

## Corpus linguistics, language learning & ELT: interviewing Sylviane Granger [Université Catholique de Louvain, Belgium] Vander Viana [APLIERJ / CAP-UFRJ / Ibeu / PUC-Rio / Redes]

This issue's interviewee, Sylviane Granger, holds an MA in English and Dutch and a PhD in English Linguistics, both from the Université Catholique de Louvain (UCL), Belgium. She is Professor of English Language and Linguistics, Director of the Centre for English Corpus Linguistics and Director of the Institute of Languages at the UCL. Her research interests include Corpus Linguistics, Second Language Acquisition, English as a Foreign Language, Lexicology and Lexicography, Contrastive Linguistics and Translation Studies, Language Testing, e-Learning and Natural Language Processing. Dr. Granger has edited several books and written more than 50 articles and book chapters.<sup>1</sup> She is known for her extensive research in Corpus Linguistics, especially in learner corpora. In this interview, she gives a brief historical overview on Corpus Linguistics and assesses the relationship between introspection and corpus analysis. She also explains how corpora can be used to help understand a given language and encourages teachers to use corpora in their classrooms. At the end of the interview, she kindly offers some useful websites to those teachers willing to learn more about Corpus Linguistics.

<sup>1</sup> A complete list of her publications can be found at <http://cecl.fltr.ucl.ac.be/publications%20by%20author.html#sylviane>.

**After the Chomskyan revolution, language description was carried out mainly by means of introspection. Nonetheless, it seems that corpus-based methodologies are becoming mainstream in the field of language studies. What would you credit this shift to?**

Some major linguists played a pioneering role in what can truly be called the corpus revolution. Among these are the compilers of the first two major computer corpora of English: Nelson Francis and Henry Kucera for the Brown corpus and Stig Johansson and Geoffrey Leech for the LOB corpus. Both corpora were built in the 60s/70s on the same design principles and contained 1 million words of writing representing a range of text types in American English for Brown and British English for LOB. These two corpora have given rise to a large number of descriptive studies of morphology, lexis, grammar and discourse, which have provided much more accurate and reliable information about language use than previous intuition-based studies. And in fact, nearly 40 years on, the Brown and the LOB are still being used, often in comparative studies with their current-

day replicas, the Frown and the FLOB. The heuristic power of corpus linguistics is highlighted by Geoffrey Leech (1992) who considers that it is a new way of thinking about language which is challenging some of our most deeply rooted ideas about language. One of the reasons why corpora are such wonderful discovery tools is that, as John Sinclair (1991) says «the language looks rather different when you look at a lot of it at once».

If the use of corpora had been confined to scientific studies published in international journals, corpora would not have had the wide impact they have turned out to have. But quite early on, major publishers started using corpora to inform their grammars and dictionaries. At the beginning, they simply used them as a way of illustrating grammatical rules or definitions with authentic examples but later, more radically, they used frequency information gained from corpus analysis as a

means of selecting and ordering the linguistic items to be included. The most dramatic developments have clearly taken place in the field of lexicography. Frequency has progressively superseded etymology as the main ordering principle and the amount of information provided has increased exponentially. Lexico-grammatical patterns have always figured prominently in English monolingual learners' dictionaries, but the systematic use of corpus data has allowed dictionary-makers to include a lot more information, especially in some of the trickiest aspects of language like collocations.

Coming back to your question, I would like to add that in my view one should not present introspective and corpus-based methods as opposed. For one thing, the two approaches do not really tap the same thing. But also, and more importantly, it is both impossible and undesirable to do away with introspection altogether. Corpus linguists need to keep a very critical mind as corpora, even native corpora, contain errors.

### **In what ways does Corpus Linguistics research help understand more about a specific language?**

In countless ways! Corpora can help us understand more about almost every aspect of a language. But I think there are three areas where they make a particularly significant contribution and those are frequency, variation and lexico-grammar.

#### *Frequency:*

As corpora are stored electronically, we can analyse them with special corpus software tools to gain access to a wide range of quantitative information on language. On the basis of corpora, we can easily find out whether a word or structure is frequent or not in any particular variety of English. In other words, we can distinguish between what is probable and improbable in language, two notions which contrast with the previous exclusive emphasis on the possible vs. impossible distinction. Frequency is important in an applied perspective because time is of the essence in teaching: there is never enough of it! Teachers just cannot afford to spend a lot of time on infrequent phenomena. I remember when I was a student, we spent quite a bit of time learning highly infrequent structures, such as the use of 'lest'

after the verb 'fear' (She fears lest she should lose courage). It is important to bear in mind, however, that frequency is only one of a number of factors that need to be taken into account when selecting items for inclusion in a syllabus.

#### *Variation:*

The availability of a wide range of corpora representing different genres and/or regional varieties of English has enabled linguists to chart the similarities and differences between the different varieties of English. These studies are very important for teachers who need to teach the language variety that relates to the course objectives and needs of the learners. If the course objective is oral communication, then those grammatical phenomena that are typical of formal writing become irrelevant. For example, non-defining relative clauses are highly infrequent in speech and devoting time to the tricky defining/non-defining relative clause distinction in an oral communication course would thus be futile. That time would be better spent on teaching the functions of discourse markers, such as 'well', 'you know', 'I mean' and 'you see', which play a major role in oral communication. The *Longman Grammar of Spoken and Written English* (Biber et al 1999) is a mine of information here as it focuses on how grammar varies with register.

#### *Lexico-grammar:*

Using text retrieval tools like concordancers, we can easily extract the recurrent lexico-grammatical patterns of any search item. We are currently using this technique in Louvain to uncover differences in the way advanced learners of English and native speakers use EAP (English for Academic Purposes) vocabulary. Concordance-based analyses reveal interesting differences that can feed into pedagogical material. For example, the patterns emerging from a careful analysis of all the occurrences of the verb 'discuss' show that native speakers mainly use the verb in the passive ('it will be discussed more fully below' or 'as discussed in chapter 3, ...') whilst learners show a predilection for active structures with first person subjects ('I would like to discuss/I will discuss...'). In addition, the corpus search has shown that many students do not realize that 'discuss' is a transitive verb ('politicians discuss \*about this problem').

**Being responsible for the ICLE (International Corpus of Learner English) project, what would you say are its main advantages? Are there any drawbacks? What are its future developments?**

The main advantage of the ICLE is that it has been compiled on the basis of strict design criteria. We haven't just taken any old student data that we had lying about on our desks. We applied very strict criteria in order to compile texts that were comparable across a range of learner and task variables. So one major advantage is that the corpus is fairly homogeneous. Another advantage is that it is well documented: we know a lot about each of the texts as learners were requested to fill in a fairly detailed learner profile questionnaire. A third advantage is that we cover a wide range of mother tongue backgrounds. At the beginning of the project, we only had data from our own French-speaking learners of English. Today the project covers no less than 19 mother tongue backgrounds! This allows us to assess whether a difficulty is specific to one learner population or more generalised. Yet another advantage is that the corpus contains data from higher intermediate to advanced level students. This is a proficiency level that we know very little about as previous studies have tended to focus on beginners or lower intermediate students. Analysis of the data helps us identify the aspects of language that still need to be addressed past the lower intermediate stage.

More generally, the major advantage of a learner corpus such as the ICLE is that it gives us access to learners' attested difficulties, which often turn out to only partly match teachers' intuitions. The major drawback of the ICLE is that it only targets one specific category of learners performing one specific task: essay writing. We need other corpora representing different proficiency levels and/or tasks (narratives, letter writing, picture description, etc.). In addition, the ICLE only contains written data, but its oral counterpart (LINDSEI) is currently under development and as it targets similar-type learners, it will enable very interesting comparisons to be made between learners' oral and written performance.

The ICLE is a dynamic project: although we have already released a CD-ROM containing data from 11 different mother tongue backgrounds, we are

preparing the second version of the CD, which will contain more data and an in-built concordancer. We have new teams joining us all the time, which means that the project will keep me busy until my retirement age and beyond! On a personal note, I would like to say that one of the nicest aspects of the project is that it is a joint project involving many teams internationally. It is very enriching, at a scientific, cultural and personal level, to work closely with researchers from so many different countries.

**How can English teachers around the world take advantage of research concerning the mapping of English? Do teachers have to wait for such results to be incorporated in teaching materials?**

It would be a pity to wait, as it will take quite some time before corpus data are fully exploited and integrated into teaching materials.

In my view, all teachers would derive tremendous benefit from using corpus tools and data to prepare

Make the **smart choice** for your school, your teachers, and your students.

**first CHOICE** **smart CHOICE 1** **smart CHOICE 2** **smart CHOICE 3**

Smart Choice — the new 4 skills American English course with a free student's Multi-ROM and teacher's Click & Change CD-ROM at every level.

*How do you teach?*  
find out at [www.howdoyouteach.com](http://www.howdoyouteach.com)

For more information contact:  
Tel/fax: (21) 3154-7206  
e-mail: [rjes.br@oup.com](mailto:rjes.br@oup.com) - [www.oup.com/elt](http://www.oup.com/elt)

**OXFORD**  
UNIVERSITY PRESS

their courses and/or correct students' work. And the good news is that there are many excellent tools and corpora available free of charge on the web. For example, on the Virtual Language Centre website (<http://vlc.polyu.edu.hk/default.htm>) there is a wide range of corpora in English, French, Chinese and Japanese, which you can search with a web concordancer (<http://vlc.polyu.edu.hk/concordance>) and then use to create corpus-based exercises. Teachers can also download the ConcApp concordancer free of charge, install it on their computers and use it to analyse their own native or learner corpora. As students usually use a word processor to do their writing assignments, it is not difficult for teachers to collect a small corpus of their own students' work. Of course these corpora will be small but the data are highly relevant as they come from the students themselves and not from some other learner population. Computer-mediated communication (CMC) is also a good source of data.

### **Would you advise English language teachers to make use of Corpus Linguistics tools in their classrooms? How can they do so?**

One very simple way of 'inviting' the corpus into the classroom is simply to have one computer in the classroom and use it in the same way as one would use a grammar or a dictionary. If a problem crops up, (eg. do you use 'increase in' or 'increase of'? In what kind of context does one use the word 'person' in the plural?), the teacher can ask a student to look up the word in a corpus. This has two major advantages. First, it empowers the students, doing wonders for their motivation. Second, it is a great help for teachers, most of whom are non-native teachers of English and quite understandably, do not always have the answer to all usage-related questions.

Another way of including corpus data in teaching is to prepare materials and ask students to do some corpus work either online or offline in preparation for the course. Good examples of data-driven learning (DDL) of this type can be found on Passapong Sripicharn's website: [www.geocities.com/tonypgnews/units\\_index\\_pilot.htm](http://www.geocities.com/tonypgnews/units_index_pilot.htm) or Tim Johns' website: [www.eisu.bham.ac.uk/johnstf/ddl\\_lib.htm](http://www.eisu.bham.ac.uk/johnstf/ddl_lib.htm).

Concordancers are easy to master (at least the basic functions). After a few training sessions in the computer lab, all students should be able to use a concordancer to retrieve words or phrases and examine their recurrent patterns of use. It is important to note however that corpus-based exercises are only one type of exercise among many. Not all learners like them because they require an inductive approach that some learners find difficult. Data-driven exercises also take a lot of time. We badly need validation studies and for this we need teachers willing to try out corpus-based methods with their own students. Teachers are the only ones who can tell us how best to incorporate corpora into teaching. Crucially however, for teachers to be able to integrate corpora into their teaching, they themselves need to be able to use them and yet most teacher training programmes still have no corpus module. This is a great pity as our own experience shows that when teachers see the tools and are trained in using them, they do so enthusiastically to the point of wondering how they could have survived without them before.

#### **Some useful websites**

- Corpus Linguistics (David Lee's website)  
<http://devoted.to/corpora>
- Information and communications technology for language teachers  
<http://www.ict4lt.org/>
- The Compleat Lexical Tutor (Tom Cobb's website)  
<http://www.lextutor.ca/>
- Centre for English corpus linguistics (contains a comprehensive learner corpus bibliography)  
<http://cecl.fltr.ucl.ac.be/>

**Sylviane Granger** is Professor of English Language and Linguistics at the University of Louvain (Belgium). She is the director of the Centre for English Corpus Linguistics, where research activity is focused on the compilation and exploitation of learner corpora and bilingual corpora. In 1990 she began to compile the *International Corpus of Learner English*, which has grown to contain learner writing by learners of English from 19 different mother tongue backgrounds and is the result of collaboration from a large number of universities internationally. She has written over 30 articles and (co-)edited several volumes on these topics and gives frequent invited talks, seminars and workshops to stimulate learner corpus research and to promote its application to ELT materials design and development. Her publications include *Learner English on Computer* (Granger ed. 1998), *Computer Learner Corpora*, *Second Language Acquisition and Foreign Language Teaching* (Granger, Hung & Petch-Tyson eds 2002) and two forthcoming volumes on phraseology: *Phraseology: an interdisciplinary perspective* (Granger & Meunier eds) and *Phraseology in language learning and teaching* (Meunier & Granger eds).