ENTREVISTA CON ... / AN INTERVIEW WITH ...

Charles F. Meyer
University of Massachusetts, Boston

by Mª Carmen Pérez-Llantada Auría
Universidad de Zaragoza

Corpus-based research in LSP: current trends and future prospects

Introduction

For the past decades, there has been an extensive use of large-scale corpora for developing quantitative and qualitative approaches to Languages for Specific Purposes both in Spain and worldwide. Concurrently, current LSP studies have also become confident in the use of small-scale corpora for addressing more specific research questions from multifarious theoretical standpoints. In addition, LSP researchers are also showing a growing interest in developing contrastive analyses of corpora thus shedding fruitful light on language variation across languages, registers, genres and academic disciplines.

From a pedagogical viewpoint, electronic collections of texts have allowed LSP practitioners to bring to the classroom real data of language use, to enquire into language teaching priorities and, accordingly, to offer more adequate input to students' specific competencies and needs. From a pedagogical viewpoint, another flourishing area of corpus linguistics which deserves special attention comprises those studies that by focusing on analyses of learner corpora attempt to bridge the gap between language research and language pedagogy. No doubt, these also stand as a promising field for corpus research in the LSP arena.

But perhaps the best way of acknowledging the importance of corpus linguistics and corpus methodology in the field of LSP is by seeking advice from one of the most outstanding representatives of corpus linguistics in the US, Professor Charles F. Meyer. Professor Meyer was coordinator of one of the major corpora projects, the International Corpus of English (ICE) and is particularly interested in using the World Wide Web as a linguistic corpus to enquire into real language behavior and language usage. In June 2005, Professor Meyer visited the University of Zaragoza to lecture on the scope and current trends of corpus linguistics. There, I had the opportunity to interview him and to profit myself from his insights into the current trends and future prospects of corpus studies. I am confident that the readers of *Ibérica* will find in this interview interesting and motivating views on the advantages of engaging in corpus methodology –either for LSP research or teaching purposes.

Carmen Pérez-Llantada Auría
Universidad de Zaragoza (Spain)

The interview

In your book *Corpus Analysis: Language structure and language use*, you state that North American symposia, every year, are a real testament of the progress being made in recent years in developing new corpora. I'd like to start this interview by asking you to comment on your research profile as a corpus analyst.

I began doing corpus analysis in the early 1980s when I was working on my doctoral dissertation. I did a thesis on the use of punctuation in the Brown corpus, and I looked at the kinds of linguistic structures that mark punctuation. For instance, I was interested in whether a clause initial adverbial like *therefore* was always followed by a comma. I found that in general while conjunctive adverbials such as *therefore* tended to always be punctuated, adjuncts such as *yesterday* or *here*, which are more integrated in the clause and express notions such as time or space, tended to be punctuated less frequently. At that time the Brown corpus existed only on computer tapes, so I worked with a programmer who would write programs searching for various strings, such as all instances of *therefore* followed or not followed by a comma. I'd then get huge piles of print outs that I literally had to go through by hand and analyze.

As the years passed and computer programs got more sophisticated and corpora got more widely available you could actually analyze your own computer evidence. It then became much easier to analyze grammatical constructions. The most recent project I worked on involved the analysis of gapped coordinations. These are constructions like "we like hamburgers and our friends fish" where some constituent, in this case *like*, is deleted in the second clause, literally creating a 'gap' in structure. I collaborated on this study with Hongyin Tao of UCLA. We used ICE-GB [the British component of the International Corpus of English] and were instantly able to find examples of gapping because this corpus is completely parsed. We wrote search algorithms that recovered all instances of gapping that also enabled us to study which genres they occurred in. In the past it would have taken hours to find what we were able to locate in a matter of minutes. It's amazing how far corpus linguistics has progressed in a matter of 25 years.

Contrastive analyses of corpora primarily rely on genre theory and, more specifically, on a genre-based sociorhetorical slant that has often found in Halliday's functional systemic theory a powerful interpretive framework. Where does your functional systemicist view of language come from?

Well, my undergraduate degree was in generative grammar and you know you read Chomsky and you analyzed sentences you invented for particular constructions and so forth. I met Sidney Greenbaum when I was doing my undergraduate degree who at the time was working with acceptability and I remember in syntax class his talking about extraposed constructions in English. These are sentences like "It is likely that we will eventually succeed" that in generative theories are derived from sentences like "That we will eventually succeed is likely." I was always puzzled by the unextraposed constructions like "That we will eventually succeed is likely" because who ever says that? I learned from Sid Greenbaum that English and other languages have this principle of end-weight, you know, you put heavier constituents at the end of the clause rather than at the beginning. This got me interested not just in the structures themselves but their usage as well and Halliday's theory of functional grammar, although I don't do strictly speaking systemic grammar. But I was interested in Halliday because it seems he had a theory that accounted for both the grammar of language and pragmatics; it was a theory not just of structure but of use. I found that very appealing, especially his three meta-functions: the ideational, the interpersonal and the textual functions of language. This theory also led me to begin using corpora as a means of explaining usage.

Corpus studies seem to represent a trend borrowed from the analytical procedures of scientific and experimental disciplines. Scientists, physicists, mathematicians or engineers rely on data to test their hypotheses and reach conclusions. In your opinion, does corpus methodology echo the scientific method of these disciplines?

I think that to an extent it does. When I was first interested in acceptability, Sidney Greenbaum was doing a series of experiments in which he would ask people to judge the acceptability of sentences. So this is definitely a more direct use of subjects in the sense that psychologists and sociologists do. With a corpus, you're sort of dealing with the end product of what people do with language and there are both advantages and disadvantages to that. The advantages are you get a real look at language use. The disadvantages are you might miss a few things that might and do occur. But I think analyzing a corpus is analogous to what is done in other disciplines in the sense that what you're doing is empirical: you are looking for some kind of evidence to support your generalizations. I think it's in line with the notion in science that you need to follow the scientific method to obtain valid and reliable results. Chomksyan linguists would disagree, saying that what linguists do has more of a parallel with physics, where you don't always investigate observable phenomena. I heard Chomsky speaking recently and he commented that real language is sometimes so messy that it's better to ignore it and just go with your intuitions.

My next question is about the relationship between corpus analysis, language structure and language use. How is this relationship reflected in both theoretical and applied linguistics? Or, in other words, would you advise LSP researchers to address their empirical research questions bearing in mind their pedagogical implications?

When you conduct a corpus analysis, you learn what's going on in language. That's the theoretical part. And once you know what's going on, you can, for instance, be a better language teacher because so many language texts are not based on how people really speak. I mean, one of the chapters of a book I recently edited deals with how French textbooks focus on many constructions that are never really used by native speakers of French. I think the corpus based approach can bring you closer to teaching the kind of structures that people really use. And this methodology is more in line with the communicative method of language teaching, although that method sometimes has no sort of formal aspects—if people just talk, many theorists argue (particularly in the United States), they'll eventually learn to speak a second or foreign language. In my estimation, corpus linguistics brings more rigor to the communicative method of language teaching.

When building up a corpus for research/teaching purposes one should first assess the validity and suitability of the data collection with regards to one's own research/teaching interests. You were coordinator of one of the most important corpora, the *International Corpus of English*. Can you briefly describe ICE, its origins, the scope of the corpus, its main research target(s) and the advantages it provides for research and pedagogical purposes?

Right, well, the ICE corpus was an attempt to have research teams from various parts of the world where English has some official status compile comparable corpora of their national varieties: one million words of spoken and written English representing various genres, such as spontaneous dialogues, telephone calls, newspaper articles, editorials, fiction, academic speech and so forth. The idea was that if you could assemble comparable corpora you can compare them and find differences between various national or regional varieties of English. It turns out that a one million corpus isn't really that large. So if you want to look, for instance, at vocabulary differences, you can find some things but not a lot. It turns out that the various ICE components were most useful for looking at were differences in grammatical structures, things like that. The next version of the British component will be enhanced so that you can read a transcription of a conversation and simultaneously hear a digitized version of the transcription. You'll also be able to search for particular words or phrases and hear how they're pronounced, which will enable the study of lexis, grammar, and intonation. The project is probably the closest we're ever going to get to developing systematically compiled comparable corpora of various varieties of English.

From my own experience I know that when using these large corpora for LSP research one often needs to learn new computer software to quantify and analyze data. The ICE corpus has, for instance, one of the most sophisticated computer applications. The Michigan corpus also has web-based browsing and searching features. What advice can you give about using corpus software?

There are two kinds of software for analyzing corpora. There are programs that people develop, say *Wordsmith* the concordancing program, which you can buy and learn how to use. Alternatively, if you are computationally savvy, you can use a programming language like Perl or Python and then write your own search algorithms or whatever. The advantages of using somebody else's program is that all you have to worry about is learning how to use the program. It's much more difficult to learn a programming language like Perl. But if you know how to do programming, you can customize what you want to do. With *Wordsmith*, you can only conduct analyses that the program is capable of doing. But software for doing linguistic analyses is getting more sophisticated. For instance, the British component of ICE (ICE-GB) comes with a program called ICECUP. ICE-GB is fully tagged and parsed. Using ICECUP, you can search for various grammatical constructions. For the study of gapping I mentioned earlier, we used ICECUP to search for all instances of gapped coordinations in ICE-GB, and we found what we wanted instantly. The disadvantage of using a tool such as ICECUP is that the corpus on which it is used has to be prepared in a specific way. Parsing the corpus was difficult, especially spoken English with all its dysfluencies. This led to many parsing mistakes that had to be corrected by hand.

Susan Hunston and Geoff Thompson describe two possible ways to do research with corpora, corpus-based and corpus-driven procedures. Douglas Biber's group in Northern Arizona also uses what they call statistical and

multidimensional analyses. In your experience as a corpus analyst, what analytical procedure(s) would you recommend to LSP researchers?

I'm in the camp that corpus linguistics is primarily a methodology. In other words, you have your own specific field, whether it's syntax, semantics, psychology or sociology, and then you use corpora as a way to provide data that serves as evidence for whatever analysis you're doing. I am sympathetic to John Sinclair's corpus-driven approach. He claims that looking at a corpus provides you with insights into language that you would just not have found if you hadn't consulted a corpus. I mean, he's done a lot of work in lexicography which obviously has been greatly facilitated by his looking at a corpus. But ultimately I really don't see the data really driving the theory.

With respect to statistics I think any time you deal with numbers you've got to do some kind of statistical analysis, and lack of statistical analysis has been a real weakness of many corpus analyses. Sociolinguistics use a program called VARBRUL which really helps you look at your results through a sophisticated statistical analysis called multiple regression. This test lets you see results that simple frequency counts would miss. Of course I've also seen corpus analyses where people overdue the statistics and the data moves to the background. I guess some happy balance is necessary.

One of the major advantages of corpus methodology is that it easily adapts to many different aspects of enquiry, from discourse features, phraseological patterning, register specificity and rhetorical organization to the analysis of social interactions, to mention just a few. Are corpus-based analyses more suitable for enquiring into all these usage trends or only into some of them?

Obviously all of them benefit. But let me talk about using corpora to study language usage in different contexts, and how this benefits teaching usage to students. If you look at, you know, planned versus unplanned discourse, for instance, you will find different structures used in each. I mean, if you want to teach *who* or *whom* it's actually good to know how people actually use these constructions. Since *whom* is on its last legs, or basically only found in more formal discourse, you may not want not to teach the *who/whom* distinction to people only wanting to speak English conversationally. But in an academic context, you may want to discuss the difference.

There are other contexts too in which corpora can prove helpful. Take conversational analysis. If you look at the traditional literature on conversational analysis, which goes back many years, conversational analysts were actually corpus linguists and they were looking at usage but they never publicly released their data. But with all the corpora of spoken English now available, people can analyze spoken English and not have to create their own corpus. To do critical discourse analysis (CDA), there is a tremendous amount of media data available, particularly on the web. Once you start looking at media data, you get a sense of the role the world media plays in disseminating information, in influencing people's views and opinions. So corpora are permitting analyses that in the past would have been much more difficult to conduct.

Considering the role of corpora such as the *International Corpus of English*, the *Michigan Corpus of Academic Spoken English*, the *TOEFL-2000 Spoken and Written Academic Language Corpus* or the *American National Corpus*, to mention the most relevant ones, can you envisage future directions both for corpus research and corpus-based pedagogical applications?

Well, actually one interesting new area is focused on using the Web as a corpus. Computational linguists are developing programs which can snatch parts of the web to create a corpus. This research is still in its infancy, but if you get the proper retrieval software and you want to do a particular analysis, you just go grab the texts and then you can do whatever you want. You can assemble whatever corpus you want, though the Web is biased towards written language. The closest you can get to spoken language is to download transcriptions of certain kinds of TV shows from, say, CNN. But creating a corpus of spoken language remains a formidable challenge. I don't see advances in working with speech in the foreseeable future. You still have to make recordings and transcribe them. There have been advances in speech recognition, but the software out there works with very restricted spoken texts, primarily those that are primarily monologic. As soon as you get any kind of messy discourse, like spoken dialogue with overlaps, you find that these programs perform very poorly.

When doing corpus research, is it advisable to ground research in a small corpus, in a large one? Or rather, is it better to use comparable corpora (i.e. parallel corpora for translation analyses, American vs. British English corpora, spoken vs. written corpora, general vs. specialised corpora, academic vs. professional corpora, etc.)?

It all depends on what type of research question you want to answer. The nice thing about a corpus like the Brown corpus is that it has been around for over 40 years now and it's become a kind of benchmark. The more people who analyze it, the more we know about it and you can compare results. So I think that it is extremely useful to have these commonly available corpora. On the other hand, if you have a very specific research question you want to look at and no corpus can answer that question, you obviously have a problem and you have to create your own corpus, which can be quite a bit of work. The size of corpus you need depends very much on the frequency of the linguistic structures you want to investigate. Gapped coordinations, for instance, occur very rarely. To conduct a study of them, you'll need a fairly large corpus. But I did a study of apposition in the early 1990s and found that in press reportage, for instance, they occurred very frequently. So I needed only short excerpts to find enough examples.

Let me move now towards more learning-oriented questions. From my own experience as LSP teacher, corpus data have become extremely helpful as pedagogical models as well as a source for deriving more realistic pedagogical materials. I also see corpus data as a very reliable source for making informed decisions about what our teaching priorities should be with regards to students' specific needs. What other functions does corpus linguistics serve in the area of second language acquisition?

I teach a lot of students who are going to be teachers of second languages, primarily English but other languages too like Spanish or Portuguese, and I find that a lot of them don't know much about language. I use corpora in my classes to raise my students' consciousness about language. I know that data-driven learning is typically thought of as a methodology based on using corpora to teach learners of English and other languages. But I use it to help my students learn more about the structure and use of the language. So many of them, especially the Americans, have been educated in a very prescriptive tradition. Having them look at real data really helps them understand English grammar. Of course, when I suggest they do the same with their students, many complain that their students are just not at a high enough level of proficiency to read let alone analyze a corpus of data. So I've been experimenting with online graded readers as a way of supplying suitable texts for beginning learners of English.

As I commented in the introduction of this interview there is a growing tendency among LSP teachers in constructing learner corpora (for instance, the works of Sylvianne Granger). How can we create or use our own corpora in order to do research in second language acquisition and, in our case, in LSP learning?

One technique I've used is to archive all the written assignments that my students send me. I mean, everything that's written these days is done on a word processor, so why not take advantage of this? After a while you end up with a fairly sizeable corpus and then you can have your students look at their own language, which they may be a little reluctant to do it first to do but I find it quite instructive to have students use, say, a concordancing program to look at a text that they or their classmates have created.

As a leading authority in the use of the World Wide Web as a linguistic corpus, I am sure you can provide us with some suggestions or practical tips to use the WWW in the teaching/learning of Languages for Specific Purposes.

Well, let me give you two examples. There are a lot of texts on the web and what I do in my own classroom is if I want to teach some element of grammar I download some texts, I strip all the html markup out of them, and throw them into a concordancing program. I bring a computer to class and then we look at things. I like to do this with current events, for instance. I did it last year with one of the Bush-Kerry debates. We explored whether George Bush was as redundant as he's claimed to be. In fact, we found that Kerry had a greater number of repeated expressions in his responses than Bush. This is one use of the Web. Another thing I have my students do is just go to a search engine and you know pick something you want to look at, say, modal verbs, and just put *may* or *might* in Google. You get three billion hits, but you can have students focus on the first ten hits and start determining the meaning of the modals.

Is it reliable to do this?

It can be overwhelming, but from my experience, students are interested in the web. It has become such an integral part of their lives that I find they are interested in how language is used on it. There are also programs designed specifically to analyze the web. For instance, Webcorp is an online concordancing program. It can be used to search for words and phrases and organizes hits into concordance lines. The Linguistic search engine allows you to search three million words of data from the Web that's been parsed. Once you get used to its interface, you can start analyzing parsed Internet data, not simply strings of words or phrases. So I think there are really interesting things happening with the Web. The obvious limitation is that the Web lacks spoken data other than transcripts of talk shows or press conferences.

When you are using the web, do you modify or adapt smaller texts for example, to make them less difficult for students?

I don't do that myself but you certainly could do that. As I said earlier, I've been collecting graded online readers for analysis by students at earlier stages of learning English.

I would like to finish this interview by requesting your advice on bibliographical references. What books can you recommend to corpus analysts? Your latest publications are an excellent source of information. What other books can be of interest for undergraduate, post-graduate students, young researchers working on their PhDs and, why not, all other

LSP researchers involved in corpus analyses?

I wouldn't be modest if I recommend a book of mine, *English Corpus Linguistics: An Introduction*. It's a good start. Tony McEnery and Andrew Wilson have their own introduction, *Corpus Linguistics*, which is good. Graeme Kennedy's book, *An Introduction to Corpus Linguistics*, provides a nice overview of the field. If you like the corpus-driven approach, Elena Tognini Bonelli's *Corpus Linguistics at Work* is a nice book. There's also Douglas Biber et al.'s *Corpus Linguistics*, which provides a good introduction to multidimensional analysis. These are all good introductions to corpus linguistics.

[This interview was held at the Department of English Studies,
University of Zaragoza (Spain) on the 3rd June 2005.]