Revista Alicantina de Estudios Ingleses 23 (2010): 51-67

Corpus Linguistics and its Applications in Higher Education

Miguel Fuster and Begoña Clavel
University of Valencia
miguel.fuster@uv.es / begona.clavel@uv.es

ABSTRACT

The aim of this paper is to review and analyse relevant factors related to the implementation of corpus linguistics (CL) in higher education. First we set out to describe underlying principles of CL and its developments in relation to theoretical linguistics and its applications in modern teaching practices. Then we attempt to establish how different types of corpora have contributed to the development of direct and indirect approaches in language teaching. We single out *Data Driven Learning* (DDL) due to its relevance in applied linguistics literature, and examine in detail advantages and drawbacks. Finally, we outline problems concerning the implementation of CL in the classroom since awareness of the limitations of CL is vital for its future success.

1. ICTs and Corpus Linguistics

It is commonplace that rapid progress in information and communication technologies (ICTs) has modified enormously the world we live in, relationships among individuals and also interaction between different speech communities. This, together with greater access by ever larger segments of the population to new technologies has in many ways gradually but inexorably affected the way we think about ourselves and others. It is safe to say that at the opening of the twenty-first century the more educated sectors of every nation feel they are part of a global community more so than a few decades back. In this new context, languages are contradictorily perceived either as defying insurmountable obstacles or as offering exciting challenges. In this same sense, the English language has reaffirmed its leadership, and has become the indisputable world language (Graddol, 1997: 4). It is somewhat early to adequately measure the full consequences of all this technological progress in language development, but a diachronician like Fennell views this revolution, which started during the 1980s and 1990s in the US, as significant an event as the Industrial Revolution of the nineteenth century, making English more global than ever before (2006: 256-8). Lavid has put it quite graphically as "a move from muscle to intelligence (2005: 31). Set against this sweeping backdrop of technological change, it would have been very surprising if pedagogy had remained excluded or unaffected. Indeed, novel methods for the teaching and learning of languages have developed and been rapidly incorporated from the onset of this technological upheaval (Braun, 2005: 51). As a corollary to these developments the role of teachers as central suppliers of information has also been disputed. Interactive multimodal packages of all sorts and formats seem to abound in the area of EFL where the physical presence of the teacher appears to be redundant. However, what these glamorous courses have to offer is, more often than not, uncritically accepted by consumers, while experts hardly ever take the opportunity of discussing the results.

Corpus development and corpus linguistics (CL) are clear outcomes of these technological advances. To start with, a corpus has been defined as 'a collection of naturally occurring language texts, chosen to characterize a state or variety of a language.' (Sinclair, 1991: 171). Large and not so large computerised collections of texts and samples of texts put together according to systematic principles have found a permanent place in the field of linguistics for a variety of purposes. Corpus studies are successfully integrated into language research today in practically every discipline (see for instance Hornero et al., 2006) and the growth of corpus-based research and analysis in practically every discipline in linguistics is immediately observable. Also, entire journals and conferences are devoted to corpus studies of various kinds, whereas traditional conferences and journals host sections to CL, if only to note developments in this area (Braun, 2005: 47; Lavid, 2005: 348-9). As Römer and Wulff suggest, no matter their persuasions, those who adhere to corpus practice:

> (...) share the common assumptions that linguistic theorizing should be driven first and foremost by (representative samples of) authentic language data, and that a solid linguistic hypothesis and theoretical claims should be based on a thorough description of these data with regard to the phenomenon under investigation (2010: 100).

2. The development of corpus-based dictionaries and grammars

Over the last decades, many reference works have been shaped and informed by corpus research. Corpus research through computer use lies at the heart of modern lexicography pioneered at the University of Birmingham by Sinclair's *Collins Dictionary of the English* (1987). This new attitude should be seen as part of a drive towards offering more realistic language descriptions in contemporary lexicography. Many facets of contemporary English dictionaries we all have, particularly those which acknowledge their debt to authentic speech, show signs of the influence exerted by corpus research. The urge to provide compelling evidence of how the native language is currently used, along contextual information, has made the role of corpora an essential part of the development of EFL dictionaries, it should be said, to the point that it would be hard to find exceptions. For instance, the pattern set by the

Longman Dictionary of Contemporary English since 1988 with the adoption of its 'defining vocabulary', that is the exclusive use of the 2000 most frequent English words has also been immediately followed by other common learners dictionaries and, to a much lesser extent, by bilingual dictionaries. Nevertheless, against this background, Teubert finds that "[i]n many ways, the *Cobuild* dictionary is still unique. While it encouraged other dictionary makers to include more corpus evidence, there is still no other dictionary exclusively based on corpus." (2004: 112). Beyond traditional lexicographical practice which offers the classic alphabetic arrangement of word entries, other products such as thesauri conspicuously show their adherence to corpus while dictionaries of collocations, pioneered by the Benson, Benson and Ilson (1986) *The BBI combinatory dictionary of English* are significant corpus-informed departures from earlier lexicography.

Descriptive English grammars have also gone through a similar process in these last decades. For instance, research initiated by Quirk, Greenbaum, Leech and Svartvik grounded in empirical work culminates in the publication of *A Comprehensive Grammar of the English Language* (1985). This grammar made extensive use of the Survey of English Usage, and although it was data-driven, the data had not been computerised (Teubert, 2004: 107). Moreover, this same empirical trend has been followed by other corpus-based descriptive and pedagogic grammars. *Collins Cobuild English Grammar* (1990) in Sinclair's own introductory words "attempts to make accurate statements about English, as seen in the huge Birmingham Collection of English Texts" (1990: v). More recent developments have focused on the linguistic description of speech, a much neglected area in earlier descriptions of the English language. New grammars, like Biber et al's *Longman Grammar of Spoken and Written English* (1999), or more recently Carter & McCarthy's *Cambridge Grammar of English* (2006) underscore that they represent authentic spoken and written English, having large corpora as their sources. So far we have outlined developments concerning printed reference works. However, the overall picture can only be obtained when newer multimedia and online reference materials are taken into account, but this is beyond the scope of the present contribution.

3. Corpus linguistics and theoretical linguistics: corpus-driven and corpus-based approaches

In sum, the earlier reliance on native speakers' intuition or native researcher's introspection in discussing language matters has given way to corpus-based language descriptions. Notions such as naturalness, authenticity, or the need for quantitative and qualitative analyses have become essential in order to decide on crucial aspects like acceptability and appropriateness. However, Partington -quoting Leech-, argues that there is a spurious theoretical controversy around the dichotomy *intuition* vs. *corpus*. In his view, both intuition and corpus linguistic evidence are essential in language research (Partington 1998: 2):

> Recent corpus users have accepted that corpora, in supplying first-hand textual data, cannot be meaningfully analysed without the intuition and interpretative skills of the analyst, using knowledge of the language (*qua* native speaker or proficient non-native speaker) and knowledge

> about the language (*qua* linguist). In other words, corpus use is seen as a question of corpus plus intuition, rather than of corpus or intuition (Leech 1991: 74).

These issues are far from being resolved even today. Flowerdew (2004: 113) detects that criticism of CL stems basically from generative quarters. For Chomsky and his followers corpora offer performance data, and represent *externalised language* (E-language). Chomsky's approach to language is rationalistic, and is based on *competence* or *internalised language* (I-language): what matters is what can be said or written, not what is actually said. Leech tries to bridge this gap (2007: 135) by suggesting that there is a connection between I-language and E-language if we assume that performance data is crucial to understand competence even though performance data in CL is made up of samples of language that might not represent the language as a whole.

In spite of these criticisms, CL is a burgeoning field and linguists of all persuasions have been attracted to it. A consequence has been the restatement of the older theoretical controversy. So there are two basic tendencies: (1) corpus-based practitioners and (2) corpus-driven linguists (Tognini-Bonelli, 2001:1). For Teubert (2004: 112), in essence:

> Linguistic findings (including the contents of dictionaries) are corpus based if everything that is being said is validated by corpus evidence. Findings are corpus driven if they are extracted from corpora, using the method of corpus linguistics, then intellectually processed and turned into results.

Divergent conceptions of the role of CL provide the blueprint for research. Whereas cognitive linguists view corpus exploitation as a *methodology*, discourse analysts like Teubert himself, view corpus as a distinctive *approach* to language. A continuum between both tendencies can also be detected among contemporary corpus linguists. The basic differences between them can be summarised as follows (see Gries, 2010; Teubert 2010; Teubert 2004):

For corpus-driven linguists:

- it is an approach;
- meaning has to be negotiated, and is found in discourse (detectable in *corpora*);
- theory is based exclusively on corpus data (it is 'bottom-up');
- corpus annotation is rejected: discourse and not an external taxonomy should supply the categories and classifications.

For corpus-based linguists:

- it is a methodology;
- meaning is first found in the minds of speakers (*corpora* are conceived as 'toolboxes');
- corpus data is used to test and/or improve prior theories (it is 'top-down');
- an external taxonomy of corpus annotation provides the categories and classifications.

This opposition is well illustrated by Teubert's position (2010). This author rejects annotations since these categories are not validated by corpus evidence. In his view, the results are preconceived by the annotator. He illustrates his rejection to the mentalist and cognitive linguistics view of corpus because they prioritise thought over language with two examples:

> A word tagged as a 'noun' will be counted as such. But it will shed no light on the "true nature" of 'noun'. This does not mean that tagging cannot be useful and simplify our work. If I am interested in synonyms and antonyms a good way to find them is to study binomials. I might look up the phrase *friends and*, and I may not know beforehand which nouns will come up as the next word on the right, words such as *allies* or *enemies*. What is important, though, is that it is discourse, and not a mental mechanism, that tells me what counts as synonyms or antonyms of *friends*. Synonymy and antonymy are discourse constructs, not emanations of a hypothetical language system (Teubert, 2010: 355-6).

These two different views on the role of corpora, with a continuum between them, will ultimately affect the field of applied linguistics, and or how the language or corpora should be introduced in teaching and learning as discussed below.

4. A corpus typology

Technical improvements of various kinds and the enthusiasm of CL have contributed to offer a very rich picture as far as corpora design is concerned. To start with, we may highlight the significant advance in terms of corpus size. The first computerised readable corpus, the *Brown Corpus*, published in the 1960s, compiled by Kucera and Nelson Francis, contained *only* one million words. Today, the *Cambridge International Corpus* contains over one billion words. Whereas the term *corpus* may refer to whole collections of texts attributable to a single writer, say Shakespeare's plays, corpus linguists have focused their attention mostly on compiling language samples. Samples are selected by corpus linguists with the view of offering small scale replicas of the whole language/discourse they aim to represent. Consequently, the results obtained from the consultation of a representative corpus can be extrapolated to represent "the whole universe of language use of which the corpus is a representative sample." (Leech, 2007: 135). Of course, if one wishes to build a cannibalistic corpus, representativeness or balance are not at stake, as long as the corpus is sufficiently large (Teubert and Èermakova, 2004). According to Leech (2007) the issue of representativeness, balance, or strict comparability of different corpora have neither been properly dealt with nor solved satisfactorily to date. Corpus design should be closely linked to its exploitation. A review of the state-of-the-art offers a wide range of corpora as shown below. Note that this classification does not necessarily represent cut-and-dried corpus types, and there may be considerable overlapping (see Hoffmann et al., 2008; Leech, 2007; Teubert, 2004; Teubert and Èermakova, 2004):

1. Historical/diachronic corpora vs. synchronic
2. Spoken vs. written
3. Specialized vs. general
4. Static vs. Dynamic

5. Plain vs. Marked-up
6. Native vs. Learner
7. Parallel
8. Comparable
9. Pedagogic corpora
10. The web as a corpus

Let us examine briefly some of these typologies in terms of relevance to corpus use in applied linguistics. For example, the contrast between *plain* (only the actual words) or *marked-up* (annotation) allows different types of exploration. Corpus mark-up refers to various kinds of taxonomies and classification (see Hoffmann et al., 2008: 24-6; Lavid, 2005: 310 & ff):

a. *Metatextual mark-up* yields details about the text production and the speaker - gender, age, the time when texts are produced, etc.
b. *Structural mark-up* and *typographical mark-up*: as for instance, paragraph boundaries, in the spoken language, speaker changes, overlaps in spoken language, interruptions, etc.
c. *Annotation* refers to linguistic features: for instance, *tagging* is used to indicate part of speech, or sentence function of words, etc.

Note that while *mark-up* enlarges the options of exploration at our disposal, it also implies manipulation. From a purely pedagogic viewpoint, it can be stated that no specific corpus type can be said to be superior to the rest. Teubert (2004) remarks, for instance, that parallel corpora, i.e. corpora containing original texts and their translations, are far superior to bilingual dictionaries:

> Even the largest bilingual dictionary will present only a tiny segment of the translation equivalents we find in a not too small parallel corpus. (...) bilingual dictionaries do not help to translate into a language we are not very familiar with. The user is left with many options and hardly any instructions for selecting the proper equivalent. From parallel corpora we can extract a larger variety of translation equivalents embedded in their contexts, which make them unambiguous (...) (Teubert, 2004: 123).

In one way or another, for those whose main concern is language teaching all corpora can offer interesting insights and can be thought of as a valuable aid (Hoffmann *et al.*, 2008: 14-5; Römer, 2008: 118; Flowerdew, 2009: 405), of course, as long as their specific compilation criteria are properly understood. General corpora, which tend to be quite large, are basically designed to represent the language of a whole speech community (Flowerdew, 2004: 12). So, for instance the *British National Corpus* (BNC), a static corpus of 100 million words, aims to represent a national variety through samples of contemporary spoken and written British standard of English. These large general corpora are often subdivided into smaller subcorpora to allow more nuanced exploitation and analysis. A similar corpus, the *American National Corpus* (ANC) is an ongoing corpus project parallels to the BNC, which aims to offer a picture of Contemporary American English. Admittedly, these two national varieties of the inner circle (Kachru, 1982, 1988) are very well represented in various other corpora. However, if we turn our attention to the issue of representativeness and balance we need to be cautious. To start with, the spoken samples in the BNC are just 10% of the total. If wished to carry out research within the specific field of, say, academic English by means of the BNC, we will observe that there is a significant imbalance. Almost 50% of all academic journal articles are medical texts (Hoffmann et al., 2008: 213). Therefore, our results about the academic language in British academic journals would be flawed. Likewise, it may be argued that small specialized corpus, which are considered as extremely easy to compile and useful in English for Academic and Specific Purposes (see Flowerdew, 2009: 397) very often cannot be taken to represent national varieties, but a mixture of contributions pertaining to the inner, outer or even the expanding circle (*ibid* Kachru). This might cast doubt on the use of L1 or native labels to refer to contributions whose authors definitely do not have English as their mother tongue, and certainly do not use it as their own national variety. Indeed, most scientific work published today is in English though much of it does not qualify as representing the inner circle. This earlier argument is unrelated to language proficiency. If we still wish to make use of these samples, perhaps it would be more appropriate to refer to them as representing collectively a community of practice of 'Successful Users of English' (Prodromou, 2003), where English acts as the lingua franca.

A final cautionary point concerns, for instance, the use of the Internet as a corpus. Here an interface like *WebCorp* has many appealing features for those who wish to explore the web. The web can be seen as a dynamic corpus whose main asset is that it may provide us with discourse samples which cannot be retrieved through conventional corpora. As a case in point, Leech mentions the search for the new coinage *deferred success*, which comes out in newspapers in 2005. This item of PC language recently coined in the UK was a conscious choice in the teachers association in order "to replace the word *fail* as a verdict on children's school work." (Renouf, Kehoe and Banerjee, 2007: 51). It is obviously impossible to find recent lexical items of this kind within a closed corpus like the BNC, whereas other dynamic corpora are not publicly available. However, Leech warns that search engines like *Google* do not offer representative segments of language: "The consensus seems to be that frequency information obtained from Google is at present seriously misleading." (Leech, 2007: 144).

5. Corpus research and EFL: indirect and direct approaches.

According to Barlow (1996) when we refer to corpus use for teaching purposes teachers have in fact, different methods at their disposal:

1. Teachers can either analyse the corpora themselves for material design or they can decide to introduce them in the classroom in order to:

i. determine frequency patterns in specific domains;
ii. enrich language knowledge;
iii. produce 'authentic data;
iv. generate teaching materials.

2. Teachers may train students in the use of corpora:
i. through research questions decided on by the teacher;
ii. by exploring an issue in a more open-ended way. (see Partington, 1998: 5-6)

Broadly speaking, two applications that we may call *direct* and *indirect* (see Römer, 2008: 113; also Braun, 2005: 51) can be identified here, that is, where either teachers themselves use corpora to teach or students are given unmediated access to the corpora. On the one hand, indirect applications can have an impact on syllabus design or teaching materials. On the other, a direct approach would offer a double option: (1) teacher-corpus interaction or (2) learner-corpus interaction. *Data Driven Learning* (*DDL*), which we will examine in some detail below, is perhaps the most widely known proposal of this direct approach.

As far as the indirect approach is concerned, beyond the publication of corpus-based dictionaries and grammars, language materials writers have also greatly benefited teachers who rely on textbooks, particularly those whose contents do not match with observations of the target language, run the risk of offering learners a distorted picture of the language. They may be offering, one might say, some kind of 'toeflese' which does not necessarily correspond to the desirable target along various parameters, such as frequency or typicality. For instance, the scrupulous analysis of the presentation of the functions of English progressives in course-books used in German schools carried out by Römer (2005: 275) allows the author to conclude that there are discrepancies and misrepresentations between the information in such textbooks and that drawn from corpus-driven analyses. Römer suggests that these EFL materials were inadequate and offered a simplified picture of this particular case.

The fact remains that various criteria and pedagogical policies have become part of corpus-based reference works alongside the authenticity offered by CL itself. Contemporary lexicographers, grammarians, textbook writers vary in their degree of adherence to the relevant authentic language available in corpora.

6. The Direct approach

A further question would concern the direct approach. If recently published reference books and textbooks rely more heavily on corpus research (Römer, 2008: 116) and are commonly used by teachers, what is the point of introducing any further corpus methodology in the classroom? In this respect, Sinclair suggests that a methodology that incorporates corpus should be seen as irreplaceable, even if the corpus selected by the instructor does not meet very strict standards (2004b: 288; see also Johns, 1991):

> Will the corpus be 100% reliable, comprehensible and representative? Of course not, but do your present textbooks match these targets? Or your reference grammars and dictionaries? Or any native speaker models? Or any combination of these? Of course not. Any source of information about language has to be evaluated carefully, but at least you will know what is in your corpus and where it came from; what is more, if any patterns or usage occurs more than once from apparently independent sources then there is a very strong possibility that it is a regular pattern in the language.

Bernardini (2004: 17) has suggested that even recent corpus-based descriptive grammars like *Longman Grammar of Spoken and Written English* (Biber et al., 1999), which have displaced more traditional grammars, and offer information about frequencies of usage do not rival the direct observation of language patterning through corpora. Römer (2008: 120) views access to corpora, if not as a substitute for other teaching methods, at least as complementary to them. There is now extensive bibliography (see McEnery and Wilson, 1996, 1997; Biber et al., 1998; Sinclair, 2004a; O'Keeffe, McCarthy and Carter, 2007; Römer, 2008; Aijmer, 2009; Campoy-Cubillo et al., 2010) offering teachers ideas and suggestions to make more efficient use of corpus as an ordinary part of their classroom methodology, particularly in higher education.

We need to return to a central argument which underlies corpus analysis and which, in turn, has consequences in our teaching. It was stated earlier that native speakers' intuitions are looked at with suspicion with reference to language use. The situation may be aggravated where the teacher in charge is, like the learners, an L2 speaker of the target language. Such a situation is far from exceptional, since probably for the vast majority of English teachers around the world English is an L2. Fuster (2010: 273) points out that descriptive foreign language teachers in higher education should be aware that their subject matter is being addressed to non-native speakers whose competence in the target language is always more limited than that of the trained native speaker. There is also sufficient research literature that testifies to the problems highly advanced learners experience concerning delicate linguistic questions. Put simply, in the foreign language classroom, the controversy among theoretical linguists as to whether one should rely on introspection or adopt an empirical approach is simply fallacious and pointless. For instance, Bernardini (2004: 16-7) mentions the difficulties advanced learners have mastering article usage. Indeed, the most frequent and basic words, typically learned at the earliest stages prove the most difficult to master, as a result of the multiple meanings they have or functions they perform. It goes without saying that in such cases on corpus observation is not just a matter of choice, but very close to an obligation. For Braun (2005: 48) the potential for the use of corpora in language teaching comes from the following assets:

1. realistic, showing language in real use;
2. rich, providing more (and more diversified) information than dictionaries or reference grammars can;
3. illustrative, providing actual patterns of use instead of abstract explanations;
4. up-to-date, revealing trends in language use and evidence for short-term historical change.

Some recent contributions have manifested the significance that the direct approach can have in higher education. Clavel and Fuster (2009) and Fuster and Clavel (2010) suggest that CL is more than welcome since it offers university students the opportunity to become autonomous learners (Boulton, 2009: 37). Various studies have underscored that a corpus approach fosters the students' role as active agents of their own learning, where the teacher becomes a mediator (McEnery and Wilson, 1997: 6). However Boulton notes that there is insufficient empirical evidence to support claims about its effectiveness (2009: 38).

It is our view that in order to address the suitability for teaching descriptive language disciplines, specific contextual factors should be carefully considered (see Fuster 2010). The volumes of Biber et al. (1998), or McEnery and Wilson (1996) have emphasized, to a greater or lesser extent, the usefulness of corpus technology in descriptive language courses. Biber et al. suggest that stimulating corpus-based activities can be designed for practically any language discipline (1998: 12). However, very few studies set out to consider issues related to the effectiveness and reception of corpus technology. We claim that both these aspects need to be seen in the light of local teaching conditions and context in order to assess the feasibility of incorporating corpus as part of our classroom methodology.

7. Data-driven learning (DDL)

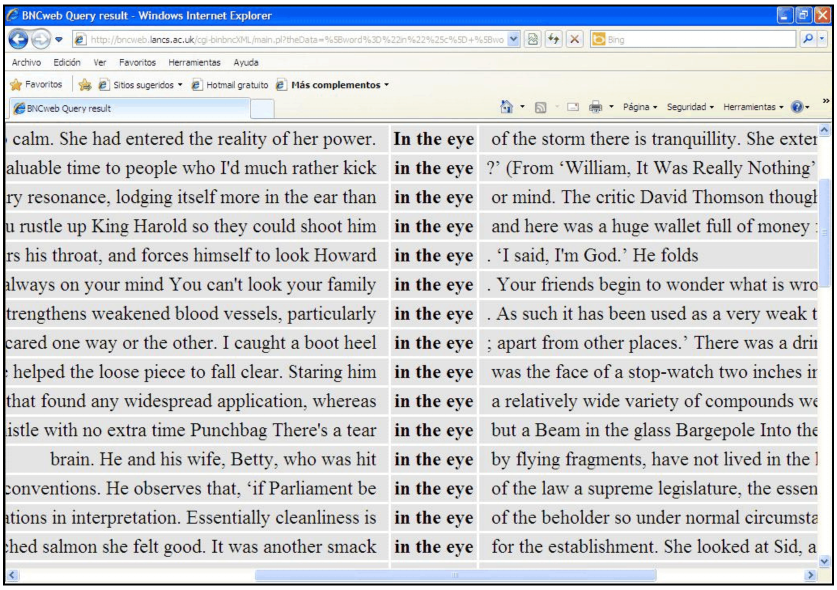
In a direct approach where corpus can have an important place, teachers are not necessarily dispensable. For Aijmer, teachers in higher education may act as mediators, guiding the students to the use of corpora so that once they are properly trained, they may come to consider corpus consultation as normal as looking up words in dictionaries or the use of grammars in solving queries about syntax (2009: 8-9). Initiated by Tim Johns, Data Driven Learning (DDL) is most probably the most well known direct pedagogical approach or methodology. What lies at the heart of this proposal is bringing direct corpus exploration into the classroom. DDL aims at turning students into active agents of their own learning process. Students are given the chance to learn by inductive acquisition of grammatical rules or regularities by way of the analysis of concordances yielded by corpora (see Partington, 1998: 6). The tenets of a DDL proposal may be summarised as follows (Johns, 1991):

1. The learner 'discovers' the foreign language through their own questions. The language-learner is essentially a researcher whose learning needs are driven by access to linguistic data, thus 'data-driven learning'.
2. The computer acts as an informant which answers the questions learners ask 'themselves'.
3. The basic computer tool is the concordancer. The concordances on a computer screen typically exhibit all instances of a word (or phrase) in the selected corpus in a *key word in context* (KWIC) format, which recovers the element of our query (the 'node') at the centre of each concordance line with surrounding cotext. (see the concordance shown in figure 1 for 'in the eye').
4. The teacher's role becomes that of a 'facilitator' of student-initiated research. Concordancers are the important focus.

Johns's own experience with concordances drawn from corpora in language teaching allows him to conclude that these (1991):

1. Stimulate enquiry and speculation on the part of learner
2. Develop the learner's ability to see patterning in the target language

3. Data come first. Students form generalisations to account for patterning through an inductive approach. This opposes traditional language teaching which works deductively
4. Through corpora students may discover rules or patterns unknown to teachers, textbooks or reference works.


BNCweb Query result - Windows Internet Explorer
http://bncweb.lancs.ac.uk/cgi-binbncXML/main.pl?theData=%5Bword%3D%22in%22%25c%5D+%5Bwo
Bing
Archivo Edición Ver Favoritos Herramientas Ayuda
Favoritos
Sitios sugeridos
Hotmail gratuito
Más complementos
BNCweb Query result
Página
Seguridad
Herramientas

calm. She had entered the reality of her power.	**In the eye**	of the storm there is tranquillity. She exte
aluable time to people who I'd much rather kick	**in the eye**	?' (From 'William, It Was Really Nothing'
ry resonance, lodging itself more in the ear than	**in the eye**	or mind. The critic David Thomson thoug
u rustle up King Harold so they could shoot him	**in the eye**	and here was a huge wallet full of money
rs his throat, and forces himself to look Howard	**in the eye**	. 'I said, I'm God.' He folds
lways on your mind You can't look your family	**in the eye**	. Your friends begin to wonder what is wro
trengthens weakened blood vessels, particularly	**in the eye**	. As such it has been used as a very weak t
cared one way or the other. I caught a boot heel	**in the eye**	; apart from other places.' There was a dri
helped the loose piece to fall clear. Staring him	**in the eye**	was the face of a stop-watch two inches i
that found any widespread application, whereas	**in the eye**	a relatively wide variety of compounds w
istle with no extra time Punchbag There's a tear	**in the eye**	but a Beam in the glass Bargepole Into th
brain. He and his wife, Betty, who was hit	**in the eye**	by flying fragments, have not lived in the
conventions. He observes that, 'if Parliament be	**in the eye**	of the law a supreme legislature, the essen
tions in interpretation. Essentially cleanliness is	**in the eye**	of the beholder so under normal circumsta
hed salmon she felt good. It was another smack	**in the eye**	for the establishment. She looked at Sid, a

Figure 1. Sample of concordance for the query in the eye retrieved from BNC, using BNCWeb.

In addition, we need to distinguish between corpus *tools* and *methods* in language teaching (Römer, 2008: 113): (1) Corpus tools are the actual collections of texts and software packages for corpus access and (2) Corpus methods: the analytic techniques that are used when we work with corpus data.

Römer (2008: 120) claims that a number of works prove the effectiveness of DDL: "These studies demonstrate that corpora nicely complement existing reference works and that they may provide information which a dictionary or grammar book may not provide."

Consequently, various options in Barlow's classification would qualify as DDL. But the way Johns (1991) describes DDL gives the impression that a more genuine approach to this methodology does not assign the language teacher a leading role in his/her interaction with a corpus. Teachers are variously referred to as mediators, facilitators or coordinators, but learners have the lions' share of the research, learning to learn through activities "that involve the observation and interpretation of patterns of use" (Bernardini, 2004: 16-7).

In our view the teacher's role in DDL is an unresolved issue, and most of what is seen in the literature reflects different approaches to this methodology. A more open version of DDL has been considered as beneficial. Certainly, depending on specific purposes, the teacher's role can become much more relevant than just acting as a 'mediator'. If DDL is seen in this light, the rich potential of corpus tools allows teachers themselves to access the intended corpus in order to offer activities "tailored to their learners' proficiency level and their particular learning needs (Römer, 2008: 120).

8. Implementation: problems and policies

Corpus exploitation has been found beneficial in foreign language teaching through direct approaches. For instance, Römer's (2005) conclusion of her own survey about corpus use by qualified English language teachers at secondary schools in Germany reveal that "many of the problems teachers have could be solved, at least, partially, if they were introduced to some of the basic corpus resources and received more support from corpus researchers." (Römer, 2009: 95). Incidentally, this also testifies to DDL's lack of implementation at these stages. Indeed, recent studies show that in spite of all the promising results, its implementation is still an unresolved issue. Aijmer claims that "the direct exploitation of corpora in the EFL classroom is unusual and the impact of corpora on syllabus and materials design has been slight" (2009: 2). In this respect, Braun (2007) notes the existence of an important contrast between tertiary and secondary pedagogical practices. Whereas most empirical work with DDL is largely confined to higher education, progress in the introduction of corpora in schools is still meager (Braun, 2007: 307-8; Römer, 2008: 123). Römer (*ibid*: 123-4) and Braun (*ibid*: 50) find it expedient to design policies in order to create DDL-friendly environments. But to do so efficiently, first and foremost we need to identify the main obstacles. The problems concerning DDL implementation may be seen as essentially affecting four areas: (1) technology; (2) training (3) methodology and (4) the addressee.

Firstly, there are technological problems related to the lack of pedagogically adequate concordance programmes (Krishnamurthy and Kosem, 2007: 369 and Braun, 2007). Here, software such as *Antconc* together with free online access in academic institutions to relevant scholarly publications might contribute to the success of corpora exploitation (Fuster, 2010: 270-1; also Römer and Wulff, 2010: 103). Moreover, while more specific corpora can be easily compiled for the ESP or EAP classroom, suitable contemporary reference corpora are beyond the reach of individual teachers. It is well known that the vast majority of those monitor corpora belong to private publishing companies (see O'Keefe et al., 2007: 17). Permission to use these large corpora is granted almost exclusively to authors working for the publishers. Thus, we are left with few publicly available reference corpora which may comply with our requirements: *the Bank of English*, although only a demo is freely accessible, *the British National Corpus* or, in future, the *Corpus of Contemporary American English*. In addition, there are some academic corpora, such as *the Michigan Corpus of Academic Spoken English* (MICASE), and a few others (O'Keefe et al., 2007: 204). These can be accessed freely online through specific interfaces.

Secondly, there is the problem of lack of training. There are different types of corpora which might be useful (general, specialized, learner) and different types of online resources (dictionaries, grammars). In this respect, students may have difficulty in selecting the most appropriate corpus and/or resource for a particular query (Flowerdew, 2009: 395). Guiding teachers and learners and providing basic training in CL is crucial. Römer points out that "corpora are not simple objects" (2008: 123). The size of many freely available corpora is certainly difficult to handle even with the appropriate software. It has also been suggested that some interfaces appear to be more adequate for research than for the classroom use (Krishnamurthy and Kosem, 2007: 368). A case in point is the *BNCWeb*, which has enormous potential (Leech, 2008: xiii) but training students to get the most of it requires considerable time. Nevertheless, *BNCWeb* offers teachers the option of extremely sophisticated guided classroom searches which cannot be performed through other interfaces (Fuster, 2010: 272). Also, it has been noted that corpus concordances can be 'messy', ambiguous or misleading and careful interpretation is required (Braun, 2005: 50; Fuster, 2010). Teachers need guidance to read concordances and advice on what types of DDL exercises they could create.

The third problem is perhaps harder to solve since it relates to methodological choices. Learning through the use of concordances may run counter to well-established teaching practices. Not all applied linguists seem to be in favour of direct exploitation methods. Flowerdew (2009) finds that opposition to DDL has centred around three problems:

1. Truncated concordance lines are examined atomistically from a bottom-up perspective.
2. Corpus data are decontextualised.
3. Corpus-based learning is typically inductive.

For Flowerdew nothing prevents us from combining bottom-up and top-down procedures. As to the problem of decontextualisation, she points out that this is not entirely true, and that more co-text can be recovered whenever it is deemed necessary, and even metatextual information can be accessed. The third point however seems to be more complex. Flowerdew acknowledges that the existence of different learning styles certainly means that DDL is not adequate for every type of student (2009: 406):

> Field-dependent students who thrive in cooperative, interactive settings and who would seem to enjoy discussion centering on extrapolation of rules from examples may benefit from this type of pedagogy. However, field-independent learners who are known to prefer instruction emphasizing rules may not take to the inductive approach inherent in corpus-based pedagogy.

Finally there is the problem of the learner and his/her context. A frequent claim in the literature is that direct approaches are more effective with advanced learners, although Boulton (2009: 51) reports positive results with intermediate students (see also Fuster, 2010). But it remains true that advanced learners' understanding of the authentic language shown in concordances makes them ideal candidates as practitioners. This student profile is typically found in higher education. Flowerdew (2009: 407, after Gardner, 2007: 255) observes that concordancing:

> (...) presupposes that learners will know most of the words (cotext) that surround a key word or phrase in context (KWIC), and that they can connect their meanings — an assumption that seems unreasonable for many groups of language learners (children, beginning L2 learners, learners with low literacy skills etc.).

Perhaps this was only to be expected since the implementation of CL in secondary education is close to non-existent in the context of foreign language teaching. This might be due not just to the probable lack of more innovative teacher training courses, but to the low competence of learners, which renders DDL ineffective in the eyes of English teachers (Fuster, 2010: 270).

However, there is no reason not to adopt direct approaches in higher education, since this is only to be expected with the new profile of student that is being promoted. There is hardly any doubt that within this new framework:

> Learner-centered teachers are guides, facilitators, and designers of learning experiences. They are no longer the main performer, the one with the most lines, or the one working harder than everyone else to make it all happen. The action in the learner-centered classroom features the students (Weimer, 2002: xviii)".

9. Concluding remarks

We have set out here to analyse the rationale of corpus research within theoretical linguistics and its pedagogical potential. An understanding of the empirical principles which have inspired CL from its initial stages is essential to capture what is implied in its later developments. The overview of its indirect applications shows that CL has exerted an enormous influence on a whole new generation of lexicographers, grammarians and perhaps, to a lesser extent, material designers. It has been shown that while DDL may offer the student valuable insights into language use, it has not become firmly established. For Flowerdew (2009: 411) there is still much to be discussed in the application of CL to foreign language pedagogy. Many applied linguists who are willing to adopt this methodology do not argue for a substitution of familiar resources, such as grammars, dictionaries, or textbooks in the classroom. Ideally students should be trained in CL at very early stages in higher education perhaps by proposing simple awareness-raising activities (see Fuster, 2010). While it is acknowledged that corpus analysis through direct approaches is fraught with difficulties, particularly when applied to earlier learning stages, there is no reason why it should not be introduced in higher education since its principles are compatible with what is expected of university students today. According to Dochy, Segers and Sluijsmans (1999: 332) university students should aspire to become self-regulated learners who have acquired relevant cognitive competencies:

> (...) such as problem solving, critical thinking, formulating questions, searching for relevant information, making informed judgements, efficient use of information, conducting observations, investigations, inventing and creating new things, analysing data, presenting data communicatively, oral and written expression (...) (*Ibid,* 332).

Direct approaches and particularly DDL activities in the framework of tertiary education can certainly contribute to promote more critical reflective practices through stimulating learner-centred teaching practices.

References

Aijmer, K. (ed.) (2009): *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Barlow, M. (1996): "Corpora for theory and practice". *International Journal of Corpus Linguistics* 1(1): 1-37.

Benson, M., E. Benson and R. Wilson (1986-1997): *The BBI combinatory dictionary of English. A guide to word combinations.* Revised Ed. Amsterdam and Philadelphia: John Benjamins.

Bernardini, S. (2004): "Corpora in the classroom: An overview and some reflections on future developments". In J. Sinclair, *How to use corpora in language teaching.* Amsterdam: John Benjamins. 15-36.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999): *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Biber, D., S. Conrad, and R. Reppen (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Boulton, A. (2009): "Testing the limits of data-driven learning: language proficiency and training". *ReCALL,* 21(1): 37-54.

Braun, S. (2007): "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora". *ReCALL* 19 (3): 307-328.

_______. (2005): "From pedagogically relevant corpora to authentic language learning contents". *ReCALL* 17(1): 47-64.

Campoy-Cubillo, M.C., B. Belles-Fortuño and L. Gea-Valor (eds.)(2010): *Corpus-based Approaches to English Language Teaching*. London: Continuum.

Clavel, B. and M. Fuster (2009): "El uso de las concordancias en la enseñanza del léxico inglés como herramienta TIC en la Educación Superior". In M. Cerezo and R. Grau, eds., *II Jornada Nacional sobre Estudios Universitarios. Los Nuevos Títulos de Grado: Retos y Oportunidades*. Castellón: Publicacions de la Universitat Jaume I, 1-11.

Dochy, F. , M. Segers and D. Sluijsmans (1999): "The use of self-, peer and co-assessment in higher education: A review". *Studies in Higher Education* 24(3): 331-350.

Fennell, B.A. (2001): *A History of English: A sociolinguistic approach.* Malden: Blackwell Publishing**.**

Flowerdew, L. (2009): "Applying corpus linguistics to pedagogy: A critical evaluation". *International Journal of Corpus Linguistics* 14(3): 393–417.

_______. (2004): "The argument for using specialized corpora to understand academic and professional language". In U. Connor and T.A. Upton, eds., *Discourse in the Professions: Perspectives from corpus linguistics.* Amsterdam: John Benjamins, 11-36.

Fuster, M. (2010): "The challenges of introducing corpora and their software in the English lexicology classroom: some factors". In I. Moskovich *et al.*, eds., *Language Windowing Through Corpora. Visualización del lenguaje a través de corpus.* Coruña: Universidad de Coruña, 269-288.

Fuster, M. and B. Clavel (2010): "Second language vocabulary acquisition and its pedagogical implications". In L. Pérez, I. Parrado and P. Tabarés, eds., *Estudios de Morfología de la Lengua Inglesa (V)*. Valladolid: Centro Buendía, Universidad de Valladolid, 205-212.

Graddol. D. (1997): *The future of English? A guide to forecasting the popularity of English in the 21st century.* London: The British Council.

Gries, S. T. (2010): "Corpus-linguistics and theoretical linguistics: A love-hate relationship? Not Necessarily". *International Journal of Corpus Linguistics* 15(3): 327–343.

Halliday, M.A.K., W. Teubert, C. Yallop and A. Èermáková (2004): *Lexicology and Corpus Linguistics. An Introduction*. London and New York: Continuum.

Hoffmann, S., S. Evert, N. Smith, D. Lee and Y. Berglund Prytz (2008): *Corpus Linguistics with BNCweb-a Practical Guide*. Frankfurt am Mein: Peter Lang.

Hornero, A.M., M.J. Luzón and S. Murillo (eds.) (2006): *Corpus Linguistics: Applications for the Study of English.* Frankfurt am Mein: Peter Lang.

Johns, T. (1991): "Should you be persuaded – two samples of data-driven learning materials". In: T. Johns and P. King, eds., *Classroom Concordancing*. Birmingham University: *English Language Research Journal,* 4: 1-16.

Kachru, B. (ed) (1982): *The Other Tongue - English Across Cultures*. Urbana, Ill.: University of Illinois Press.

_______. (1985): "Standards, codification and sociolinguistic realism: the English language in the outer circle". In R. Quirk and H. G. Widdowson, eds., *English in the World: Teaching and learning the language and literatures.* Cambridge: Cambridge University Press for The British Council.

Krishnamurthy, R., and I. Kosem (2007): "Issues in creating a corpus for EAP pedagogy and research." *Journal of English for Academic Purposes,* 6: 356-373.

Lavid, J. (2005): *Lenguaje y nuevas tecnologías: Nuevas Perspectivas, métodos y herramientas para el linguista del siglo XXI.* Madrid: Cátedra.

Leech, G. (2008): "Foreword". In S. Hoffmann, S. Evert, N. Smith, D. Lee, and Y. Berglund Prytz, *Corpus Linguistics with BNCweb-a Practical Guide*. Frankfurt am Mein: Peter Lang, xiii-xvi.

_______. (2007): "New resources, or just better old ones? The Holy Grail of representativeness". In M. Hundt, N. Nesselhauf, and C. Biewer, eds., *Corpus Linguistics and the Web.* Amsterdam and New York: Rodopi, 133-150.

McEnery, T. and A. Wilson (1997): "Teaching and language corpora". *ReCALL* 9(1): 5-14.

_______. (1996): *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Keeffe, A., M. McCarthy and R. Carter (2007): *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Partington, A. (1998): *Patterns and Meaning*. Philadelphia: John Benjamins.

Prodromou, L. (2003): "In search of the successful user of English". *Modern English Teacher,* 12(2): 5-14.

Renouf, A., A. Kehoe and J. Banerjee (2007): "WebCorp: an integrated system for web search". In M. Hundt, N. Nesselhauf, and C. Biewer, eds., *Corpus Linguistics and the Web.* Amsterdam and New York: Rodopi, 47-68.

Römer, U. and S. Wulff (2010): "Applying corpus methods to writing research: Explorations of MICUSP". *Journal of Writing Research,* 2(2): 99-127.

Römer, U. (2009): "Corpus research and practice. What help do teachers need and what can we offer?" In Aijmer, K., *Corpora and Language Teaching*. Amsterdam: John Benjamins, 83-98.

_______. (2008): "Corpora and language teaching". In A. Lüdeling and M. Kytö, eds., *Corpus Linguistics. An International Handbook* (volume 1). [HSK series] Berlin: Mouton de Gruyter, 112-130.

_______. (2005): *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English progressive Forms, Functions, Contexts and Didactics.* Philadelphia: John Benjamins.

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

_______. (2004a): *How to use corpora in language teaching.* Amsterdam: John Benjamins.

_______. (2004b): "New evidence, new priorities, new attitudes" In J. Sinclair, *How to use corpora in language teaching.* Amsterdam: John Benjamins.271-299.

Teubert, W. (2010): "Our brave new world". *International Journal of Corpus Linguistics* 15(3): 354-358.

_______. (2004): "Language and Corpus Linguistics". In M.A.K. Halliday, W. Teubert, C. Yallop, A. Èermakova and R. Fawcett, *Lexicology and Corpus Linguistics. An Introduction*. London and New York: Continuum., 73-112.

Teubert, W. and A. Èermakova (2004): "Directions in corpus linguistics". In M.A.K. Halliday, W. Teubert, C. Yallop, A. Èermakova and R. Fawcett, *Lexicology and Corpus Linguistics. An Introduction*. London and New York: Continuum, 113-165.

Tognini-Bonelli, E. (2001): *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Weimer, M. (2002): *Learner-centred teaching: Five Key Changes to Practice*. San Francisco: Jossey-Bass.