CORPORA FOR ENGLISH LANGUAGE TEACHING AND LEARNING

Carmen Santamaría García
Escuela Universitaria de Magisterio de Guadalajara
Universidad de Alcalá

Resumen

Los corpora informatizados de lengua inglesa abren un amplio campo de posibilidades tanto a profesores como a estudiantes de dicha lengua. Nos ofrecen un banco de datos de lengua real con un enorme potencial para la enseñanza. Después de una breve introducción sobre corpora y los programas informáticos que nos permiten su estudio, presentaré algunas de sus aplicaciones a la enseñanza y aprendizaje de la lengua inglesa.

Abstract

Computer corpora of English Language open up a wide range of possibilities for teachers and learners of this language. These corpora contain a powerful databank of real language awaiting to be discovered and applied to ELT.
After a brief introduction on corpora and the computer programmes that may be used to exploit them, I will present some of their applications to English language teaching and learning.

1 ENGLISH LANGUAGE CORPORA

In this paper I am referring to *computer* corpora, i.e. collections of naturally-produced spoken or written texts assembled to contain a representative sample of a language (or a particular type of texts or variety of language) and stored in machine-

readable form. Their production and size have been increasing over the last two decades. The first computer corpora consisted of a million words but the latest projects are certainly more ambitious. The *British National Corpus*, for instance, contains a 100-million-word sample of modern English and the current central *Bank of English* of the COBUILD project is 211 million words of text and «has no final extent because, like the language itself, it keeps on developing» (Sinclair 1991:25).

Such projects are very expensive and are usually collaborative ventures between leading publishers and important institutions. The members of the *British National Corpus* consortium are: Oxford University Press, Longman, Chambers-Harrap, Lancaster University's Unit for Computer research in the English language, Oxford University Computer Services, and the British Library (Rundell 1995:2). The *Cobuild* project is a collaborative venture between Harper Colins and the University of Birmingham.

For a survey of the English Machine-readable Corpora compiled up to 1991 you can read Taylor, Leech and Fligelstone (1991).

2 CONCORDANCING TOOLS

Concordancers are computer programmes that allow the researcher to search corpora for words and word patterns. There are several programmes available on the market. To quote but a few, you can use Oxford Concordance Program, KAYE, Longman Miniconcordancer, Microconcord, Wordcruncher, Free Text Browser or TACT. All of them offer the same basic utilities, for instance:

a) Instant *Word lists* of all the word types in the corpus sorted by their beginnings or endings and listing their frequency.

b) *Searching commands* that search for all the occurrences of any word.

Wildcard operators are usually provided to select words matching the criteria specified by the user. For instance, an asterisk means *any character*, and thus the expression **ed* will search for all the forms ending in *-ed*. It is also possible to search for co-occurring words either occurring in immediate adjacency or nearby.

c) *Several display formats* of search words or sequences of characters. The most common display format is KWIC (acronym for *key word in context*) usually called a *concordance*. This display presents all the occurrences of the search word highlighted in the middle of the screen and surrounded by their context. The amount of context displayed (from one word up to several lines) can be specified by the user. It is also possible to switch to a whole text context for closer study of the search word. A *collocation* display shows the words that co-occur with the key word most frequently.

d) *Saving command.* It enables the user to save concordances for future use and retrieve them with a word processor.
e) *Print out command.* It allows the user to obtain concordance sheets. It is extremely useful when it is not possible that students use a concordancer directly on a computer: their teacher can provide them with computer printouts.

3 APPLICATIONS OF CORPORA CONCORDANCING IN ENGLISH LANGUAGE TEACHING AND LEARNING

Corpora store real language and constitute an important source of actual examples for the compilation of grammars, dictionaries and coursebooks. It is nowadays assumed that learners must study real instances of language. Sinclair is a keen supporter of this idea: (Sinclair et al. 1990:vii) «I am convinced that it is essential for a learner of English to learn from actual examples, examples that can be trusted because they have been used in real communication». And he adds: «There is no justification for inventing examples» (*op.* cit.:xi).

These grammars, dictionaries and coursebooks are necessary learning tools. However, corpora may also be used directly in ELT. Their use presents several advantages over the above mentioned traditional sources of actual examples and above the use of introspection, -another source used to supply examples in the classroom-. I will attempt now to summarise these advantages:

- *Grammars, dictionaries* and *coursebooks* offer a limited selection of examples due to an obvious limitation of space. Corpora, on the other hand, contain a great number of instances of actual language in use that may be used: a) to supply additional information when required, and b) to be incorporated in the design of exercises.

- Reference materials show their authors' reflections on language. But we may want to explore language ourselves. In this case, the evidence in corpora may help our purpose. Our findings may even challenge the presentation of a particular point of grammar in grammars and coursebooks, especially in those which are not based on real language evidence. (See section 3.2.1). With access to corpora, students and teachers have the possibility to interpret raw data and offer a different presentation.

-*Introspection* is another source of examples used by teachers, especially to answer questions of the type «what is the difference between x and y?» or «Is x correct or incorrect?» However it is not advisable to `make guesses' based on intuitive data. Conversation and Discourse Analysis have proved that intuitions

about language may not be accurate nor sufficient although intuition is always prerequisite to analyse language. Corpora contain evidence of language in use and are much more reliable than even native speakers' intuitions. Their data may help teachers answer students' questions. Why making guessess when a reliable source containing authentic information can be accessed?

However, intuitions about language continue to be essential. As Knowles (1990:45) says: «If intuitions are insufficient without a corpus, a corpus is also insufficient without intuitions. A corpus provides masses of data, and intuition is needed to analyse it».

From this brief presentation of some of the advantages of corpora as sources of authentic language, we can derive some of their applications to language teaching and learning.

Corpora may be used both in and outside the classroom:
a) for research by teacher or learner
b) to devise activities that complement the information on reference grammars, dictionaries and teaching materials either
- focusing on a particular item, or
- helping answer students' questions

These applications may serve the study of grammar, vocabulary and discourse from both inductive and deductive approaches to language learning. Murison-Bowie (1993:39-44) elaborates on the usefulness of these two approaches. Inductive reasoning (also called *bottom-up*) consists in giving students examples taken from a corpus and encouraging them to discover regularities and rules: «(...) one takes the evidence as a starting point and by a series of inductive steps tries to discover the patterns in the language and the rules which govern those patterns» (Murison-Bowie, 1993:40).

This approach is very related to *discovery techniques* (Harmer 1983) and favour discovery learning, i.e. «they present language in a way that enables learners to discover new knowledge for themselves, rather than being spoon-fed» (Tribble and Jones 1994:35).

Deductive reasoning (or *top-down*), on the other hand, «starts at the top -with a generalization, a theory, a hypothesis about something, and then looks at the data» (Murison-Bowie, 1993:42).

I will try now to illustrate the above mentioned applications of corpora.

3.1. Research by teacher or learner

With the help of corpora and concordancing programmes both teachers and learners may take the role of language researchers. Johns (1988:14), elaborating on Seliger (1983), suggests that language learning and linguistic research are close parallels:

Like a researcher, the learner has to form preliminary hypotheses on the basis of intuition and scanty evidence: those hypotheses then have to be tested and rejected or refined against further evidence, and finally integrated within an overall model.

The data obtained through corpora concordancing serve language researchers as evidence to prove their hypotheses, to discover new patterns in language and as a stimulus to challenge established language descriptions.In a similar fashion, teachers and students can use concordance printouts to test hypotheses formulated in the classroom, to look for language patterns or to challenge language descriptions in their coursebooks.

This sharing of researcher role by teacher and students may derive in a change of attitudes. Students may realise that teachers are not `databanks' or `founts of all knowledge' but collaborators in a never-ending learning process. Teachers and learners are faced with the same task: the exploration of language through data evidence. The teacher can help learners to make discoveries about language but s/he is also learning. This fact increases the element of risk in the teaching situation, i.e. up to a certain extent, it is not predictable what students are going to discover in the data and their discoveries may challenge teacher's knowledge. But we should not be afraid of risk because, according to Johns (1988:11) « (...) the effectiveness of the teacher is potentially greatest when he or she is most at risk.» And the least stimulating learning situation is that of minimum risk:

An extreme example of minimum risk is the scenario in which the teacher ploughs through a textbook reading out the explanations and checking students' answers to exercises against the teacher's key. (...)

Taking some risk then, may be very possitive for the learning situation. Corpora concordancing may thus have an important impact on language learning if students become aware that both teachers and learners should speculate about language and test their knowledge against evidence.

3.2. Complement to reference materials and coursebooks.

Apart from traditional reference materials as grammars, dictionaries and coursebooks where a selection of examples is provided, teachers and learners can access to real examples of language by means of concordancers. It is easy, quick and motivating. We will see examples of the application of corpora both to focus on a particular item of language and to answer students' questions. We will consider their usefulness in the study of grammar, vocabulary and discourse combining inductive and deductive methods.

3.2.1. Focusing on a particular item

A particular grammatical item may be presented through a concordance. As an example in the study of grammar, students can learn the different uses of reflexive pronouns from close observation of a concordance printout. Following an inductive approach, they can attempt to assign a different category for each different use. By searching for **self/*selves* in the London Lund Corpus (LLC henceforth) you can get some hundreds of sentences like these:

We must ask *ourselves* a personal question.
We don't want to compromise *ourselves.*
I think I'm quite good at abstratcing *myself.*

The categories assigned by students can be discussed with partners and finally, with the teacher. This activity gives learners responsibility for their learning. Knowles (1990:45) points out the benefits of learning from concordances: «Instead of assimilating information and theories, learners can actually test theories and find things out for themselves».

Coursebooks, on the general, devote small sections to present grammar. After completion of a unit, students may feel they still need more examples to check whether they understand the item in question. At this stage, concordance printouts may be a helpful source of examples. It may also be the case that the presentation of an item may be challenged by the data in a corpus. To give an example, one of my students of first year wanted to know if the use of *any* as a general determiner used to talk about some non-specific person or thing was a frequent use. Up to now they had only been taught the use of *any* as general determiner used in negatives and questions and therefore they thought these two would be the most frequent uses.

To find out, we prepared a search for *any* in a corpus of 31.500 words and we found 39 occurrences of *any.* Students had to assign a number from 1 to 7 to each different use of *any* according to the following categories:

1. *Any* as general determiner used in questions asking whether something exists or not.
2. *Any* as general determiner in negative statements to say that something does not exist.
3. *Any* as general determiner to talk about someone or something when you do not want to mention a specific person or thing.
4. *Any* as pronoun.
5. *Any* as submodifier in comparison.
6. *Any* as a quantifier.
7. *Any* as general determiner referring to a quantity of something which may or may not exist.

In our data, 24 instances of *any* belonged to category 3. There were only 3 instances of *any* used in questions (category 1) and 6 instances of *any* used in negative statements (category 6). From this evidence it seems that textbooks should devote more attention to the use of *any* in category 3 than they do.

The teaching of vocabulary may also benefit from corpora concordancing. A concordance presents the same word in different contexts simultaneously. This fact facilitates that learners deduce the meaning of the key word when it is a new word for them. I gave my students these sentences obtained from a concordance of *taste* from which they successfully deduced its meaning:

1. Guinness has a slightly sharp *taste*.
2. I was going to make a very stupid remark. I was going to say that nothing that *tastes* nice is poisonous but of course the things that are poisonous we don't eat, so we don't know if they taste nice or not, do we?
3. A kingfisher is probably not going to *taste* very good.
4. Slugs don't *taste* much better.

The context in concordances also facilitates activities on polysemy and confusable words. Learners can be asked to classify the different meanings of a polysemic entry like *bank* or to distinguish the meaning of confusable words like *able* and *capable*, *above* and *over*, *actual* and *real*, *actually* and *really*, etc. It will be easy to prepare a gapfill exercise by merging concordances of confusable words and then blanking out search words.

The collocation display is a useful facility to study collocations. A concordance of **self/*selves*, for instance, reveals grammatical words as *to, of* and lexical words as *find, ask* as frequent collocates. A KWIC display highlights language patterns. This feature may be exploited to

devise activities on syntactic patterns of keywords. The prepositions that may follow a particular verb or adjective are apparent from a right-sorted KWIC display. For instance a search for *consist/consists/consisted/consisting* gives plenty of examples followed by *of* and *in*.

Wildcard operators are extremely useful for learning derivations. We can prepare exercises on affixes by retrieving all the examples we need. A search for *anti*/ counter*/ de*/ dis*/ ex*/ in*/ non** will retrieve words containing negative prefixes. It will also throw up plenty of words where these beginnings are not prefixes (like *distant*). But you can easily eliminate the unwanted words.

The study of discourse depends on corpora. Its object of study are authentic texts either in spoken or written form. Therefore, the retrieval of whole texts and conversations can provide useful material for the study of, for instance, cohesive devices or conversation mechanisms. The relationship between form and function is one of the main concerns of discourse analysis. Corpora enable teachers and learners to study language forms (grammatical and lexical ones) in context, therefore enabling them to study function. Retrieval of whole conversations may be specially useful to expose learners to authentic conversations because textbooks, on the general, only include them for extensive listening.

3.2.2. Helping learners answer their questions

Let us see now, another way to complement reference materials, i.e. to help learners answer their questions. Johns (1988:11) explains how a situation of high risk for the teacher may arise in the classroom «if a student has the temerity to ask a question: `Please, what is the difference between *therefore* and *hence*?´» He continues to reveal the potential danger of evading questions:

What often happens in practice, of course, is that teachers develop strategies for avoiding risk by evading questions, and in so doing suppress the innate curiosity that is a precondition -perhaps the precondition- for successful language learning.

> The teacher, either native or non-native speaker of English, may have problems to illustrate an answer but s/he should not evade questions nor use introspection to improvise an answer. Thus, the teacher has at least two safer alternatives: a) direct students to read the information in their reference materials, b) prepare a corpus search for the words or word forms in question.

Presumably, the effort that learners have to make to find out an answer for themselves in a concordance output will result in better retention of the information. The novelty will be a stimulus too. As an example, one of my students was puzzled when she read in Collins COBUILD English Grammar (Sinclair, et al. eds. 1990:57) that the determiner *either* usually indicates that only one of two is involved but it can also mean both of two things: «especially when it is used with `end' and `side'». She was afraid that she would not be able to recognise which of the two senses of *either* was operating in other examples. I prepared an exercise eliciting the meaning of *either* in ten sentences retrieved from LLC. Some of them were:

1. (...) being carried down the steps the heralds flanking on **either** side (...)
2. There is this beautiful Cross of Westminster with two taperers standing on **either** side and the mourners follow after.
3. I read on the back page of **either** The Sunday Times or The Observer (...)
4. I believe this is approached by two carved and gilded doors on **either** side o the altar.

From a sample of thirty one instances of *either* as a determiner, *either* meaning *both of two things* was only found seven times and in all the seven cases *either* was followed by *side*. As a conclusion students realised that it was not such a problem to distinguish the two meanings of *either* in authentic texts.

* * *

REFERENCES

BONGAERTS, T., P. de HAAN et al. (eds.). (1988). *Computer Applications in Language Learning.* Dordrecht: Foris.

HARMER, J. (1983) *The Pratice of English Language Teaching.* London: Longman.

HOFLAND, K. (1991). «Concordance Programs for Personal Computers» in Johansson and Stenström (eds.), 283-306.

JOHNS, T. (1988). «Whence and Whither Classroom Concordancing?» in Bongaerts, de Haan et al. (eds.) 9-35.

JOHANSSON, S. and A. STENSTRÖM. (1991). *English Computer Corpora. Selected Papers and Research Guide.* Berlin, New York: Mouton de Gruyter.

TRIBBLE, C. and G. Jones (1994) *Concordances in the Classroom.* London: Longman.

KNOWLES, G. (1990). «The Use of Spoken and Written Corpora in the Teaching of Language and Linguistics» in *Literary and Linguistic Computing,* 5:1, 45-48.

MURISON-BOWIE, S. (1993). *MicroConcord Manual.* Oxford: Oxford University Press.

RUNDELL (1995) «The British National Corpus» in *Longman Language Review,* London: Longman, 2-5.

SELIGER (1983). «The Language Learner as Linguist : of Metaphors and Realities» in *Applied Linguistics* 4/3.

SINCLAIR, J., G. FOX, et al. (eds.). (1990). *Collins COBUILD English Grammar.* London: Collins.

SINCLAIR, J. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

SVARTVIK, J. and R. QUIRK (1980). *A Corpus of English Conversation.* Lund Studies in English 56. Lund CWK Gleerup.

TAYLOR, L., G. LEECH and S. FLIGELSTONE. 1991. "A Survey of English Machine-readable Corpora" in Johansson and Stenström (eds.) 319-354.

WILLIS, D. (1990) *The Lexical Syllabus.* London: Harper Collins.