Computer learner corpora, or how can we turn our students' interlanguage into a resource for EFL research and teaching?

Maria Teresa Prat Zagrebelsky
Università di Torino

Abstract

The essay presents recent developments in the field of learner corpora, with special reference to some written corpora that have been collected in European academic contexts. One particular corpus, "The International Corpus of Learner English" (ICLE), is dealt with in more detail since it allows cross-linguistic comparison among 11 different national subcorpora. The research that has already been carried out on ICLE ranges from *ad hoc* case studies of mistakes made by specific groups of learners to wider and more systematic investigations of areas of difficulty and of features of "foreign-soundings" across different mother tongue groups, and from the discussion of corpus design criteria to the empirical testing of Second Language Acquisition hypotheses and findings. Using an already existing learner corpus, joining a project in progress or building a new learner corpus appear to be useful and challenging enterprises which may help reconcile research and teaching needs especially in the field of English, which is the most highly required foreign language in European universities.

Introduction

At present, in many Italian and other European universities, to be a lecturer in Modern English implies that you are required to do two things at the same time. On the one hand, you teach courses on various aspects of modern English and introduce your students to one or several methods of linguistic analysis, with a growing emphasis on translation skills. On the other hand, you are also responsible for students' language improvement in cooperation with native speaker experts and language centres. The latter responsibility means to set internal language standards, to compare them to international standards, to plan final written and oral tests and to evaluate students for EFL proficiency several times a year. If you teach in a large university, marking and evaluating students' productions will take a large part of your professional life. A good background in error analysis, contrastive analysis and interlanguage studies is therefore desirable as well as the awareness of recurrent and typical mistakes and areas of difficulty to be dealt with in remedial teaching materials or to be focused upon

in courses. Therefore, storing students' productions electronically and analysing them through software programmes may prove extremely useful and help reconcile teaching and research needs.

1. What is a computer learner corpus?

A learner corpus is the product of one of the branches —probably the least known so far— of the flourishing field of corpus linguistics. To start with a general definition, a computer learner corpus is a computerized textual database of the language produced by foreign language learners. The data can be exploited through the application of specific software programmes in order to obtain information on learner language on a larger scale and in a more systematic and reliable way than through manual search or teacher intuition.

2. What computer learner corpora are available for English?

Learner corpora are made of either oral or written learner productions. The two should be complementary; however, so far oral corpora have been rare because the recording and the phonological transcription, and even the simple orthographic transcription, are extremely time consuming processes. A notable example is the Lindsei Corpus, collected in Louvain la Neuve by Sylvie De Cock, which contains 50 interviews of advanced French and Italian EFL learners while other national subcorpora and an English comparable corpus are being developed (for more information , see the web page of the "Centre for English Corpus Linguistics ").

Ten written learner corpora are available for English today, according to Pravec (2002) and there are probably many more which have not been publicized. In her recent survey in the ICAME Journal nº.26, Pravec gives detailed information for six of them according to a set of corpus design criteria (see Table 1).

Table 1.

COMMERCIAL LEARNER CORPORA

NAME	SIZE	TYPE OF TEXT	LANGUAGE BACKGROUND	WEBSITES
Cambridge Learner Corpus (CLC)	15.000.000 words and expanding; about one quarter tagged for errors	Cambridge exam scripts at different proficiency levels	From all over the world	http://www.cambridge-efl.org/rs-notes/0001/rs-notes1-6.cfm
Longman Learner's Corpus(LLC)	10.000.000 words (and expanding)	Essays and exam scripts at different proficiency levels	From all over the world	http://www.longman.com/dictionaries/corpus/lccont.html

ACADEMIC LEARNER CORPORA

NAME	SIZE	TYPE OF TEXT	LANGUAGE BACKGROUND	WEBSITE AND AVAILABILITY
International Corpus of Learner English (ICLE)	2.200.000 Words	Advanced academic essays, mainly argumentative, by university students	11 languages: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish	http://juppiter.fltr.ucl/ac.be/FLTR/GERM/ETAN/CECL/cecl.html A CD-ROM edition, with an accompanying handbook, will be released by October 2002. C/o S. Granger, Centre for English Corpus Linguistics, Université Catholique de Louvain-la-Neuve Place Blaise Pascal 1, B-1348 Louvain-La-Neuve Email: granger@lige.ucl.ac.be
Janus Pannonius University (JPU)	About 400.000 Words	Advanced essays by university students	Hungarian	http://www.geocities.com/writing_site/thesis/ Portions accessible on line http://www.geocities.com/jpu_corpus.
USE. Uppsala Student English	About 1.000.000 words and expanding	Written academic texts of advanced level	Swedish	http://www.hit.uib.no/icame/ij24/use.pdf
The Polish Learner English Corpus (The PELCRA Project)	500,000 words and expanding	Exam essays from beginning to post-advanced levels	Polish	http://www.uni.lodz.pl/pelcra/corpora.htm Samples available at http://www.uni.lodz.pl/pelcra.samples.htm

A useful distinction made by Pravec is the one between commercial and academic learner corpora. The two commercial learner corpora are the Cambridge Learner Corpus (CLC) and the Longman Learner's Corpus (LLC). They are the largest (many million words and still expanding), as they should represent learner written productions from all over the world, at different levels of proficiency and reflecting both timed examination and untimed practice conditions. They are being collected mainly for in-house use and can be currently accessed by the EFL authors working for these two publishers. However, it is interesting to read their web sites to find out more about their design. Both corpora contain some information on learner backgrounds and variables (e.g. nationality, text type and level of competence), and are partially coded for types of mistake. They allow, therefore, the identification of areas of success and of difficulty, and of recurrent mistakes made by learners from

different language backgrounds. They are meant to improve the quality of teaching materials and pedagogical dictionaries and, in the case of the Cambridge Learner Corpus, also of the Cambridge EFL exams by making the teaching materials more sensitive to the recurrent difficulties of different national language backgrounds.

Four academic learner corpora have been singled out for more detailed presentation, as they have been developed in European university contexts and should be more easily available for research. The four corpora (ICLE, JPU,USE and PELCRA) have been collected from1990 to the present, and some of them are still expanding. They vary in size from over 2 million to 400 thousand words . Size and representativeness are important issues in corpus linguistics: they depend on the purpose of the data collection and on the type of research questions to be asked, whether on more frequent language phenomena (e.g. grammatical) or on more rare ones (e.g. lexical). The texts they are made of are mainly academic essays, often of an argumentative type, but they differ in the students' levels of proficiency, length, topic and writing conditions (timed exam versus un-timed practice; with or without reference tools). Three out of the four corpora refer to a specific European language background, Hungarian, Swedish and Polish respectively. The ICLE is the notable exception because it comprises 11 national contexts, from Bulgaria to Sweden, and the number of countries will be extended further to include extra-European language backgrounds such as Chinese and Japanese. Information about learner backgrounds and variables is provided in all corpora. While the ICLE corpus will soon be available for purchase as a CDROM, the other corpora can be accessed, at least partially, on the net, often by negotiating with the researchers responsible for them. Another important feature of the learner corpora to be further investigated through the web sites is the presence of various types of annotation. For instance, the ICLE corpus can also be tagged for parts-of-speech through the Tosca Tagger software, and several types of error annotation projects are being developed on the various corpora. To analyse the corpora several software programmes can be used, Wordsmith Tools being the most popular at present.

3. The ICLE Corpus: what can we do with it?

ICLE is the only learner corpus so far which allows us both to search 11 individual national subcorpora (Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish) and to carry out extensive cross-linguistic comparison between them. Each subcorpus is made of 200,000 words from academic essays written by students in their 3rd or 4th years of study

in foreign languages and literatures degrees in European universities. The essays are argumentative, between 500 and 1,000 word long, and deal with controversial topics (e.g. Attitudes to crime, abortion or artificial insemination), and in a limited percentage, with literature.

Each text is accompanied by an anonymous learner profile, which includes 21 variables. The most important are gender, age, country, mother tongue, language(s) used at home, years of study of English at school and at university, other foreign languages studied, months spent in English-speaking countries, writing conditions and use of reference tools. These variables will be searchable through a user-friendly interface in the forthcoming CDROM edition. This will allow the researcher to select specific national subcorpora (for instance, the Italian subcorpus or the subcorpus of Romance languages) and to extract and compare different socio-linguistic and learning variables (for instance, the performance of males and females and the performance of two groups of learners who have had different years of study of English). To complement this type of information an accompanying handbook will provide a socio-linguistic and educational profile of each country. The software that is currently used is Wordsmith Tools, which allows the usual operations of word count, concordances and key words search.

Well before its publication, the ICLE corpus was exploited for research by the ICLE research teams, as is evidenced by the more than a hundred titles listed in the web page of the *Centre for English Corpus Linguistics* or quoted in Granger' book *Computer Learner Corpora*. Research papers based on the ICLE corpus deal with grammatical, lexical, phraseological and discourse phenomena; they may take a more quantitative and/or a more qualitative approach; or they may focus on corpus design criteria, theoretical issues and/or pedagogical applications. The number of contributions dealing with computer learner corpora is constantly increasing in corpus linguistics conferences such as ICAME and TALC and in general conferences of applied linguistics such as AILA, and also in ESSE.

The research that has been carried out so far follows three different approaches which answer different research questions and range from the most "local" to the most general, from "ad hoc" pedagogical applications to second language acquisition theoretical issues.

A first type of research aims to collect extensive and systematic evidence of lexico-grammatical errors made by learners from a specific national and language background, e.g. Italian learners. The need for this type of investigation may start from casual observations in a specific teaching context.

To give an example, the writer of this article has collected a small corpus of e-mails addressed to her by her students on several academic matters where the word *informations" was used instead of "information". This is a common mistake for Italian EFL learners at an initial level, since this noun is countable and often plural in Italian. By searching ICLE-IT, it was possible to find out that the mistake was still present in the productions of advanced students of English in 4 out of 21 occurrences of the lemma. (see Table 2).

Table 2. Concordances for "information/informations" in ICLE-IT

1. comes close to us just to ask **an information** because we feel he or
2. vehicle* of personal knowledge **and information**, and it underlines ever
3. of education based on **both information** and discussion of the
4. program is stopped for **commercial informations**. How can we avoid then
5. a look to past times, we can **find information**, read and see a series
6. it was becoming hard to **find informations** between all these
7. of the TV, where you are **given information** you sometimes assimilat
8. but because they have much **more information** than the common citizen
9. first was considered as means **of information** and entertainment and
10. can witness very dangerous leak **of information** and magistrates'
11. *means* for the divulgation **of information**, which marks the advent
12. a powerful means and source **of information**, which contributes to
13. of television, all the methods **of information**: films, soap operas,
14. science among the huge quantity **of informations** that they provide
15. of both lack of time and **proper information**. Moreover, we are

16. the couple' s life, giving **some information**: he explains where the
17. : it takes more. It **takes information**. It takes education. If
18. in the important functions **that information** can absolve: only if
19. also because **the information** about the semen' s
20. common citizens. The use of **the information** those people do have
21. and social life. It gives all **the information** people need. Watching
22. quickly conveyable than **written information**, for example books. In

The rearrangement in alphabetical order of the first word on the left of the search term, made possible by Wordsmith Tools, allows us to spot wrong patterns more easily, like in line nº 1 "* an information". It is interesting to note that, if we carry out the same search in the Dutch corpus, we do not find one single mistake out of 107 occurrences of "information".

Another example refers to the fact that many Italian advanced students of English use "even if" instead of "even though", thus blurring the prototypical semantic distinction between hypothetical and concessive clauses. The search in this case proved to be considerably longer, as it was necessary to expand the one line concordance into more extended context, which Wordsmith allows us to do, as can be seen in Table 3.

Table 3. Selected expanded concordances for "even though/even if" from ICLE -IT

1. **be more sense of duty and responsibility among parents**. Today the increase in crime is one of the greatest problems **even though** the police try to prevent law-breakings suggesting safety measures such as controlling schools, traffic jam in

2. expensive cars and beautiful houses. In other words they prefer their economic stability as well as their economic power **even though** this leads to a lack of responsibility towards their children. As long as a child receives the early education at

.....

1. has increased tremendously. Teenagers start at the age of ten or seventeen burgling, robbing friends and neighbours **even if** they belong to a middle class family. Youth courts have suggested, in order to solve this horrible society plague,

2. to ask only to give help to social services in the payment of juvenile crime. Children will continue to commit burglaries **even if** their parents will be obliged to pay for their offences. Families have the right to look after their sons and to

After consulting pedagogical and descriptive grammars, native informants, and other English native corpora (in particular, the *Louvain Corpus of Native English Essays* (LOCNESS), a comparable corpus of American and British academic essays), what was considered a very clear-cut distinction was considerably more "fuzzy" than expected, also across the British and the American varieties of English, and that some of the uses that were initially marked as wrong were instead to be considered acceptable.

Another example is a lexical investigation made on 5 lexical items belonging to the semantic field of "work" by Aurelia Martelli at the TALC, 2000. The items *work, job, employment, occupation, career* were frequently used in ICLE-IT as many essays dealt with unemployment and the difficulty of finding a job for young people. The search allowed to observe that many Italian students seemed to consider these words as interchangeable synonyms(e.g. "*to look for a work" instead of "to look for a job"). This resulted in the production of unacceptable or unidiomatic collocations, often related to mother tongue transfer such as " *to make a job" instead of "to have a job" or "*to learn a job" instead of "to learn how to do a job". In particular, the comparison of the uses of "career/careers " in ICLE-IT and in the already quoted *Louvain Corpus of Native English Essays* (LOCNESS) has revealed some interesting differences (see Table 4). By observing the concordances of ICLE-IT one can notice the underuse of idiomatic collocations, such as "to pursue a career" or "a career woman" and the presence of collocations that are influenced by Italian usage (e.g "*making career" that corresponds to the Italian "far carriera"). One can also observe in the Italian corpus the prevailing of a negative connotation ("career" often collocates with "dangerous" and "criminal"), which is less evident in LOCNESS (an exception is "career trap"). The negative connotation of "career" in ICLE-IT may be due to cultural factors, since in Italian the corresponding lexical item (*"carriera"*) often implies the idea of individual competitiveness, and the loss of more humane

values, which is considered negative especially for women. In Italian, the expression "*fare carriera*" is often ironical and suggests that success may be achieved through lack of fairness and even bribery.

Table 4. Concordances of "career/careers" from ICLE-IT and from LOCNESS, reordered in alphabetical order according to the first word to the left of the search word .

ICLE-IT

1. the starting point for a *__criminal career__* . That is why I consider
2. of dangers and risks that such __a career__ could imply, F. Shashkova
3. the cultural basis to build up __a career__, it focuses only on
4. decide to devote their lives to __a career__, rather than their family.
5. social status through a job and __a career__ were deliberately stopped by
6. women to choose between family __and career__ is still alive. For women on
7. concern is to have a __brilliant career__ and to have the freedom to
8. an important place or a __brilliant career__ in a society dominated only
9 encouraged to go on in their __crime career__. Certainly the problem of
10. and philosophies of a __criminal career__. For the criminal behaviour
11. boys who started their __criminal careers__ when they were 12 and 10
12. two different sport, __dangerous careers__ and after having taken into
13. Moreover women with __dangerous career__ force their children to live
14. need to pursue a __dangerous career__ in order to prove her
15. give up their previous __dangerous careers__ in order to assure to the
16. mothers to renounce to __dangerous career__ in the world of sport. The
17. the right to pursue a __dangerous career__ regarded a man that was a

18. their children a more **dignified career** than that of the criminal.
19. example, a mother that lives **for career** and work; probably this
20. time for herself, or even a **good career**. In the long term, this can
21 the same time to continue with **her career**, even when she takes risks
22. not necessarily renounce to **her career** in a dangerous sport is that
23. because he' s too worried with **his career**, his power, his politics,
24. plans, which could be: **making career**, getting married, having a
25. all her own interest and her **own career**. A mother continues to be a
26. have the right to pursue their **own career**, even if it dangerous, and
27. by the wires of money, **power**, **career**, success and personal
28. sport represent their **professional career**, and as all the people that
29. men, as for example the **religious career** or airplain pilots (to
30. contributes to the **scholastic career**. Being a relatively young
31. dedicate themselves to **successful careers** in management, politics,
32. more interested in a **successful career** than getting married and
33. students into the **teaching career**. People are persuaded that
34. mother and the realization in **the career**, in particular when she is
35. as in the sixties, climbing up **the career** ladder in the eighties and
36. because women have reached **the career** ladder. Yet, some of them
37. children are interested in **the career** of crime? If children are
38. woman should be free to pursue **the career** she has chosen. First of all
39. today to find a woman who has **the career** she wanted and the family

40. about the options they make in **the career** versus their family. The
41. for them: to continue with **their career** and consequently to renounce
42. majority of cases, started **their careers** many years before to have a
43. children, should forfeit **their careers** when they are too dangerous
4. at the age of 12 has started **this career** of crime that begun shoplift

LOCNESS

1. childcare and housework over **a career**, and few men want to risk
2. A woman today, if she pursues **a career** as expected by her peers,
3. who returned to college and made **a career** for herself. Those outside
4. that they may one day practice **a career** in the field of their major.
5. mother and disregard or postpone **a career**). Is it a harder life to for
6. I have grown up in an age where **a career** man and a career woman have
7. satisfaction of pursuing **a career**. Many women followed this
8 , and if a woman wants to pursue **a career**, she should be allowed to do
9. they did not particularly want **a career** still tended to stay in
11. hole (not be a mother but choose **a career** vs. be a mother and
12. in an age where a career man and **a career** woman have become
13. prestige that usually accompany **a career**. Yet if a woman decides on
14. ill young enough to choose **another career** if he failed. What if Ralph
15. he same generation of newly **formed career** moms. I remember my mother a
16. He had a good life. He had a **good career** in a fortune 500 company. He
17. children are in school, begins **her career** later in life? It seems to

18. can guarantee a **high-ranking career** and and accelerated
19. becoming a corporate lawyer? **His career** might not have been as
20. guilt and can ruin a **journalist's career**. A judge asking a journalist
21. a baseball player. He later **made career** choices that increasingly to
22. they can pursue a **meaningful career**, relationship, or anything
23. and contented himself with a **party career** in journalism. He was not
24. of higher education, and **pursue careers** without being judge as bad
25. Without patronage, **several careers** related to the arts profess
26. r future. She establishes a **sewing-career**, re-unites with her children
27. **the career** take a minor role or vice
28. are the only ones caught in **the career** trap. There is a lot of
29. who works outside the home. **The career** with a capital C is
30. witnessed the evolution of **the career** woman who worked from 9 to
31. they were ready to interrupt **their careers** for a family. This is not
32. ended up dissatisfied with **their career**, their marriage, and a good
33. off having children until **their careers** were established, only to
34. train an elite for future **top careers** in administration, industry
35. chastising these women for **wanting careers** outside the home. Nowadays,

From small-scale *ad hoc* searches like the ones presented above, one can move to larger and more systematic group projects of interlanguage and error analysis, which should be at the base of the development of more effective syllabuses and teaching/learning materials. Two examples will be given. One comes from Asia and was started in Hong Kong at the Hong Kong University of Science and Technology and the TELEC Centre for Teachers of English Language Education. Large-scale investigations of the areas of difficulty of

secondary school students of Chinese(Cantonese) mother tongue background were carried out on the two large learner corpora, mentioned by Pravec (2002) (the HKUST and TSLC corpora). These projects refer to the needs of those specific institutions, but they offer useful models in order to exploit learner corpora for pedagogical purposes. Several corpus-based EFL tools have been developed such as a programme for electronic composition, an interactive on-line English tutorial for students and a network of grammatical explanations and graded exercises to be used by secondary EFL teachers (cf. Milton, 1998).

A similar project is being developed on the ICLE corpus through the use of an error-tagger software. The aim is to identify and tag lexico-grammatical problem areas in several national EFL varieties, and then develop an electronic pedagogical tool that will deal with both common core and L1-specific mistakes, by providing both concordance-based exercises and grammatical explanations.

Such pedagogical applications will answer the practical needs of remedial work with large numbers of students of English and are in character with the growing tendency to encourage learners' autonomous learning, often through on–line resources.

A second type of research carried out on ICLE aims to identify areas of "foreign- soundingness" or "non-idiomaticity", that is, learner language which, even though not incorrect in itself, does not sound "natural" or appropriate in register. This can be achieved by comparing a non-native and a native corpus, thus discovering phenomena of "overuse", "underuse" and even "avoidance" of particular language features and stylistic choices (the already quoted LOCNESS, has often been used as the native control corpus).

Several examples of this type of research can be quoted. For instance, Ringbom (1998) has carried out a search on vocabulary frequencies in advanced learner English in seven different nationalities. When he observes high-frequency main verbs (i.e. *think, get, make, become, want, take, find, know, use, go, live)*, he discovers that some are systematically overused by non-native speakers (e.g. *think, get, find, want,* know) while others are underused (e.g. *use , believe and feel)*. These quantitative findings may be the starting point for more qualitative considerations, such as some general features of learner language, which is described by Ringbom and others as "dull, repetitive and verbose "or "more informal in register" An example of inappropriate register is the very frequent use of "I think" in academic essays on the part of foreign learners of different nationalities.

Another example comes from a key area in argumentative essays, which Petch Tyson (1998) and Migliorero (1999) call "writer/ reader visibility". By searching several linguistic features which should reveal the presence of the writer or the reader (e.g. first person singular and plural reference, second person reference, verbs of mental processes, expressions used to monitor the information flow , fuzziness words, emphatic particles, reference to situation and discourse markers), it has been found that the presence of reader/writer in non-native essays is more overtly coded than in the native speakers essays. It is therefore possible to draw the conclusion that a higher degree of interpersonal involvement in argumentative essays is a cross-linguistic feature, which appears to differ from the prevailing native use.

It should be noticed that some of the types of research reported follow an hypothesis- finding data-driven procedure while others use already existing linguistic categories and a more data-based approach. Research also moves from quantitative comparisons (through frequency word counts) to more qualitative considerations, by observing language in context through concordances. Concordances are particularly useful in exploring the area of phraseology and the phenomena of recurrent collocations. Both teachers and students can take advantage of this type of research and compare learner productions with native and, hopefully, more idiomatic models. This comparison can be the starting point for what is called a data-driven approach, whereby the learners are encouraged to work out rules and idiomatic patterns by being exposed to non-native and native usage.

A third and more ambitious type of research aims to find answers to the always open question of how foreign language learning takes place and by what variables it is affected. Learner corpora provide the data to explore well-known learner phenomena such as the types and the causes of mistakes, whether related to L1 transfer or universal learning strategies, or the more covert phenomenon of avoidance of those structures and uses that are considered difficult by learners. The ICLE corpus offers the opportunity to compare 11 different national corpora and to extract a variety of learning variables. Other corpora, the longitudinal ones, may allow to verify some findings of second language acquisition research such as the order of acquisition of English grammatical morphemes, as can be seen in Tono (2000), or the presence of developmental patterns.

However, along with the pedagogical usefulness and the theoretical potential of learner corpora, some disadvantages and methodological dilemmas of learner corpora should also be mentioned. First of all, and very obviously, this type of research and pedagogical application requires an amount of technological

resources not all universities can afford. The amount of computational and statistical expertise is an important educational issue to be taken into account in planning new curricula for the training of applied linguists. Secondly, the representativeness and the validity of the corpus for specific research questions should always be kept in mind and the danger of over-generalising findings should be avoided. In this respect the present prevailing of written over oral corpora should be overcome. Besides, the choice of native models to compare to non-native corpora is a hotly debated issue. In the case of argumentative academic essays, should one favour a comparable native learner corpus, like LOCNESS, or an expert/professional corpus or, in any case, an edited corpus, such as one of editorials in quality newspapers and periodicals, be they written by native speakers or by non-native proficient users of English as a *lingua franca*?

A final problem has to do with the use of learner language. Should it be used only by teachers or also presented for comparison to learners in the so-called data-driven mode? The idea of introducing **native corpora** in EFL teaching and learning has been gaining ground in the last few years. The idea of using **learner corpora** is more recent and perhaps more difficult to assimilate because of the risk of reinforcing wrong habits in learners. The presentation of unedited learner language should indeed be supported by corrective feedback and cognitively satisfactory explanations. Granger and Tribble (1998) argue in favour of the role played by form-focused instruction within a communicative approach, and give many convincing examples of the usefulness of comparing native and non-native speakers' concordances to stimulate language awareness and learning by discovery.

4. To buy or to build a computer learner corpus?

Building a computer corpus is a time consuming enterprise and requires the solution of linguistic, statistical and computational problems. It is therefore a good idea to start by using one of the existing corpora, if they apply to one's specific teaching and research needs. It may also be advisable to join one of the existing projects (ICLE, for instance) and extend its scope to other nationalities or levels of competence by following already existing guidelines.

However, building new corpora is also necessary in order to update the existing ones, to cover other registers, genres and learning contexts. It is very important that applied linguists decide to embark on new learner corpora projects to answer new and different research questions and needs. To this purpose Granger's book on computer learner corpora is a very clear and comprehensive introduction ranging from the compilation to the use of learner

corpora, from theoretical debate to pedagogical applications. Some future developments that would fill some of the gaps in the existing panorama are the following:

1) to collect spoken learner corpora, a type of corpus of which there is shortage;
2) to develop a corpus for a new national group of EFL learners, to compare with other already existing corpora
2) to collect a longitudinal corpus to study the development of the same learners'—or of different learners' — interlanguage through time;
3) to build bilingual corpora, such as a corpus of English-Italian argumentative essays, to compare the types of rhetoric used in the mother tongue and in the foreign language;
4) to develop a corpus of a specialised genre, such as scientific reports or business letters, to suit the students' learning needs.

5. Learner corpora, Second Language Acquisition (SLA) Research and EFL teaching

Is the development of computer learner corpora a revolution which will upturn the foundations of SLA and EFL research and practice , or is it a powerful methodology to be used in conjunction with other traditions of Applied Linguistics research?

A similar discussion is going on among corpus linguists on the pros and cons of the "computer linguist" versus the "chair linguist". This sharp contrast should be overcome. Computer learner corpora provide large amounts of interlanguage data for analysis and observation, thus avoiding the limitations of small-scale tightly controlled elicited experiments on the one hand and vague and impressionistic considerations on the other. This empirical perspective should be welcomed by all those who work in EFL teaching and research because it may counterbalance the role of intuition and introspection in SLA research. Learner corpora have encouraged scholars to reconsider the contribution of Contrastive Analysis (CI) by establishing what Granger calls "Contrastive Interlanguage Analysis" (CIA), which is not the traditional comparison between different language systems but between what native and non–native speakers of a language do in a comparable situation. Learner corpora offer tools to extend Error Analysis by focusing not only on mistakes but also on general features of learner language, such as avoidance, overuse or underuse of specific language

choices. Learner corpora are made of data that are produced in a classroom context and under a teacher's control, thus complementing the tendency of much SLA research to observe natural types of acquisition.

However, the complexity of analysing interlanguage data, especially at discourse level, should not be underestimated. Convincing explanations for learner data require, in my opinion, the use of several sources , such as learner introspection, teachers' intuitions and broader contextual and cultural information. Different approaches should be experimented and combined such as hypothesis-finding and hypothesis-checking searches, data-based and data-driven approaches, raw and tagged corpora, quantitative and qualitative analyses, attention to the individual learning process and to group performance, machine versus manual approaches, product and process perspectives, synchronic and diachronic approaches, textual and contextual considerations

6. *Conclusion*

Computer learner corpora are a very interesting and fast developing branch of corpus linguistics. They offer rich empirical data for the study of language acquisition and learning. It is to be hoped that the different research communities of corpus linguistics and SLA/ELT studies would take a convergent rather than a divergent course.

7. *Selected references.*

Granger, S. (ed) 1998. *Learner English on Computer*. London: Longman

Granger, S. and C. Tribble.1998. "Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning". in Granger (ed) 199-209

Martelli, A. forthcoming. "Lexical errors in EFL writing: a corpus-based approach". poster at the TALC 2000, Graz

Migliorero, S. 1999. "Interpersonal involvement in argumentative essays: a corpus-based analysis of Italian EFL and ENL students". Unpublished MA dissertation, University of Torino, Italy

Milton, J. 1998 . "Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment" in Granger (ed) 186-198

Petch Tyson, S. 1998. "Writer-reader visibility in EFL written discourse" in Granger (ed) 107-118

Prat Zagrebelsky, M. T. forthcoming. "Even if and/or even though ? A corpus-based investigation of written learner productions in the Italian subcorpus of the International Corpus of Learner English". Poster presented at ICAME 2001, Louvain-la Neuve, Belgium

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

Ringbom, H. 1998. "Vocabulary frequencies in advanced learner English: a cross-linguistic approach" . in Granger (ed) 41-52

Tono, Y. 2000 "A computer learner corpus based analysis of the acquisition order of English grammatical morphemes". in Burnard and McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang

For more references see the web page of the *Centre for English Corpus Linguistics* (CECL): http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html