist who wants to assess the part played by individual learner variables such as age, sex or task type, or if he wants to be in a position to carry out both cross-sectional and longitudinal studies, then he should give priority to the quality rather than the quantity of the data. This is not to say that size is not a consideration. As de Haan (1992) demonstrates, optimum corpus size depends on the specific linguistic investigation to be undertaken: for some linguistic studies, for instance those involving high frequency words or structures, relatively small samples of c. 20,000 words may be sufficient, while for others much larger samples are needed. In ICLE we have opted for a size of 200,000 words per national subcorpus. This is obviously very small in comparison with current NS corpora, but is nevertheless a real improvement over the narrow empirical foundation of most SLA studies.

Things are very different, however, if the researcher works within an ELT framework. A learner corpus compiled by an ELT publisher with a view to improving ELT tools, such as a learner's dictionary, for example, needs to be very large because it is supposed to be representative of a whole learner population. The 10 million word Longman Learners' Corpus, where the level of detail on learner attributes is kept to a minimum, is a good illustration of this type of corpus.

3.4. Data capture, mark-up and documentation

Data capture, mark-up and documentation are to a large extent similar for learner corpora and native corpora. As these stages are well documented in the literature (see in particular Barnbrook's 1996 excellent introductory textbook), I will limit myself in this section to a few practical hints for the would-be learner corpus builder.

Of the three methods of data capture - downloading of electronic data, scanning and keyboarding - it is keyboarding that currently seems to be most common in the field of learner corpora. Indeed it is the only method for learners' handwritten texts[6].

Fortunately the fast-growing number of computers at students' disposal both at home and on school/university premises is improving this situation and researchers can expect to get a higher proportion of material on disk in the near future.

Both keyboarded and scanned texts contain errors and therefore require significant proofreading. In the case of learner corpora this stage presents special difficulties. The proofreader has to make sure he edits out the errors introduced during keyboarding or scanning but leaves the errors that were present in the learner text, a tricky and time-consuming task. One useful way of starting the process is to run the learner texts through a spellchecker. Highlighted forms can then be compared to the original texts and eliminated or retained according as the case may be. Because of the limitations of the spellchecker, this process will only eradicate erroneous word-forms. Other errors such as omissions, additions or homonyms (*their/there, it's/its*) can only be spotted by careful manual editing. In any case, as errors can escape the attention of even the most careful of proofreaders, it is advisable to keep original texts for future reference.

Using a standard mark-up scheme called SGML (Standard Generalized Markup Language), it is possible to record textual features of the original data, such as special fonts, paragraphs, sentence boundaries, quotations, etc. Markup insertion is a very time-consuming process and researchers should aim for 'a level of mark-up which maximizes the utility value of the text without incurring unacceptable penalties in the cost and time required to capture the data' (Atkins and Clear 1992: 9). In the case of learner corpora, which tend to contain few special textual features, this stage can be kept to a minimum[7] although it should not be bypassed. For some types of analysis it would be highly advantageous to have textual features such as quotations, bold or underlining marked up in the learner corpus.

In order to be maximally useful, a corpus must be accompanied by relevant documentation. Full details about the attributes described in section 3.1. must be recorded for each text and made accessible to the analyst either in the form of an SGML file header included in the text files[8] or stored separately from the text file but linked to it by a reference system[9]. Both methods enable linguists to create their own tailor-made subcorpora by selecting texts which match a set of predefined attributes and focus their linguistic search on them. On this basis, learner corpus analysts are able to carry out a wide range of comparisons: female vs. male learners, French learners vs. Chinese learners, writing vs. speech, intermediate vs. advanced, etc.

4. Computer-aided linguistic analysis

4.1. Contrastive Interlanguage Analysis

A learner corpus based on clear design criteria lends itself particularly well to a contrastive approach. Not a contrastive approach in the traditional sense of CA (Contrastive Analysis) which compares different languages, but in the totally new sense of 'comparing/contrasting what non-native and native speakers of a language do in a comparable situation' (Pery-Woodley 1990: 143). This new approach, which Selinker (1989: 285) calls a 'new type of CA' and which I refer to as CIA - Contrastive Interlanguage Analysis - lies at the heart of CLC-based studies. James (1994: 14) sees

this new type of comparison as a particularly apt basis for a 'quantificational contrastive typology of a number of English ILs.'

CIA involves two major types of comparison:
(1) NL vs. IL, i.e. comparison of native language and interlanguage;
(2) IL vs. IL, i.e. comparison of different interlanguages.
As the two types of comparison have different objectives, it is useful to examine them separately.

NL/IL comparisons aim to uncover the features of non-nativeness of learner language. At all levels of proficiency but especially at the most advanced ones, these features will not only involve plain errors, but differences in the frequency of use of certain words, phrases or structures, some being overused, others underused. Before CLCs became available, work on learner production data had focused mainly on errors, but now, SLA specialists can also investigate quantitatively distinctive features of interlanguage (i.e. overuse/underuse), a brand new field of study which has important implications for language teaching. To take just one example, writing textbooks and electronic tools such as grammar checkers, even those designed for non-native speakers, advise learners against using the passive and suggest using the active instead. A recent study of the passive in native and learner corpora (see Granger forthcoming a) however, shows that learners underuse the passive and that they are thus not in need of this type of inappropriate advice.

NL/IL comparisons require a control corpus of native English and as clearly appears from section 1, there is no lack of them. A corpus such as ICE (International Corpus of English) even provides a choice of standards: British, American, Australian, Canadian, etc. One factor which analysts should never lose sight of, however, is the comparability of text type. As many language features are style-sensitive, it is essential to use control corpora of the same genre. As demonstrated in Granger and Tyson (1996: 23), a comparison of the frequency of three connectors - *therefore, thus* and *however* - in the LOB, a corpus covering a variety of text types, and ICLE, which only includes argumentative essay writing, leads to a completely distorted view: it fails to bring out the underuse of these connectors by learners, which clearly appears when a comparable corpus of native speaker argumentative essays is used instead. The corpus used in this case was the LOCNESS (Louvain Corpus of Native English Essays), a 300,000 word corpus of essays written by British and American university students. Whilst this corpus has the advantage of being directly comparable to ICLE, it has the disadvantage of being relatively small and containing student, i.e. non-professional writing. Criticisms can be levelled against most control corpora[10]. Each has its limitations and the important thing is to be aware of them and make an informed choice based on the type of investigation to be carried out.

The second type of comparison - IL vs. IL - may involve comparing ILs of the same language or of different languages. As the focus of this volume is on English interlanguages, I shall, in this section, deal exclusively with the former type. The main objective of IL/IL comparisons is to gain a better insight into the nature of interlanguage. By comparing learner corpora or subcorpora covering different varieties of English (different in terms of age, proficiency level, L1 background, task type, learning setting, medium, etc.), it is possible to evaluate the effect of these variables on learner output. Lorenz (this volume), for example, draws interesting conclusions on the effect of

age/proficiency on written output by comparing German learners of English from two different age groups with a matched corpus of native English students. On the other hand, the influence of the learner's mother tongue can be studied on the basis of corpora such as ICLE or LLC, which cover a high number of L1 backgrounds. In the field of grammar, the underuse[11] of passives mentioned above has been found to characterize advanced learners from three different mother tongue backgrounds - Swedish, Finnish and French - which might indicate that it is more of a cross-linguistic invariant than a transfer-related feature, as I would intuitively have thought on the basis of the French learner data only. Note, however, that for transfer to be unambiguously established, the researcher has to have access to good contrastive descriptions of the languages involved. Lack of reliable CA data casts doubt on the reliability of results of previous interlanguage investigations (see Kamimoto *et al.*'s 1992 criticism of Schachter's influential paper on avoidance). In other words, if we wish to be able to make firm pronouncements about transfer-related phenomena, it is essential to combine CA and CIA approaches[12].

4.2. Automated linguistic analysis

4.2.1. Linguistic software tools

One of the main advantages of computer learner corpora is that they can be analyzed with a wide range of linguistic software tools, from simple ones, which merely search, count and display, to the most advanced ones, which provide sophisticated syntactic and/or semantic analysis of the data. These programs can be applied to large amounts of data, thus allowing for a degree of empirical validation that has never been available to SLA researchers before.

Text retrieval programs - commonly referred to as 'concordancers' - are almost certainly the most widely used linguistic software tools. Initially quite rudimentary, they have undergone tremendous improvement over the last few years and the most recent programs, such as *WordSmith*, for example, have reached a high degree of sophistication, enabling researchers to carry out searches which they could never hope to do manually. They can count words, word partials and sequences of words and sort them in a variety of ways. They also provide information on how words combine with each other in the text. Finally, they can also carry out comparisons of entities in two corpora and bring out statistically significant differences, a valuable facility for CIA-type research. Gillard and Gadsby (this volume) illustrate well how CLC-based concordances can be used to discover patterns of error, which in turn can be converted into useful hints for learners in ELT dictionaries or grammars.

The value of computerization, however, goes far beyond that of quick and efficient manipulation of data. Using the appropriate electronic tools, SLA researchers can also enrich the original corpus data with linguistic annotations of their choice. This type of annotation can be incorporated in three different ways: automatically, semi-automatically or manually.

Part of speech (POS) tagging is a good example of fully automatic annotation. POS taggers assign a tag to each word in a corpus, which indicates its word-class membership. The interest of this sort of annotation for SLA researchers is obvious,

making it possible for them to conduct selective searches of particular parts of speech in learner productions.

Semi-automatic annotation tools enable researchers to introduce linguistic annotations interactively, using the categories and templates provided or by loading their own categories (see Meunier this volume).

If software does not exist for a particular type of annotation, researchers can always develop and insert their own annotations manually. This is the case for error tagging. Once an error taxonomy has been drawn up and error tags inserted into the text files, the learner corpus can be queried automatically and comprehensive lists of specific error types can be produced (see Milton and Chowdhury 1994, Milton this volume and Dagneaux *et al.* in preparation).

The corpus linguistics literature contains many good introductory surveys to linguistic software (see Barnbrook 1996 and McEnery and Wilson 1996 for a general survey, and Meunier, this volume, for a review of software tools for interlanguage analysis). Rather than duplicating these surveys, I will devote the final section of this article to some general methodological issues.

4.2.2. CLC methodology

One issue of importance in CLC methodology is how to approach learner corpora. Research can be hypothesis-based or hypothesis-finding. Using the traditional hypothesis-based approach, the analyst starts from a hypothesis based on the literature on SLA research and uses the learner corpus to test his hypothesis. The advantage of this approach is that the researcher knows where he is going, which greatly facilitates interpretation of the results. The disadvantage is that the scope of the research is limited by the scope of the research question.

The other approach is defined as follows by Scholfield (1995: 24): 'in more exploratory, "hypothesis-finding" research, the researcher may simply decide to gather data, e.g. of language activity in the classroom, and quantify everything he or she can think of just to see what emerges.' This type of approach is particularly well-suited for CLC, since the analyst simply has to feed the data into a text analysis program and wait to see what comes out. This approach is potentially very powerful since it can help us gain totally new insights into learner language. However, it is potentially a very dangerous one. SLA specialists should avoid falling prey to what I would call the 'so what?' syndrome, which unfortunately affects a number of corpus linguistics studies. With no particular hypothesis in mind, the corpus linguist may limit his investigation to frequency counts and publish the 'results' without providing any interpretation for them. 'So what?' are the words that immediately come to mind when one reads such articles.

There is no way, however, in which one approach is better, in absolute terms, than the other. Depending on the topic and the availability of appropriate software, the analyst will opt for one or the other or combine the two. Moreover, as rightly pointed out by Stubbs (1996: 47) 'The linguist always approaches data with hypotheses and hunches, however vague.' What matters then is that the CLC enables SLA researchers to approach learner data with a mere hunch and let the computer do the rest.

Another important methodological issue is the role of the statistical-quantitative approach in computer-aided analysis. Figures and linguists can be an explosive mixture

and the cry of 'it's statistically significant' is heard all too often in contexts where it has no real meaning. Scholfield's (1995) book 'Quantifying Language' aims to heighten linguists' awareness of the principles underlying a field he calls 'linguometry'. He warns them that 'No investigation, however clever the design or complex the statistical analysis, is any use if the "grassroots" measurement of the variables involved is unsatisfactory in some way' (ibid.: 29). But in the field of learner corpora, the notion of statistical significance should be weighed against that of pedagogical significance. A teacher analysing his learners' output with the help of computer techniques may well come up with highly interesting new insights based on quantitative information which may in itself not be statistically significant but which nevertheless has value within a pedagogical framework. A much greater danger than not being 'statistically significant', in my opinion, is to consider figures as an end in itself rather than a means to an end.

Lastly, a computerized approach has linguistic limitations. It is better suited to some aspects of language than others and SLA researchers should resist the temptation of limiting their investigations to what the computer can do. Ideally suited for the analysis of lexis and to some extent grammar, it is much less useful for discourse studies: 'many textual and discoursal phenomena of interest are harder to get at with the help of existing software, and a manual analysis of the texts then seems the only possibility' (Virtanen 1996: 162). SLA researchers should never hesitate to adopt a manual approach in lieu of or to complement a computer-based approach. As Ball (1994: 295) remarks, 'given the present state of the art, automated methods and manual methods for text analysis must go hand in hand.'

5. Conclusion

By offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will help researchers to get more of the facts right. They will contribute to SLA theory by providing answers to some yet unresolved questions such as the exact role of transfer. And in a more practical way, they will help to develop new pedagogical tools and classroom practices which target more accurately the needs of the learner.

Notes

[1] I use 'SLA' as a general term referring both to the description of learner language and the explanation of its characteristics. This stands in sharp contrast to the more restricted UG (Universal Grammar)-centred approach to SLA.

[2] We have started compiling a corpus of informal interviews. At present only the French learner variety is covered, but there are plans to extend it to other national varieties (see De Cock *et al* , this volume).

[3] Free compositions have a somewhat ambiguous status in SLA data typology. Though considered as elicited data, they are usually classified at the lower extreme of the +/- control continuum. They are different from clearly experimental data in that learners are free to write what they like rather than having to produce items the investigator is interested in. For

this reason they represent a special category of elicitation which Ellis (1994: 672) calls 'clinical elicitation' as opposed to 'experimental elicitation'.

[4] The corpus also contains a small proportion of literature exam papers.

[5] In one section of the questionnaire, students are requested to give their permission for the data to be used for research purposes.

[6] For the transcription of spoken material, see Edwards (1992) and Crowdy (1994).

[7] Going against an earlier decision (see Granger 1993: 62), it was later decided not to use mark-up to normalize any errors in the ICLE learner corpus because of the high degree of subjectivity involved.

[8] See Johansson (1994) for detailed encoding guidelines.

[9] Though both methods are valid, the latter is often preferred by corpus users because SGML-sensitive text retrieval software is not widely available (however, see Bradley 1996 and Quinn and Porter 1996 for recent developments in this field).

[10] A good candidate, however, would be a corpus of newspaper editorials, a text type which combines the advantages of being argumentative in nature and written by professionals.

[11] In my terminology 'underuse' is a neutral term, which merely reflects the fact that a word/structure is less used, while 'avoidance' implies a conscious learner strategy.

[12] For a description of the Integrated Contrastive Model, which combines CA and CIA, see Granger 1996a.