Chapter XIV
EFL through the Digital Glass of Corpus Linguistics

Vander Viana
Catholic University of Rio de Janeiro, Brazil

Sonia Zyngier
Federal University of Rio de Janeiro, Brazil

ABSTRACT

Like the advent of the telescope, computers today can provide ways of looking into language patterns that cannot be seen with the naked eye. From this perspective, this chapter argues for the centrality of corpus use in English as a foreign language (EFL) classrooms. It shows how the computer can offer new ways of looking at language and of relying on real data to see how language is used. A historical background is provided so as to enable an approach to corpus linguistics, one which moves away from reliance on intuitions and abstract examples. Having made the claim for the strengths of corpus linguistics as a way to develop students' autonomy in language learning, an online corpus and some concordancers are provided, and examples are offered of how they could work in a classroom. The chapter ends with research and educational prospects in the area.

INTRODUCTION

In *Through the Looking-Glass,* Alice suggests a new way of perceiving the world. She tells the cat: "First, there's the room you can see through the glass—that's just the same as our drawing-room, only the things go the other way" (Carroll, 1972, p. 181). This chapter follows the example by going through the looking-glass of language to see what can be revealed. As Sinclair (1991) stated, "the most exciting aspect of long-text data-processing, however, is not the mirroring of intuitive categories of description. It is the possibility of new approaches, new kinds of evidence, and new kinds of description" (p. 36). Later, he added that "we must gratefully adjust to this new situation

[the growing evidence of language in use] and rebuild a picture of language and meaning which is not only consistent with the evidence but also exploits it to the full" (Sinclair, 2004d, p. 10).

Having said that, for a long time, the input available to language students had been restricted to what could be contained in a single book. As Frankenberg-Garcia (2005) explains:

Not very long ago, learners had to be content with little more than traditional dictionaries, which focused more on what words meant than on how they were used…Language learners wishing to overcome this limitation had to rely to a large extent on what native speakers believed sounded right. (p. 335)

Along the same line of thought, Johns (1991) had already stated that most grammars used to be based on intuitions rather than on authentic data. In sum, in those days, students had to rely on their course books, reference books, or teachers as the only sources of learning and solving their doubts (Aston & Burnard, 1998, p. 20).

That was not an ideal world for at least two reasons. First, no matter how complete a textbook was, it could not possibly keep up with the constant changes of a given language. This means that most course books ran the risk of becoming anachronisms when made available to the market. There are also other issues these publications did not address explicitly, such as the kind of language they used as models (McEnery & Wilson, 1996, pp. 103-104), the choice of these models, the extent to which they represented the different registers students would have to deal with, the making up or editing of texts, and the extent of their authenticity. As a result, many of these textbooks did not reflect the real-world situations students would have to face when interacting with other people in English.

Second, research in the corpus tradition can now offer solid ground to claim that speakers' intuitions are not reliable enough in language teaching. Sinclair and Renouf (1988) illustrated this point by showing that the most frequent use of 'see' is not the action performed by one's eyes, as most speakers of English as a first language (henceforth EL1) would have it, but in the expressions 'you see' and 'I see'. Over a decade ago, Sinclair (1991) argued that "the problem about all kinds of introspection is that it does not give evidence about usage…Actual usage plays a very minor role in one's consciousness of language and one would be recording largely ideas about language rather than facts of it" (p. 39). The same view was later reinforced by Hunston (2002), who stated that "corpora can give information about how a language works that may not be accessible to native speaker intuition" (p. 13).

The problem with intuition and introspection is further explained by Tsui (2004) in that "they describe what people *know about* language, or what they *perceive* language to be, rather than how language *is used*" (p. 39). It is true that speakers can judge whether sentences are well-formed based on grammatical rules. However, this is not enough, especially if we understand form and meaning to be inseparable (Stubbs, 1993, p. 2). When it comes to the possible combination of words, grammatical rules may not explain what occurs. Besides, they do not tell us anything about word or pattern frequency in a given register (Granger, 2002, p. 4). In a word, as Francis (1993) puts it, "the evidence of the ways in which language is really used is available in plenty, and there is no longer any need to invent example sentences in the time-honoured way" (p. 138).

Actually, the age of sole reliance on books and intuition began to fall with the rise of corpus linguistics in the 1980s. According to Meunier (2002), "corpus research has highlighted the patterned nature of language, both lexically (collocations, recurrent word combinations) and grammatically or syntactically" (p. 121). Similarly, Granger (2002) argued that "it is quite clear therefore that the enriched description of the target language provided by native corpora is a plus for foreign

language teaching" (p. 21). With the compilation and later availability of corpora, both students and teachers began to have access to real language use. Language could then be presented in an authentic context. As Alderson (1996) put it, corpora "make it possible for learners not only to have access to real (rather than contrived) language data, but also to explore language data on their own and to generate their own hypotheses and rules about the language" (p. 258). Learners could now examine a specific word in its context and see the use of that particular word.

Therefore, we argue for the use of concordances in English language teaching following the discussions provided by Johns (1988, 1991), Tribble and Jones (1990), Hardisty and Windeatt (1989), among others. More specifically, the aim of this chapter is to show how the principles and tools in corpus linguistics may be used in the EFL setting. The focus here will lie on the resources available for free on the Internet, thus enabling anyone to make use of them.

This text is divided into four parts. In the first, corpus linguistics is discussed from a theoretical point of view before its use in language teaching is discussed. The second part presents an online corpus and some concordancers, together with some pedagogical applications to the language classroom. The third section proposes future work in this field of study. Conclusions are drawn in the last part of the chapter.

BACKGROUND

The use of language data in linguistic analysis goes back to the first half of the 20th century. At that time, American structuralists and ethnographers focused on the study of Indian languages. With no writing records, these languages could only be investigated by means of what was collected from oral data. Although Bloomfield (1933) argued for an inductive research method based on actual language analysis, the corpora used in those days, consisting of only a few sentences, differed widely from the multi-million-word corpora available nowadays.

In the second half of the 20th century, the inductive method proposed by Bloomfield (1933) was replaced by a deductive approach to language.[1] Chomsky, an influential linguistic at that time, changed the focus of linguistic research in the U.S. and around the world by advocating a mentalist view[2] to language studies. His perspective contrasted with the previous empirical approach. In Chomsky's (1957) terms, the focus was not on performance—that is, an individual's use of language—but on competence,[3] the knowledge one has about a given language. Language competence was seen as dependent on an innate faculty which all human beings are born with (Chomsky, 1999). In other words, emphasis was placed on the way speakers processed language cognitively. Generativism aimed at mapping individual competences with a view to finding linguistic universals which would be valid to each and every language in the world. Chomsky's concepts of 'competence' and 'performance' bore some similarity to Saussure's (1916) distinction between '*langue*' and '*parole*' (Lyons, 1981; Weedwood, 2002). 'Competence' would pair up with '*langue,*' whereas '*parole*' would correspond to 'performance'.[4] Both Saussure and Chomsky claimed that it was impossible to study performance or *parole* given their fuzzy nature. To them, corpus-based studies would be aimless as they would not be able to account for all instances of a language. In addition, corpora would encompass hesitations, repairs, and mistakes. In this perspective, studies based on samples of language were left aside for some time.[5] At this time, linguistics was basically derived from made-up sentences (Stubbs, 1993, p. 8). As Sampson (2005) puts it:

Those of us who came of age as scholars of linguistics in the 1960s and 1970s were surrounded by people who urged that linguistics does not need empirical data, and that it gets on faster and more

efficiently if it bypasses painstaking observation of natural usage and relies instead on speakers' intuitive 'knowledge' of their language. (p. 16)

Chomskyan theory was based on an ideal speaker's intuition about language. What Chomsky meant by 'ideal speaker' and how he or she would be identified in practical terms is still unknown. This ideal speaker, actually a linguist in most cases, would be in a position to judge what was grammatical or ungrammatical, thus determining how language worked. Sinclair (1991) commented that "linguistics languished" by summarizing the setting as follows:

The tradition of linguistics has been limited to what a single individual could experience and remember. Instrumentation was confined to the boffin end of phonetics research, and there was virtually no indirect observation of measurement. (p. 1)

Change would only take place in the 1980s when linguistics was ready to turn back again to empirical data. However, this time, corpus-based investigations were greatly helped by the advances in technology, especially the design and popularization of personal computers. As a matter of fact, technology played an important role in bringing language data back to linguistic research: it put an end to the major criticism to earlier corpus-based work carried out by researchers who did not have the means to do so. According to McEnery, Xiao, and Tono (2006), "Using paper slips and human hands and eyes, it was virtually impossible to collate and analyze large bodies of language data" (p. 4). One example is the compilation of the Survey of English Usage, which was used for the writing of *A Comprehensive Grammar of the English Language* (Quirk, Greenbaun, Leech, & Svartvik, 1985).[6] Today, the use of computers makes this type of research easier, faster, and more reliable. The second advantage is that it is possible to investigate ever-growing corpora (McEnery et al., 2006, p. 4). It should be stressed that the word 'corpora' nowadays refers to collections of *machine-readable* material (McEnery & Wilson, 1996, p. 14; Aston & Burnard, 1998, p. 5).

One of the advantages of using corpus linguistics is that what can be obtained by observable data differs greatly from what intuition tells us. Sinclair (1991) explains:

Indeed, the contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language. (p. 4)

Corpus linguistics highlights language in use and leaves aside artificial or contrived examples (McEnery et al., 2006, p. 6). If, on the one hand, the accuracy of explanations it provides depends largely on the quality and the size of the corpus used, then, on the other hand, flaws in a corpus may be relevant either because they will not form a pattern and will thus not represent the language (variety/register) under study, or because they are so frequent that they are in fact characteristic of language use.

The notion of frequency is central to corpus linguistics. It is only by probing language with computers that it is possible to get to know how frequent its features are. Frequency then is a much more objective way to arrive at conclusions than intuition can offer. McEnery and Wilson (1996) agree with it when they state:

Empirical data enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalized cognitive perception of the language. (p. 87)

At the moment, corpus-based investigations have become worldwide and researchers have been compiling corpora to suit the most different needs. However, today's widespread use does not come without problems. Defining a corpus, for instance, has become complex. It is not any collection of texts as, for example, common sense would hold (Cowie, 1994, p. 265). In order to be considered a corpus, a collection of texts needs to contain a selection of naturally occurring language. The selection must follow certain principles and there must be some design criteria to match what researchers want to investigate. A corpus compiled for linguistic analysis should be representative of the language to be investigated (or of one of its varieties). It must also be computer-readable, that is, it must be formatted in a way that it can be probed with the help of computer programs.

In addition to allowing the researcher to look into language in new ways, corpus linguistics has an interdisciplinary nature. For instance, there are studies that combine corpus linguistics with translation (cf. Baker, 1993; Olohan, 2004), lexicology (cf. Halliday, Teubert, Yallop, & Cermakova, 2004), systemic-functional linguistics (cf. Thompson & Huston, 2006), and stylistics (cf. Zyngier, 2002; Semino & Short, 2004), to cite a few. As far as language learning is concerned, there are many more studies that investigate the potential of corpora (cf. Granger, 1998, 2004).

The vast number of corpus-based research carried out so far may lead us to conclude that corpora have impacted the academic world and created the need, for instance, of different journals such as the *International Journal of Corpus Linguistics* and *Corpora: Corpus-Based Language Learning, Language Processing and Linguistics,* among others. However, it seems that the interface between corpus linguistics and EFL language teaching still needs to be refined in a way that the research developed so far also reaches the EFL classroom.

The use of corpora in English language teaching has a number of advantages. First, it provides learners with natural input, that is, with language as it is used by EL1 speakers (Gavioli & Aston, 2001; Scott & Tribble, 2006). Second, it reveals patterns of use that remain obscure to the naked eye. It is only possible to identify most patterns with the help of large corpora combined with the strengths of computer tools. According to Scott and Tribble (2006), "The new hardware and software used in corpus-based methods are opening up exciting possibilities which could not have been envisaged without them" (p. 4). Third, it fosters a student-centered approach as the teacher guides the students to find the answers by themselves (Stevens, 1995; Meunier, 2002). As Johns (1991) puts it, "We simply provide the evidence needed to answer the learner's questions, and rely on the learner's intelligence to find answers" (p. 2). In other words, "Increasing access to corpora may modify the traditional role of the teacher as an authority about the use of the language to be learned" (Aston & Burnard, 1998, p. 43). Finally, it promotes autonomy (Benson, 2001; Bernardini, 2004; Kaltenböck & Mehlmauer-Larcher, 2005), one of the goals of language teaching.

One of the first volumes to exploit the relationship between corpora and language teaching was Tribble and Jones's (1990) *Concordances in the Classroom,* which offered practical applications to the pedagogical use of corpora. Although providing only 11 entries in the bibliography, the volume was highly important at the time it was published. It covered aspects such as how to design and run concordances, how to compile corpora, and how to use concordances in a pedagogical way both in language and literature teaching.

Later, Granger and Tribble (1998) argued for the use of learner corpora in EFL teaching. Similar to Tribble and Jones (1990), students were asked to carry out the analysis on their own. They were given a number of concordance lines in order to come to some conclusions about the accurate use of English. The difference was that these students were also given learner corpus data. Two sample tasks were offered in the article, focusing on (a)

'accept'/'possibility' and (b) 'important'/'critical'/'crucial'/'major'/'serious'/'significant'/'vital'. In both examples, students had a number of questions that either guided their studies or presented a common problem in EFL writing. They were then asked to take a careful look at some concordance lines to try to check what differences there were between the EFL and EL1 writing that had been analyzed.

Although also working with a learner corpus, Milton's (1998) aim differed from that of Granger and Tribble (1998). The former reported on the development of a computer program to be used by EFL learners whose components are:

- An error recognition (i.e., 'proofreading' or 'editing') exercise intended to sensitize learners to the most common or most 'serious' errors exposed by the first analysis;
- A hypertext online grammar designed to give context-sensitive feedback and based on these errors;
  Databases of the 'underused' lexical and grammatical phrases exposed by the second analysis, and made interactively available to learners from their word processor; and
- List-driven concordances that interact with text in these programs and databases (p. 192).

For instance, while proofreading an essay, users must decide what kind of error can be found in each line, if any. There is also the possibility of resorting to help, which is divided into four levels: 'grammar category', 'full explanation', 'show error', and 'answer'. At any time, users can have access to a detailed analysis on their progress.

Finally, attention is drawn to a number of books that have been dedicated to the use of corpora, wordlists, and concordances in language teaching (e.g., Wichmann, Fligelstone, McEnery, & Knowles, 1997; Aston, 2001; Granger, Hung, & Petch-Tyson, 2002; Kettemann & Marko, 2002; Aston, Bernardini, & Stewart, 2004; Sinclair, 2004a; Scott & Tribble, 2006).

MAIN FOCUS OF THE CHAPTER

There is no doubt that corpora are becoming increasingly more important in the EFL environment. As Milton (1998) suggests, learning cannot be completely effective without the help of corpora. Based on the discrepancy between the description of language by means of corpus-based investigation and one's intuition about language use, he states:

Conventional classroom methods are often inadequate for conveying to learners our growing understanding of language features, and inappropriate for providing learners full access to, or significant experience with, the features of target language behaviors and how particular features of their own production deviate from these targets. (p. 186)

The use of corpus data in the classroom enables students to master the language by themselves. The concept of language here means real language in use as in a communicative event. The importance of such approach is also highlighted by Granger and Tribble (1998) who hold that "the authenticity of the data ensures that learners are presented with samples of language which reflect the way people actually speak or write" (p. 201).

The introduction of such methodology demands a shift in perspective. The focus on form brought about by the use of corpus-based language teaching materials seems to be a necessary counterpart for the communicative approach because "it has been accompanied by a loss of accuracy, especially grammatical accuracy" (Granger & Tribble, 1998, p. 199). This means that the use of corpora in the classroom may help students become more accurate when it comes to language use, which will then help them express themselves more naturally in English.

This chapter suggests the use of tools that are available on the Internet and can be accessed by teachers worldwide. By this token, the proposal differs from Tribble and Jones's (1990), for they did not mention online corpora/concordancers simply because they were not available at the time their book was written. Granger and Tribble's (1998) article shows an interesting way by means of which EFL teachers can make use of learner corpora; however, these corpora may not be available to EFL practitioners and/or they may not have the necessary skills or time to compile their own corpora. Milton's (1998) suggestion may not be put into practice by EFL instructors due to the fact that they may not be in a position to purchase the program he refers to. This is why free-of-charge online tools are introduced here.

Online Corpus: Theoretical and Practical Aspects

The British National Corpus (BNC) is probably the largest and most renowned corpus available through the Internet nowadays.[7] Although it is not possible to download the corpus for free,[8] one can access it by means of wordlists and concordances either by the interface on its own site or by other links in the Web.

The BNC represents British English in use in the last quarter of the 20th century and consists of both written and oral texts.[9] The corpus, which contains more than 100 million words, "corresponds to roughly 10 years of linguistic experience of the average speaker in terms of quantity" (Aston & Burnard, 1998, p. 28). Its written part, accounting for 90% of the corpus, contains books, periodicals, brochures, advertising leaflets, letters, and essays to cite a few. Texts written to be spoken, such as political speeches, plays, and broadcast scripts, are also included in it. The oral part, standing for 10% of its words, contains transcriptions of conversations, lectures, tutorials, interviews, sales demonstrations, sports commentaries, radio phone-ins, among others (Burnard, 2007).

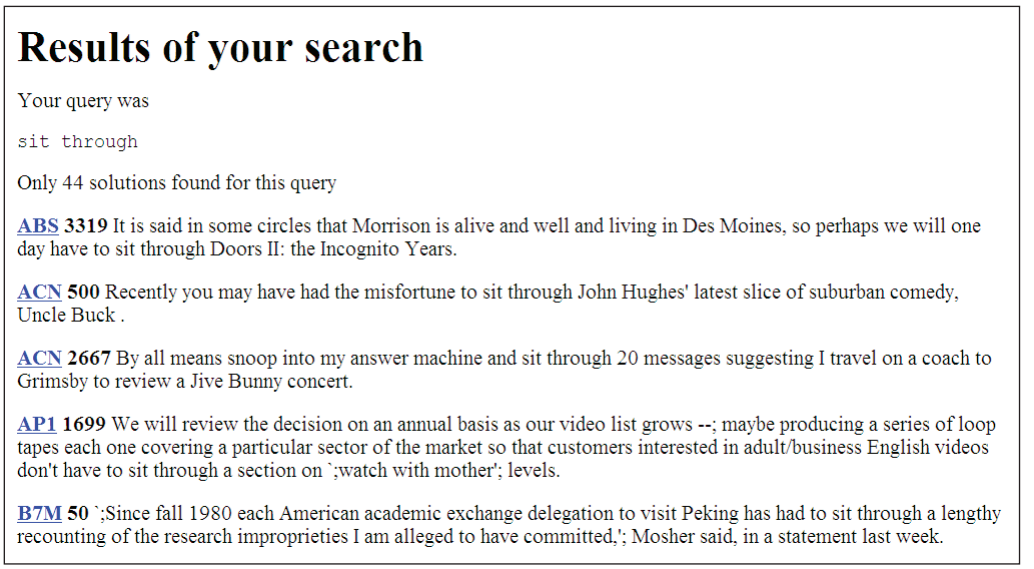
The BNC is a general corpus as it attempts to represent both the written and the oral registers, covering a wide range of subjects. Even though the final result is not balanced (90% written/10% oral), there is a major concern to include a variety of text types. The BNC may be described as monolingual since it only contains text in one language—English—and in only one variety—British English. When it comes to time aspect, it is a synchronic corpus for it represents a specific period of time, that is, the last part of the 20th century (from 1975 onwards). The only exception here remains with the selection of literary texts, which include texts dating back to 1960. However, such inclusion is justified by the compilers because of "their longer 'shelf-life'" (Burnard, 2007). The BNC is considered a static corpus since texts cannot be added or removed from it. Complete texts were included if they amounted to a limit of 45,000 words. When this limit was not met, only a sample of the text corresponding to the previously mentioned limit was included in the corpus.

From the BNC official site (*http://www.natcorp.ox.ac.uk/*), it is possible to assess the corpus for free. Users just need to type in what they are looking for, which can be a word or a sequence of words. It is also possible to restrict the search to a part of speech, for instance. The drawback of this system is that users need either to know the manual (see Burnard, 2007) by heart or to refer to it in order to find out how such restriction works and which codes are used for each and every part of speech.

The results page of the BNC adopts the format of full-sentence concordance as illustrated in Figure 1.

In this specific case, the search string was 'sit through'. As there are only 44 instances of such string in the BNC, all of them are shown in the results page.[11] The code on the left-hand side of the screen refers to the source from which the sample sentence has been taken. The major problem in this page is that the search words in the concor-

Figure 1. Results for 'sit through' at the BNC official site[10]



Results of your search

Your query was

sit through

Only 44 solutions found for this query

ABS 3319 It is said in some circles that Morrison is alive and well and living in Des Moines, so perhaps we will one day have to sit through Doors II: the Incognito Years.

ACN 500 Recently you may have had the misfortune to sit through John Hughes' latest slice of suburban comedy, Uncle Buck .

ACN 2667 By all means snoop into my answer machine and sit through 20 messages suggesting I travel on a coach to Grimsby to review a Jive Bunny concert.

AP1 1699 We will review the decision on an annual basis as our video list grows --; maybe producing a series of loop tapes each one covering a particular sector of the market so that customers interested in adult/business English videos don't have to sit through a section on `;watch with mother'; levels.

B7M 50 `;Since fall 1980 each American academic exchange delegation to visit Peking has had to sit through a lengthy recounting of the research improprieties I am alleged to have committed,'; Mosher said, in a statement last week.

dance lines are not highlighted in any way, nor are they in the middle of the line, which makes it difficult for their instances to be identified. In addition, users must read the complete sentences looking for the words themselves. So, it will be less likely that they will be able to concentrate on finding a pattern and more likely that they will focus on understanding the meaning of the sentences displayed.

In order to work with the BNC in a friendlier way in the classroom, it is advisable to use another interface, developed by Mark Davies at Brigham Young University, which can be accessed at *http://corpus.byu.edu/bnc/*. Although it may look difficult at first sight, the interface is easy to operate. The variety of the options it provides actually adds to its validity.

For instance, when searching for 'sit through', the output is different from the previous one reported above.[12] The format used in this case is the 'keyword in context' (KWIC), as can be seen in Figure 2.

These concordance lines are better to work with because 'sit through' is roughly in the middle of each line. It is also in bold type and underlined, helping users identify it and start studying its patterning.

Similar to the BNC site, in Davies's interface, it is possible to identify the source of each concordance line. In addition to the different layout, there is also some extra information on the source such as its genre, subject, and medium. Another difference is that it is possible to access more text than what is shown in the concordance line. Users may resort to an expanded context if they need to read more on a specific sample.

The most important aspect, however, of working with concordance lines is to start noticing patterns that would remain obscure in other ways. Therefore, while working with the concordance lines for 'sit through', one may notice its complements[13]:

...maybe producing a series of loop tapes each one covering a particular sector of the market so that customers interested in adult/business English videos don't have to sit through a section on 'watch with mother' levels.

'Since fall 1980 each American academic exchange delegation to visit Peking has had to sit through a lengthy recounting of the research improprieties I am alleged to have committed,' Mosher said, in a statement last week.

And I could never sit through all these interviews, transcribing them off the tape afterwards. It was too boring.

Maybe I'm just being traditional in my tastes; it is probably important to watch the compilation in small doses and not sit through the whole three hours as I did.

It was backs to the wall for most of the 90 minutes, and I hope I don't have to sit through too many games like that.

As can be seen in the five instances retrieved from the BNC, that which collocates with 'sit through' is generally something unpleasant and/or something that the person is not willing to do. As Hunston (2002) states, "Because it is often used with items that indicate something lengthy and boring, connotations of boredom tend to attach to the phrasal verb itself" (p. 141). This is generally referred to as semantic prosody, which is defined by Louw (1993) as "a consistent aura of meaning with which a form is imbued by its collocates" (p. 157). Therefore, it could be said that 'sit through' has a negative semantic prosody.

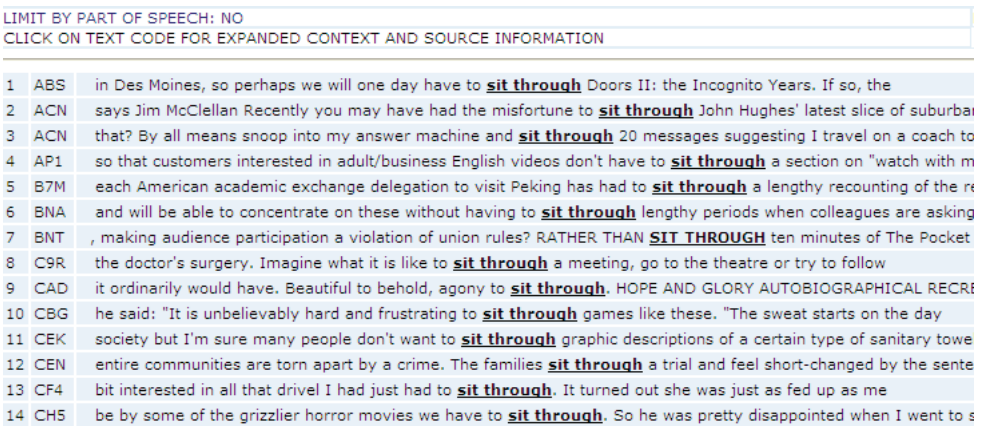
Another possible investigation may be carried out with collocations, as in the case of adjectives preceding nouns. One may also be interested in checking whether the word 'problem' has been placed with an adequate collocate in the following example[14]:

Nowadays there is a great gap between social classes. The majority of the population lives with less than one dollar a day, while the minority spends thousands of dollars buying foolish things in shopping malls. Two different realities, side by side, in the same world, in the same neighborhood. How to deal with this reality? How to manage with this terrible problem? [our emphasis]

In this case, the option 'adj.ALL' should be selected from a pull-down menu '(insert tag)' before the specific noun is typed as shown in Figure 3.

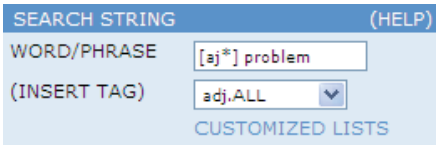
The pull-down menu offers an array of possibilities: nouns, verbs, adjectives, adverbs, negatives, articles, determiners, pronouns, possessives, prepositions, conjunctions, numerals, interjections, and punctuation marks. Most of these options are subdivided, supplying informa-

Figure 2. Results for 'sit through' in Davies's interface


LIMIT BY PART OF SPEECH: NO
CLICK ON TEXT CODE FOR EXPANDED CONTEXT AND SOURCE INFORMATION

1	ABS	in Des Moines, so perhaps we will one day have to **sit through** Doors II: the Incognito Years. If so, the
2	ACN	says Jim McClellan Recently you may have had the misfortune to **sit through** John Hughes' latest slice of suburba
3	ACN	that? By all means snoop into my answer machine and **sit through** 20 messages suggesting I travel on a coach to
4	AP1	so that customers interested in adult/business English videos don't have to **sit through** a section on "watch with m
5	B7M	each American academic exchange delegation to visit Peking has had to **sit through** a lengthy recounting of the re
6	BNA	and will be able to concentrate on these without having to **sit through** lengthy periods when colleagues are asking
7	BNT	, making audience participation a violation of union rules? RATHER THAN **SIT THROUGH** ten minutes of The Pocket
8	C9R	the doctor's surgery. Imagine what it is like to **sit through** a meeting, go to the theatre or try to follow
9	CAD	it ordinarily would have. Beautiful to behold, agony to **sit through**. HOPE AND GLORY AUTOBIOGRAPHICAL RECRE
10	CBG	he said: "It is unbelievably hard and frustrating to **sit through** games like these. "The sweat starts on the day
11	CEK	society but I'm sure many people don't want to **sit through** graphic descriptions of a certain type of sanitary towe
12	CEN	entire communities are torn apart by a crime. The families **sit through** a trial and feel short-changed by the sente
13	CF4	bit interested in all that drivel I had just had to **sit through**. It turned out she was just as fed up as me
14	CH5	be by some of the grizzlier horror movies we have to **sit through**. So he was pretty disappointed when I went to s

Figure 3. Search string for adjective + problem[15]



tion for different searches. For instance, in the case of adjectives, which are the focus here, users may choose to work with all adjectives (adj.ALL), adjectives in the comparative form (adj.CMP), or adjectives in the superlative form (adj.SPRL).

Table 1 contains the first 10 results for such query.

The most common collocate is 'major', occurring 4.01 times per million words in the BNC. The adjective 'terrible' does not appear on top of the list. As a matter of fact, it is in the 63rd place, totaling 16 instances and 0.16 times per million words. When its distribution across registers is checked, another difference can be noticed.

The first five registers in the list are all from the spoken continuum. The one that had the highest number of raw occurrences was 'broadcast discussion', totaling three instances. It is true that the complete list of registers contains written ones as well, but they refer to newspapers, biographies, and non-academic texts, which do not resemble the register of compositions.

By studying this information, EFL students may become aware of the registers in which the

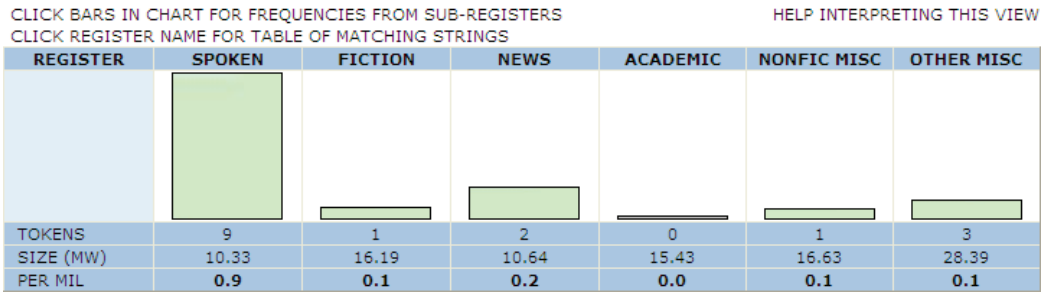
Table 1. First 10 adjectival collocations of 'problem'

Distrib	**Word/Phrase**	**Rokens Reg1**	**Per mil in Reg1 [100,000,000 words]**
1	major problem	401	4.01
2	main problem	301	3.01
3	real problem	287	2.87
4	only problem	249	2.49
5	serious problem	217	2.17
6	particular problem	215	2.15
7	big problem	125	1.25
8	biggest problem	112	1.12
9	social problem	85	0.85
10	further problem	82	0.82

Table 2. Distribution of 'terrible problem' across registers

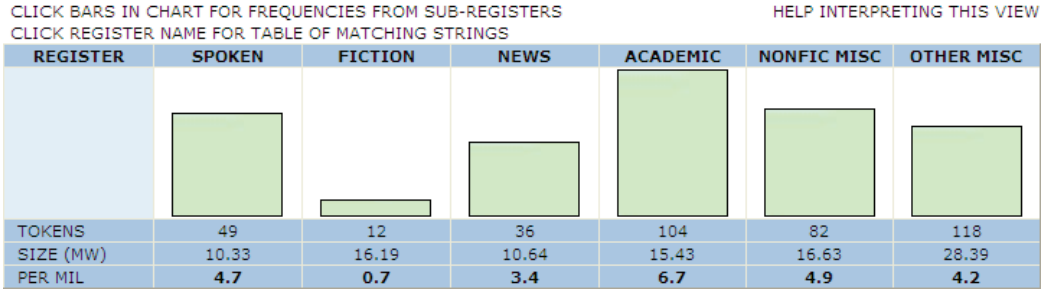
#	**Register name**	**# per million**	**# tokens**	**# words**
1	S_consult	7.2	1	138011
2	S_lect_soc_science	6.3	1	159880
3	S_brdcast_discussn	4.0	3	757317
4	S_brdcast_news	3.8	1	261278
5	S_unclassified	2.4	1	421554

Figure 4. Distribution of 'terrible problem' across registers (chart display)



REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
TOKENS	9	1	2	0	1	3
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**0.9**	**0.1**	**0.2**	**0.0**	**0.1**	**0.1**

Figure 5. Distribution of 'major problem' across registers (chart display)



REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
TOKENS	49	12	36	104	82	118
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**4.7**	**0.7**	**3.4**	**6.7**	**4.9**	**4.2**

search word is more generally used. If users select the 'chart' display, they may have the same information in a visual way.

Figure 4 shows that the academic register does not make use of 'terrible problem', whereas the spoken register is the one in which this word may be found most frequently.

The result found for 'major problem', the first combination of an adjective preceding the word 'problem', is strikingly different as may be observed in Figure 5. It seems that 'major problem' is used in more formal contexts since it is more commonly found in academic texts. As a matter of fact, it may be found 6.7 times per million words.

Davies's interface offers a number of other facilities to the EFL classroom, but only a few will be highlighted here. Instead of working with the whole BNC, users can search for specific registers according to their needs. The use of a word (sequence) can also be contrasted between two registers. Finally, it is also possible to search for words surrounding a specific search item in a span of 10 words to the left and to the right of the node.

In sum, Davies's interface facilitates the use of the BNC and helps the user to investigate collocations, colligations, semantic prosodies, and phraseologies. If students are asked to do this by themselves, they will probably learn what they search faster since they are doing it on their own instead of being spoon-fed by their teachers. As Frankenberg-Garcia (2004) argues, concordances are, in most cases, meaningful to learners as they are the ones who decide on what to investigate, that is, they choose what to look up in corpora based on their own language doubts. In the long run, it will also make them more autonomous and independent.

FUTURE TRENDS

As the discussions which have been carried out so far indicate, it is highly probable that corpus-based materials will become commonplace. Such change has already begun. In the field of lexicography, a number of English monolingual dictionaries are based on corpora. This means that words are objectively selected based on their frequency and not on lexicographers' intuitions. In addition, the order in which the meanings of a word are presented reflects their frequency. Dictionaries also offer specific hints based on what has been investigated in learner corpora.

Corpus-based grammars have also been developed for a couple of years now. The first corpus-based grammar was the *Collins COBUILD English Grammar,* as Sinclair (1990) explains:

The information in this book is taken from a long and careful study of present-day English. Many millions of words from speech and writing have been gathered together in a computer and analyzed, partly by the computer and partly by a team of expert compilers. It is the first grammar of its kind. (p. v)

Another corpus-based grammar is the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999). It is more sophisticated than the previous one because the description it provides includes four different registers, namely, conversation, fiction, news, and academic texts. The future prospect is that grammars will become even more specific when it comes to describing a language. That means linguistic features will be described in relation to the registers in which they are most frequently used.

The last type of publication to be informed by corpus is that of textbooks. The six-level course by Saslow, Ascher, and Kisslinger (2006) includes 'corpus notes' from the very first level, indicating the most common problems learners face while learning English. As described in the methodology for the course, it "provides concise and useful information about frequency, collocations and typical native speaker usage" (Saslow et al., 2006, p. Txiii). Another course book has been written by McCarthy, McCarten, and Sandiford (2005) which "draws on the *Cambridge International Corpus*, a large database of conversations and written texts, to build a syllabus based on how people actually use English" (back cover; (emphasis in the original). Other corpus-based textbooks are in line.

It seems that in some years to come, more computer programs will be made available to EFL teachers, facilitating the use of corpus data in the classroom. At the same time, it is expected that an ever-growing number of corpora will be on offer on the Internet so that they can be used by any EFL practitioner. In addition to the corpora already online, such as the BNC, the Bank of English, and the Michigan Corpus of Academic Spoken English, for instance, it is anticipated that new corpora will be either fully compiled or made available to public access.[16]

Having stated that, it is quite unlikely that corpora will reach the classroom if teachers are not offered special guidance. Those EFL teachers who have been away from university centers in the last few years may resist using corpora in their everyday classes because they are unaware of how to do it. Training sessions and/or courses are of the utmost importance so that corpora finally reach the classroom. Sinclair (2004b) warns us that "to make good use of corpus resources a teacher needs a modest orientation to the routines involved in retrieving information from the corpus, and—most importantly—training and experience in how to evaluate that information" (p. 2).

Research on how students actually react to corpus-based approaches is also needed and should be carried out more thoroughly. Although empirical classroom-based investigations in second language acquisition have shown that drawing students' attention to form yields better results

than implicit learning (Meunier, 2002, p. 120), it should be evaluated to what extent corpus-based materials also result in positive achievements. It should be investigated whether students actually want to work with concordances and whether they realize that corpus-based work makes them more autonomous. Research is also needed to assess the effects of corpora study on students' achievements.

Finally, it should be highlighted that as "corpus linguistics is still in its infancy" (Scott & Tribble, 2006, p. 3), a number of other possibilities are yet to come in the very near future.

CONCLUSION

This chapter has argued that intuitions about language cannot be considered authoritative if research wants to avoid false theories (Sampson, 2005, p. 16). Although empirical data was left dormant for some time under the influence of Chomskyan theory, it has come back with the development of new technology. In Sinclair's (2004b) words, "For a quarter of [a] century, corpus evidence was ignored, spurned and talked out of relevance, until its importance became just too obvious for it to be kept out in the cold" (p. 1).

The main advantage of corpora is perhaps that it allows linguistic research and language teaching to rely on empirical evidence. The use and study of real data opens up a new vista to those interested in the language phenomenon. Quoting Sinclair (1991), to whom corpus linguistics owes much of the standing it has today, "Without relinquishing our intuitions, of course, we try to find explanations that fit the evidence, rather than adjusting the evidence to fit a pre-set explanation" (p. 36).

Even when the evidence runs counter to what has been held for centuries, corpus linguistics is to be trusted as providing the way language really works. One can now look through the digital glass and see the language that people actually speak and write.

REFERENCES

Alderson, J.C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research* (pp. 248-259). London: Longman.

Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics, 1*(1), 93-111.

Aston, G. (Ed.). (2001). *Learning with corpora.* Houston: Athelstan.

Aston, G., Bernardini, S., & Stewart, D. (Eds.). (2004). *Corpora and language learners.* Philadelphia/Amsterdam: John Benjamins.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233-250). Philadelphia/Amsterdam: John Benjamins.

Benson, P. (2001). *Teaching and researching: Autonomy in language learning.* London: Longman.

Berber Sardinha, T. (2004). *Lingüística de corpus.* Barueri, SP: Manole.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Bloomfield, L. (1933). *Language.* New York: Holt, Rinehart & Winston.

Burnard, L. (Ed.). (2007). *Reference guide for the British National Corpus (XML edition).* Oxford: Research Technologies Service at Oxford University Computing Services. Retrieved May 20, 2007, from http://www.natcorp.ox.ac.uk/XM-Ledition/ URG/

Carroll, L. (1972). *The annotated Alice.* Middlesex: Penguin.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N. (1999). *O programa minimalista.* Lisboa: Editorial Caminho.

Cowie, A.P. (Ed.). (1994). *Oxford advanced learner's dictionary.* Oxford: Oxford University Press.

Francis, G. (1993). A corpus-driven approach to grammar—principles, methods and examples. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 137-156). Amsterdam/Philadelphia: John Benjamins.

Frankenberg-Garcia, A. (2004). Lost in parallel concordances. In G. Aston, S. Bernardini, & D. Stewart (Eds.). *Corpora and language learners* (pp. 213-229). Amsterdam/Philadelphia: John Benjamins.

Frankenberg-Garcia, A. (2005). A peek into what today's language learners as researchers actually do. *International Journal of Lexicography, 18*(3), 335-355.

Gavioli, L., & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal, 55*(3), 238-246.

Granger, S. (Ed.). (1998). *Learner English on computer.* London: Longman.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam/Philadelphia: John Benjamins.

Granger, S. (2004). Practical applications of learner corpora. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers: PALC 2003* (pp. 291-301). Frankfurt am Main: Peter Lang.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam/Philadelphia: John Benjamins.

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (p. 199-209). London: Longman.

Halliday, M.A.K., Teubert, W., Yallop, C., & Cermakova, A. (2004). *Lexicology and corpus linguistics.* London: Continuum.

Hardisty, D., & Windeatt, S. (1989). *CALL*. Oxford: Oxford University Press.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Johns, T. (1988). Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe, & H. Wekker (Eds.), *Computer applications in language learning* (pp. 9-27). Dordrecht: Foris.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *ELR Journal, 4,* 1-16.

Käding, J. (1879). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: privately published.

Kaltenböck, G., & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL*, *17*(1), 65-84.

Kettemann, B., & Marko, G. (2002). *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics, 29*(3), 333-347.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Amsterdam/Philadelphia: John Benjamins.

Lyons, J. (1981). *Language and linguistics: An introduction*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone 2: Student's book*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* London/New York: Routledge.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 119-141). Amsterdam/Philadelphia: John Benjamins.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer* (pp. 187-198). London: Longman.

Olohan, M. (2004). *Introducing corpora in translation studies.* London/New York: Routledge.

Quirk, R., Greenbaun, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Sampson, G. (2005). Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics, 10*(1), 15-36.

Saslow, J., Ascher, A., & Kisslinger, E.J. (2006). *Top Notch fundamentals: Teacher's edition and lesson planner.* New York: Pearson Longman.

Saussure, F. (1916). *Cours de linguistique general.* Paris: Payot.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education.* Amsterdam/Philadelphia: John Benjamins.

Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing.* London/New York: Routledge.

Sinclair, J. (Ed.). (1990). *Collins COBUILD English grammar.* London/Glasgow: Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. (Ed.). (2004a). *How to use corpora in language teaching.* Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. (2004b). Introduction. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 1-10). Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. (2004c). Preface. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers* (pp. 7-11). Frankfurt am Main: Peter Lang.

Sinclair, J. (2004d). *Trust the text: Language, corpus and discourse.* London/New York: Routledge.

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160). London: Longman.

Stevens, V. (1995). Concordancing with language learners: Why? When? What? *CAELL Journal, 6*(2), 2-10.

Stubbs, M. (1993). British traditions in text analysis—from Firth to Sinclair. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 1-33). Amsterdam/Philadelphia: John Benjamins.

Thompson, G., & Huston, S. (Eds.). (2006). *System and corpus: Exploring connections*. London: Equinox.

Thorndike, E. (1921). *A teacher's wordbook*. New York: Columbia Teachers College.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.

Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource book for teachers.* London: Longman.

Tsui, A.B.M. (2004). What teachers have always wanted to know—and how corpora can help. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 39-61). Amsterdam/Philadelphia: John Benjamins.

Viana, V.P. (2006). Modals in Brazilian advanced EFL learners' compositions: A corpus-based investigation. *Profile, 7,* 77-86.

Weedwood, B. (2002). *História concisa da lingüística*. São Paulo: Parábola Editorial.

Wichmann, A., Fligelstone, S., McEnery, A., & Knowles, G. (Eds.). (1997). *Teaching and language corpora.* London: Longman.

Zyngier, S. (2002). "Smudges on the canvas"? A corpus stylistics approach to Macbeth. In I. Biermann, & A. Combrink (Eds.), *Poetics, linguistics and history: Discourses of war and conflict* (pp. 529-546). Oxford: Oxford University Press.

Xiao, Z. (forthcoming). Well-known and influential corpora. In A. Ludeling, M. Kyto, & A. M. McEnery (Eds.), *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter. Retrieved September 11, 2007, from http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpus%20survey.htm

KEY TERMS

Colligation: The co-occurrence of a word with a pattern or the co-occurrence of two patterns in language use.

Collocation: The co-occurrence of two or more words within a short span in a text.

Competence: A concept proposed by Chomsky which refers to the knowledge an individual has about his or her first language.

Concordance: A listing of all the occurrences of a word (sequence) in a given corpus which gives access to language patterns.

Concordancer: A computer program by means of which corpora can be probed with a view of generating concordances.

Corpus: A collection of naturally occurring language text which can be read by computers and which has been selected to represent a particular type of language (variety/register) and with a specific research objective.

Corpus Linguistics: The linguistic study of large samples of language in use by means of computer tools.

Natural Language: Language as it is spoken and/or written by real speakers, that is, language which occurs in everyday use.

Performance: From a Chomskyan perspective, it refers to the use of language made by speakers.

Semantic Prosody: The general meaning a word assumes because of its collocates, which can be positive, negative, or neutral.

ENDNOTES

1 Although one may argue that Chomsky's influence in language studies has decreased considerably in the past, the differences between deductive and inductive approaches to languages are still worthy of attention. For instance, a number of recent publications on corpus linguistics revisit such distinction (e.g., Tognini-Bonelli, 2001, Berber Sardinha, 2004; McEnery et al., 2006). In addition, although corpus-based studies have become more popular, there are some scholars who still develop investigations based on Chomskyan tradition, as Sinclair (2004c) reminded us. More recently, Chomsky has argued in an interview that studies based on corpora do not mean anything (Andor, 2004).

2 This approach was not first proposed by Chomsky. In fact, Saussure (1916) had already worked with it in the beginning of the structuralist movement.

3 In Chomskyan theory, this concept refers to syntactic competence.

4 Nonetheless, it should be stressed that the concepts are, in fact, diverse. For instance, although being a mentalist, Saussure never argued for the existence of an innate biological factor.

5 Some areas, nevertheless, continued making use of such data. This is the case of phonetics and child language acquisition, for example, where introspection and intuition did not help much and/or could not be sought.

6 Some other pre-electronic corpora include those compiled, for instance, by Käding (1879) and Thorndike (1921).

7 The status of the Web as a corpus itself is a controversial issue. Some scholars argue that it may not be considered a corpus for a number of reasons, including the issues of representativeness and balance. There are, however, some researchers who argue for its use as a corpus. Such discussion, although of interest, is out of the scope of this chapter. For a better understanding of the latter tradition, see Kilgarriff and Grefenstette (2003).

8 Both the BNC XML Edition (containing the full content of the BNC) and the BNC Baby (containing a four-million-word selection from the BNC) can be purchased online at http://www.natcorp.ox.ac.uk/ getting/index.xml.ID=order.

9 An American correspondent of the BNC is being compiled. The ANC, American National Corpus, will total 100 million words of American English in texts from 1990 onward. By the time this chapter was being written, its second release, containing 22 million words, had already been made available. Additional information on this corpus may be found at http://americannationalcorpus.org/.

10 Data cited in all examples in this chapter have been extracted from the British National Corpus Online service (http://www.natcorp.ox.ac.uk/), managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

11 The maximum number of concordance lines displayed in this page is 50, despite how frequent the search string is.

12 Although the number of concordance lines shown does not apply to this case since there are only 44 instances of 'sit through' in the BNC, it should be mentioned that Davies's interface shows twice as many lines as the BNC interface, that is, 100 lines at a time. In addition, there is not a limit as there is at

the BNC official site. Here users can read as many concordance lines as they want to.

13 The emphasis in the following examples is ours.

14 This example has been retrieved from a corpus of compositions written by advanced students in private language courses in the city of Rio de Janeiro. More information on this corpus, which was compiled by the main author of this chapter, may be found in Viana (2006).

15 The layout of this interface may have changed in March 2008, but has not affected the information in this chapter.

16 For a description of existing corpora, see Xiao (forthcoming).