In Nerbonne J. (ed.) (1997) *Linguistic Databases*. CSLI Publications, Stanford, 175-188

# THE COMPUTER LEARNER CORPUS: A TESTBED FOR ELECTRONIC EFL TOOLS

Sylviane Granger
Centre for English Corpus Linguistics
Université Catholique de Louvain

**DRAFT**

## 1. CURRENT EFL TOOLS

Most current EFL (English as a Foreign Language) tools, be they dictionaries, grammars, grammar and style checkers or CALL (Computer Assisted Language Learning) software, have two things in common. Firstly, they no longer use invented examples, opting instead for examples taken from corpora of authentic language. And secondly, the tools seem in general to be designed for all learners of English, irrespective of their mother-tongue.

The focus on authenticity originated with the Collins Cobuild dictionary project, which gave rise to a whole range of EFL tools based on authentic data. Underlying the Collins Cobuild 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, M. 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.

As for the 'generic' nature of most EFL tools, there seem to be a variety of reasons for this, some theoretical, some practical and not least of which is the principle of unilingualism in ELT - ie the exclusive use of English in the teaching/learning process - which has undoubtedly played a large role. Although this principle may no longer be the dogma that it once was, it still nevertheless dominates the ELT scene. The general focus on universal features of learner language may also play a part. But the main reason is probably a practical, a commercial one. There is a much bigger market for all-round ELT tools than for L1-specific tools, which are obviously much more of a commercial risk for publishers. But this attitude is no longer acceptable at a time when all specialists recognize the importance of transfer in second language acquisition. Luckily, there are signs that things are changing, with the recent appearance of several bilingual reference books aimed at specific language groups. (1) It has to be said though that these bilingual tools are still the exception rather than the rule.

This creates something of a paradox. On the one hand is the belief that ELT materials should be based on solid, corpus-based descriptions of native English. On the other hand, materials designers are content with a very fuzzy, intuitive, non-corpus-based view of the needs of an archetypical learner. However, the efficiency of EFL tools could be improved if materials designers had access not only to authentic native data but also, as suggested in Figure 1, to authentic learner data, with the NS (native speaker) data giving information on 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.

**Computer corpora of native English**

**Computer corpora of learner English**

**EFL TOOLS**

Figure 1: Source data for EFL tools

The aim of this paper is to demonstrate the usefulness of using computer learner corpora in the design of EFL tools, with specific reference to electronic grammar and style checkers.

## 2. COMPUTER LEARNER CORPORA

The main advantage of the computer learner corpus (CLC) is that, benefitting from the techniques and tools developed in corpus linguistics, it can provide the EFL sector with a much more solid and versatile empirical foundation than has previously been available. CLCs are a relatively new development in corpus linguistics, and very few have been compiled to date, but the enthusiasm with which existing CLCs have been met suggests there will be rapid proliferation.

### 2.1. Corpus design criteria

As in other areas of corpus linguistics, it is important to ensure that certain criteria are established for the building and analysis of CLCs. In the discussion of these criteria which follows, particular reference will be made to the **International Corpus of Learner English (ICLE)**, a computerised corpus of writing of learners from different backgrounds (2).

In their seminal 1992 article S. ATKINS et al laid down the basic standards for corpus design. Their aim was to establish a set of criteria "necessary to foster the creation of high-quality compatible corpora of different languages, for different purposes, in different locations, and using different types of software and hardware" (p.1). The need for explicit corpus design criteria is particularly acute in the case of learner language because of the inherent heterogeneity of learner output. Among the text attributes listed by S. ATKINS et al, the following are of particular relevance to EFL corpus building: mode (written vs spoken), genre (essay, letter, conversation, etc.), function (narrative, expository, reflective, etc.) and technicality (general, technical, semi-technical). Most current EFL corpora cover the same mode - writing - but differ in other features of the writing assignment. The **Longman Learner Corpus** for example, contains a variety of genres, functions and degrees of technicality. **ICLE** on the other hand, a smaller, specialized corpus, contains only one task type, non-technical argumentative essay writing. Thus while it cannot make statements about all learner writing, it can be used to make reasonably definitive statements about a particular genre of learner writing.

Because of the major role played by transfer in second language acquisition, the language of the writer, an attribute which is, according to S. ATKINS et al (ibid:8) "in many cases unavailable or irrelevant", assumes particular importance in the CLC framework. An L1-undifferentiated EFL corpus would therefore be of relatively little interest. EFL corpus builders should either focus on one specific group of learners (the **Hong Kong University of Science and Technology Learner Corpus**, for instance, covers writing by Chinese (primarily Cantonese) learners) or cover a variety of mother tongue backgrounds (the option chosen by both the compilers of the Longman Learner Corpus and ICLE). (3) The advantage of the multi-language background learner corpus is that it makes it possible to distinguish between areas of difficulty specific to one language group and features common to several, perhaps all language groups, a distinction which would make it possible to improve both generic and bilingual EFL tools.

## 2.2. A computer-based methodology

A machine-readable raw (unannotated) learner corpus can be submitted to a whole range of text-handling software tools, thereby providing EFL analysts with a wealth of quantitative and qualitative data which has hitherto been inaccessible. The quantitative measures which are most easily provided are frequency counts of morphemes, words and phrases. Comparisons with NS (native speaker) frequency counts highlight patterns of over- and underuse which often provide impetus for further analysis. Concordancing software enables the researcher to take a more qualitative look at the learner data and detect cases of syntactic, semantic or stylistic misuse. Though research in this area is still in its infancy, this computer-aided methodology has already shed new light on several aspects of learner language: use of connectors (MILTON & TSANG 1993; GRANGER & TYSON 1996), collocations and lexical phrases (CHI et al 1994; GRANGER forthcoming), verbs of saying (TRIBBLE 1991), modal expressions (DAGNEAUX 1995). Table 1, which displays concordances of the word **anyway** extracted from the ICLE database (French learner subcorpus) and a comparable corpus of native English writing, demonstrates clearly that while native speakers use the word **anyway** primarily in clause or sentence final position, French learners distinctly favour sentence initial position.

**Non-native writing**

| | |
|---|---|
| some time dreaming or not. | >Anyway we cannot affirm |
| some kind of bygone days. | >Anyway, a solid amount |

| | |
|---|---|
| ecological policies. | >Anyway, they do not have |
| orientations has to be found. | >Anyway, "wait and see" see |
| am I going too far? | >Anyway what must be kept in |
| labyrinth with reality... | >Anyway, when the dreams |
| death, life will go on | >anyway. Those numbers have |
| whether it is possible and | >anyway, it is not true. |

**Native writing**

| | |
|---|---|
| The fact that he married her | >anyway, and the photos of |
| should have thrown it | >anyway. By saving those two |
| death to a crime of passion, | >anyway, as yet another |
| they're going to die | >anyway. This shows that |
| when we are going to die | >anyway. The laugh causes |
| being looked on favourably | >anyway, after accepting the |
| were not eligible for it | >anyway. Then, as we can |
| Greenland was an exception | >anyway because it's reason |

Table 1: Concordances of **anyway** in NS and NNS writing

NS/NNS concordance-based comparisons frequently highlight erroneous patterns in learner writing, such as the use of the marked infinitive after **accept** alongside the grammatical **that**-clause complementation in Table 2.

**Non-native writing**

| | |
|---|---|
| families, the parents accept | >that new visions of things |
| don't always accept | >that their children also |
| nor the children accept | >to recognize that |
| young. He could never accept | >to be inferior. |
| Feminists have to accept | >to be treated as men |

**Native writing**

| | |
|---|---|
| not being able to accept | >that fulfilment of |
| the act. Hugo cannot accept | >that the party line has |

Table 2: Concordances of **accept** in NS and NNS writing

Results of such comparisons can be used by materials designers to draw users' attention to frequent pitfalls. Specialists in EFL pedagogy have also suggested using such concordances in the classroom as starting points for discussion of grammar and usage in the context of data-driven learning (see C. TRIBBLE & G. JONES 1990; V. PICKARD 1994; T. JOHNS 1994).

Concordancers are of great value in highlighting marked and/or erroneous use in learner data. However current concordancers have limitations. As J. KIRK (1994: 25) points out, one of their major weaknesses is that they operate on the basis of unlemmatized word forms. This means for example, that anyone interested in NNS use of the verb **be** will have to search for 9 different forms (**be, 's, is, isn't, was, wasn't, were, weren't, been**). A lemmatizer would automatically group together all inflected forms of one and the same lemma. Regrettably, none of the currently available concordancing packages contain a fully automatic lemmatizer, although Wordsmith, the new package from Oxford University Press, contains a lemmatization facility, which enables users to link several entries.

Another difficulty faced by EFL analysts working with raw, learner corpora is that words in English frequently belong to more than one word category. A search for the major prepositions in learner writing will bring up all the occurrences of **in, over, after, etc.** as adverbial particles, thus forcing the analyst to undertake a long and tedious process of manual disambiguation. This can be avoided if the learner corpus is tagged. The ICLE corpus is currently being tagged with the TOSCA analysis system. (4) Preliminary investigation brings out the advantages of a tagged CLC over a raw CLC: it is both more efficient, in that it only gives the analyst what he/she wants (prepositions and not adverbial particles, for instance), and more powerful, in that it allows for grammatical as well as lexical searches. A search for the grammatical tag AUX (modal) in the French subcorpus of ICLE by F. MEUNIER (1995) and a similar-sized native corpus using the same tags brings out a significant overuse of modal auxiliaries by the learners. A subsequent lexical search proves this to be to a large extent due to overuse of **can**. Such comparisons are sure to shed new light on learners' use of grammar. It is also possible to search for sequences of tags. Table 3 (taken from F. MEUNIER 1995) illustrates a search for the verb-adverb-noun sequence in NS and NNS data, a structure which is both overused and misused by French learners of English.

**V (montr,*) + ADV (ge/intens/partic) + ART/N(com)**
Verb (monotransitive, pres/past/-ed participle) + Adverb (general, intensifier or particularizer) + Article or Common Noun.

**Examples of NNS structures:**
*...and to select **carefully** the programmes.*
*...people like maecenas able to support **materially** artists.*
*...This justifies **naturally** the success of all these games.*

**Example of NS structure:**
*This illustrates **emphatically** the folly of l'Optimisme.*

Table 3: Combination-of-tags search in NS and NNS data

## 3. A COMPUTATIONAL MODEL OF EFL PERFORMANCE

Though it is rarely stated explicitly, the underlying assumption behind most grammar checkers is that they are intended for any user, native or otherwise. However, several studies have demonstrated that existing grammar checkers are "not reliable writing aids for non-native speakers" (MILTON 1994:66) and indeed, if one analyses their performance in detail, it becomes apparent that they cater primarily for the needs of the native writer.

This was confirmed by a recent study which compared the effectiveness of two widely-used checkers, Correct Grammar and Grammatik, in analyzing samples native and non-native data (GRANGER & MEUNIER 1994). Whilst problems shared by both native and non-native speakers such as concord or double negation had a reasonable success rate, the rate of detection for typically non-native problems was generally low. The whole area of lexico-grammar is a case in point. The success rate of dependent preposition errors (**discuss *about sth, depend *to sth**) or non-finite verb complementation errors (**prevent sb *to do sth, insist *to do sth**), for instance, proved to be extremely low.

What is lacking is an adequate computational model of EFL performance. According to CATT & HIRST (1990) it is this that is holding back the whole field of computer-assisted language instruction (CALI). Current CALI systems, which merely match students' answers with a predefined list of responses, should be replaced by more intelligent CALI (ICALI) systems which can deal with learners' free-form input. For that however, it is necessary to construct a computational model of EFL performance: "Only if the learner's competence can be modelled computationally is there hope of developing CALI systems capable of dealing intelligently with learner language" (CATT & HIRST 1990:6).

The Scripsi system developed by CATT & HIRST remedies this deficiency "by incorporating a credible model of the second language learner's linguistic competence, one that takes into account the phenomena of transfer and overgeneralization" (1990:22). This rule-based error diagnostic tool detects transfer errors by allowing the application of L1 rules in the parsing process. It detects overgeneralization errors by relaxing constraints on feature values (for a full description of the system, see CATT 1988:27-53). Scripsi is a prototype system, which only caters for a very limited number of errors from French- and Chinese-speaking learners of English. Other error-diagnosing parsers, which combine 'correct' and 'lenient' rules, are being developed (COVINGTON & WEINRICH 1991; LUOMAI 1994), but they suffer from the same weakness as Scripsi, ie they have a very low lexical and syntactic coverage.

If they are to be truly effective, error diagnostic tools need to cater for the most typical errors in a particular learner population. As TSCHICHOLD (1994:198) rightly points out: "The fact that a grammar checker can correct an error the user is not ever likely to make may be impressive, but it is not very useful to users". In constructing their model of learner competence, CATT & HIRST have used data from well-known EA (error analysis) studies. Such data suffer from two major weaknesses: size and heterogeneity. Most EA studies are based on very small samples of learner language, sometimes no more than 2,000 words representative of a dozen or so learners. In addition, EA researchers have often not paid attention to the variety of factors which can influence learner language, such as mother tongue background, medium, level of proficiency, etc. EA data may therefore be a good source of data to build a prototype but will be of little use beyond that stage. In the following section I will show how computer learner corpora, which contain large quantities of carefully collected learner data, can contribute to improving the diagnostic capabilities of grammar and style checkers.

## 4. CONTRIBUTION OF CLC DATA TO GRAMMAR AND STYLE CHECKERS

This section reports on some preliminary results of an ongoing two-year project to adapt current English grammar and style checkers for French-speaking users. The data used in the project is the 300,000 word component of the ICLE database, which contains samples of writing from 450 different advanced French-speaking learners. With a view to widening the proficiency range of the data this corpus was supplemented with a 75,000 word corpus of writing by intermediate learners.

The first stage of the project involved manually correcting and error tagging a 150,000 word subset of the overall corpus, evenly distributed between intermediate and advanced. The error tagging system devised for the project is based on the following principles. Each error tag carries a general category tag: G for grammatical, L for lexical, F for formal, etc. and one or more specific codes. For instance, the GVT code refers to grammatical verb tense errors. Table 4 shows a sample of error tagged material: errors are preceded by the relevant error tag and followed by the corrected form. (5)

The first positive aspect would be the age of the child. (CLS) Actually $In fact$ studies carried out by eminent linguists have proved that the ideal (FS) adge $age$ for learning a foreign language is between 3-10 years. (CLC) Of course $0$ no consensus (GVAUX) could be reached $has been reached$ about this subject as others (LSF) pretend $claim$ that this trend would lead to mental fatigue, overworking and that it could disturb the children's (GNN) mind $minds$. In their opinion, it would be preferable to learn a second language only after having a (FS) throughout $thorough$ knowledge of the first.

Table 4: Sample of error-tagged text

Once a text has been tagged for errors, it is possible to draw up comprehensive inventories of specific error types. Table 5 shows the concordance of error-tagged data sorted on the error tag GVPR, which marks errors made with dependent prepositions following verbs. Table 6 contains a list of LSF errors, ie lexical single word errors due to the influence of a formally similar word in the user's mother tongue (the so-called 'faux amis') and Table 7 contains count/uncount noun errors.

| | |
|---|---|
| the fact that we could | (GVPR) argue on $argue about$ the definition |
| want to be parents, do not | (GVPR) care of $care about$ the sex |
| is rising. These people who | (GVPR) come in $come to$ Belgium |
| Family planning | (GVPR) consists on $consists of$ |
| have the possibility to | (GVPR) discuss about $discuss$ their problems |
| which the purchaser cannot | (GVPR) dispense of $dispense with$ |
| the health. Nobody | (GVPR) doubts about $doubts$ that. |

| | |
|---|---|
| harvest they got is often | (GVPR) exported in $exported to$ countries |

Table 5: Error tag search: dependent prepositions

| | |
|---|---|
| seems to us, thanks to its | (LSF) actual $modern$ style of |
| West leads the | (LSF) actual $present$ world of economy |
| fall into despair, the truth will | (LSF) conduct $drive$ him to suicide |
| goes to his or her | (LSF) course $lecture$ but because |
| to others, it is rather | (LSF) deceiving $disappointing$ |
| it is not impossible that such | (LSF) experiences $experiments$ |
| employ an | (LSF) important $large$ number of |
| be shown thanks to a certain | (LSF) mark $brand$ of cigarettes |
| since they are allowed to | (LSF) penetrate $enter$ our country |
| health of stressed | (LSF) pork $pigs$ |

Table 6: Error tag search: faux amis

| | |
|---|---|
| of advice on | (GNUC) a $0$ better health care |
| for years. Undoubtedly | (GNUC) a $0$ big progress has been made |
| you from breathing | (GNUC) a $0$ pure air |
| seems to be different. | (GNUC) A $0$ clear evidence is the percentage |
| characteristic | (GNUC) behaviours $behaviour$ |
| It provides | (GNUC) employments $employment$ |
| combining study life and | (GNUC) leisures $leisure facilities$ |
| a balance between work and | (GNUC) spare times $spare time$ |
| need to do some | (GNUC) works $work$ or simply for your personal |

Table 7: Error tag search: uncount nouns

By running these lists of errors through current checkers, it is possible to give a precise assessment of which rules - be they grammatical, lexical, orthographic or stylistic - need to be refined or added to the existing stock. One area among many where improvement is badly needed is that of the count/uncount distinction illustrated in Table 7. This distinction is responsible for many errors in learner writing: it affects number (***leisures** vs **books**), article usage (***without car** vs **without passion**), use of indefinite determiners (***many leisure** vs **many books**) or quantifying expressions (***a great deal of books** vs **a great deal of money**), etc.

Our research, however, is not limited to the analysis of error lists. Besides identifying errors, a good grammar and style checker should also detect and correct stylistic infelicities in learner writing. The clichés, circumlocutions and other infelicitous words and expressions highlighted by current checkers are not very useful because they are typical of native usage. A phrase such as **I don't doubt but that** is not likely to be used by learners. Here too CLC data prove very useful. The whole 300,000 word database can be scanned for recurring words and phrases and the results compared with a corpus of native writing of the same genre. Such a search brings out overused phrases such as **we can say that, we must not forget that, from the point of view of, as far as (x is concerned**) which make learner writing sound verbose and pompous (see DECOCK et al forthcoming).

Though the project is not complete, some preliminary conclusions can already be drawn.

First of all, the high number of L1-dependent errors detected would seem to require that separate programs be developed for each mother tongue background. The LSF category alone, ie the so-called 'faux amis' illustrated in Table 5, makes up one third of all the lexical errors in the corpus. Only when several mother tongue backgrounds have been covered will we be able to determine the nature and extent of the common core of EFL problems.

Secondly, it would seem to be essential that an effective grammar and style checker should cater for learners' lexical errors. Our research demonstrates that the majority of EFL errors are lexical. A key objective of the project will be to assess how best to cater for lexical errors.

In general, it seems reasonable to say that CLC - both in their tagged and untagged versions -prove to be a particularly rich source of data for improving grammar checkers. Error-tagged CLC bring out the typical errors in a given learner population, informing us of the relative weight of each category type and spelling out the most error prone members in each category. Untagged CLC bring out stylistic infelicities in learner writing.

It should be noted, however, that although our research so far shows clearly that all components of grammar checkers can be substantially improved by the use of CLC data, there is still a large proportion of errors which will remain beyond the capabilities of grammar checkers for the foreseeable future, because they are semantic in nature. This restriction affects lexis as well as grammar: many grammatical errors are undetectable because they display no formal sign of error. For instance, article errors due to the generic vs specific distinction (**I like the carrots** vs **I like carrots**), modal auxiliary errors (the difference between **I can do it** and **I may do it**) or discourse-level pronoun errors (**it is true** vs **this is true**) all escape detection. As a consequence, grammar checkers need to be viewed as one component in a wider writing workstation, which would include a series of interactive lookup facilities, such as an online grammar, an online learners' dictionary and an online collocational database. As MILTON (1994:68) rightly points out: "students need more than just pruning shears to cut away errors (...) They also require, in the absence of human assistance, a better online expert than the type that is now available".

## 4. CONCLUSION

The native speaker perspective occupies a central position in current ELT material. It is typicality of usage in native English that determines syllabus design; it is native speaker language that serves as a testbed for grammar checkers. This is understandable since native proficiency is what all learners are after. In fact, EFL materials designers need to have access to more detailed descriptions of native English use than are currently available, enabling improvements along the lines of the frequency data which has been incorporated in the latest editions of the Collins Cobuild Dictionary and the Longman Dictionary of Contemporary English. However, pedagogic practice shows that the native perspective needs to be supplemented with a learner perspective. Computer learner corpora are the best way of finding out about learners' difficulties and they will undoubtedly play a major role in the development of future EFL tools.

## ACKNOWLEDGEMENTS

We acknowledge the support in this research provided by the Belgian National Scientific Research Council, the Ministry of Education of the French-speaking Community of Belgium and the University of Louvain Research Fund.

## NOTES

1. Cambridge University Press has just brought out a new series called Cambridge Word Routes, which is a sort of bilingual thesaurus. And Collins has just published a student's dictionary specially developed for Brazilian learners of English.
2. For more information on the ICLE database, see Granger 1993, 1994 and 1996.
3. The International Corpus of Learner English contains eleven subcorpora which correspond to the following mother tongue backgrounds: French, Dutch, German, Spanish, Swedish, Finnish, Polish, Czech, Russian, Japanese and Chinese.
4. The TOSCA analysis system was developed by Jan Aarts and his team at the University of Nijmegen within the framework of the International Corpus of English (ICE) project, with which the ICLE corpus has close links.
5. The insertion of the corrected form seems to contradict a statement made in a previous article according to which “Errors should not be normalised, as this involves a high degree of subjectivity, given that many errors - particularly lexical, stylistic and textual ones - can be corrected in many different ways” (GRANGER, MEUNIER & TYSON 1994:105). Corrected forms were inserted within the framework of this project because it was useful for the researchers to have access to them when analysing and categorizing the error data.

## BIBLIOGRAPHY

Aarts J., P. de Haan & N. Oostdijk (eds.) (1993), **English Language Corpora: Design**, **Analysis and Exploitation**, Rodopi (Amsterdam/Atlanta).

Atkins S., J. Clear & N. Ostler (1992), Corpus Design Criteria, **Literary and Linguistic Computing** 7/1: 1-16.

Catt M. (1988) **Intelligent Diagnosis of Ungrammaticality in Computer-Assisted Language Instruction**. Technical Report CSRI-218. Computer Systems Research Institute: University of Toronto.

Catt M. & G. Hirst (1990) An Intelligent CALI System for Grammatical Error Diagnosis. **CALL**, Volume 3, 3-26.

Chi A.M., Pui-yiu K.W. & Chau-ping M.W. (1994), Collocational problems amongst ESL learners: a corpus-based study, in Flowerdew & Tong (eds.): 157-165.

Covington M. & K. Weinrich (1991), Unification-based Diagnosis of Language Learners' Syntax Errors, **Literary and Linguistic Computing** 6/3: 149-154.

Dagneaux E. (1995), **Epistemic modal expressions in native and non-native writing**, unpublished MA Dissertation, Université Catholique de Louvain: Louvain-la-Neuve.

Decock S., S. Granger, G. Leech & T. McEnery (forthcoming) An Automated Approach to the Phrasicon of EFL Learners, in S. Granger (ed.) **Learner English on Computer**, Addison Wesley Longman (1997).

Flowerdew L. & A.K.K. Tong (eds.) (1994), **Entering Text**, The Hong Kong University of Science and Technology.

Fries U., G. Tottie & P. Schneider (eds.) (1994), **Creating and using English language corpora**, Rodopi (Amsterdam/Atlanta).

Granger S. (1993), International Corpus of Learner English, in J. Aarts, P. de Haan & N. Oostdijk (eds.): 57-71.

Granger S. (1994), The Learner Corpus: a Revolution in Applied Linguistics, **English Today** 10/3: 25-29.

Granger S. (1996), Learner English around the World, in S. Greenbaum (ed.) **Comparing English World-wide**, Clarendon Press: Oxford, 13-24.

Granger S. (forthcoming), Prefabricated patterns in advanced EFL writing: collocations and lexical phrases, to appear in: T. Cowie (ed.), **Phraseology**, Oxford University Press.

Granger S. & F. Meunier (1994), Towards a grammar checker for learners of English, in Fries et al (eds.): 79-91.

Granger S., F. Meunier & S. Tyson (1994) New Insights into the Learner Lexicon: a preliminay report from the International Corpus of Learner English, in L. Flowerdew & A. Tong (eds.), 102-113.

Granger S. & S. Tyson (1996), Connector usage in the English essay writing of native and non-native EFL speakers of English, **World Englishes**, Volume 15, No. 1, 19-29.

Johansson S. & A.B. Stenström (eds.) (1991) **English Computer Corpora**, Mouton de Gruyter: Berlin & New York.

Johns T. (1994), From printout to handout: Grammar and vocabulary teaching in the context of Data-driven Learning, in T. Odlin (ed.): 293-313.

Kirk J. (1994), Taking a byte at Corpus Linguistics, in Flowerdew & Tong (eds.): 18-49.

Ljung M. (1991), Swedish TEFL meets reality, in Johansson & Stenström (eds.): 245-256.

Luomai X. (1994), Smart Marker - an efficient GPSG parser, in Flowerdew & Tong (eds.): 144-156.

Meunier F. (1995), Tagging and parsing interlanguage, in: L. Beheydt (ed.) **Linguistique appliquée dans les années nonante**, ABLA Papers 16, 21-29.

Milton J. (1994), A Corpus-Based Online Grammar and Writing Tool for EFL Learners: A Report on Work in Progress, in Wilson & McEnery: 65-77.

Milton J. & K.S.T. Tong (eds.) (1991), **Text Analysis in Computer-assisted Language Learning**, The Hong Kong University of Science and Technology.

Milton J. & E.S.C. Tsang (1993), A corpus-based study of logical connectors in EFL students' writing: directions for future research, in Pemberton & Tsang (eds.): 215-246.

Odlin T. (ed.) (1994**), Perspectives on Pedagogical Grammar**, Cambridge University Press.

Pemberton R. & E.S.C. Tsang (eds.) (1993), **Studies in Lexis**, The Hong Kong University of Science and Technology.

Pickard V. (1994), Producing a concordanced-based self-access vocabulary package: some problems and solutions, in Flowerdew & Tong (eds.): 215-226.

Tribble C. (1991), Electronic Text Analysis in Curriculum Design and Evaluation, in Milton & Tong (eds.): 4-14.
Tribble C. & G. Jones (1990) **Concordances in the Classroom**, Longman.
Tschichold C. (1994) Evaluating second language grammar checkers, **TRANEL** 21, 195-204.
Wilson A. & T. McEnery (eds.) (1994), **Corpora in Language Education and Research**, Unit for Computer Research on the English Language: Lancaster.