Corpus Linguistics: Corpora as an aid in the ESP classroom

Paschalina Kellartzi
Aristotle University of Thessaloniki, Greece

Abstract

The purpose of this paper is to demonstrate how electronic corpora can prove a useful tool in the EFL – and particularly ESP – classroom. It is useful for teachers looking for diversity and authenticity in context-based language instruction. The paper bases its discussion and suggestions on the Collins Cobuild Bank of English and proposes its use in the classroom either in research-orientated way or through teacher-generated tasks and activities.

Keywords: corpora, corpus linguistics, Bank of English, ESP, authentic material

1. Introduction

Teaching a foreign language primarily depends on the learners' attitudes and purposes. Among the millions of people learning English as a foreign language there is a large number who learn it for a very specific purpose. Learners have various motivations and attitudes towards the second language and, ideally, instruction has to accommodate learners' needs and priorities. In learning English for Specific Purposes, the motives are quite different from those of English as a foreign language (EFL) learners. To examine how corpora help in teaching English for Specific Purposes (ESP), it is vital to state that the motivation of ESP learners is often very strong. Especially in the light of the modern globalisation, the need to communicate in all fields becomes more and more persistent.

Contemporary corpus linguistics is primarily concerned with the exploration
and use of language issues mainly through lists of concordances that derive from a large selection of data. Corpus linguistics has lately undergone a significant change thanks to a “basic software tool used in data processing, a concordancer, a programme which identifies the pattern of a given word, a part of a word, or a combination of words, and displays a list of all its occurrences along with the immediate context known as a concordance” (Mallikamas 1999). With the aid of concordancers, searching corpora becomes fairly sophisticated. Possible queries include combinations of words – apart from isolated words – and corpora can provide lists of frequency and of the commonest collocates. Arriving at conclusions about linguistic features has always drawn linguists’ interest and various ways of using corpora in the classroom have been suggested.

In this paper some methods of incorporating corpora in ESP teaching will be examined. First, it will be explained why ESP classrooms are used as examples and why Collins COBUILD corpus has been chosen for reference. Most of what is suggested in the paper applies not only in ESP teaching but in any EFL environment. The paper intends to trigger teachers’ imagination on how to exploit electronic corpora in language learning. Thus, the paper suggests activities to teachers and at the same time demonstrates the possible uses of the features of modern electronic corpora. Among the possible corpus-based activities, gap-filling and guessing activities based on concordance lines (the corpus output) are mostly examined. Finally, the advantages of vocabulary building through the corpus and of learners' direct access to the corpus are examined.

2. Approach-delimitation

2.1 Principles - learners of ESP

In this paper, ESP students of medicine will be used for reference, since medical students are quite numerous and the study of English is fairly important for their studies. Medical students – and those of para-medic faculties – are likely to need access to English bibliography, to attend conferences held in English, or for a variety of other reasons. Regarding the age and level of students, it is assumed that the age of a typical group of ESP students ranges between 19-35 years old; these are mostly people who choose to enhance their current knowledge of English and combine it with their professional interests, or simply need English as an extra qualification.
2.2 COBUILD

The term 'corpus' in general refers to any large collection of data – for instance, the whole works of Shakespeare. There are special corpora: literature ones or those comprising of texts of specific jargon such as medical or engineering. Although literature corpora could be an extra tool in the teaching of languages, what is most important to foreign language instructors is the use of general language corpora.

The body of linguistic data is taken from many sources – both spoken and written – and represents as accurately as possible the particular language. It is taken here that corpora are annotated, that is, marked up with different types of information, such as author, genre, title etc. Language orientated users are likely to be more interested in the linguistic information – world class, syntactic structure, etc.

From the variety of English corpora currently developed, Collins COBUILD has been selected for the purposes of this paper. It is one of the largest corpora and has been developed and improved by the University of Birmingham and HarperCollins publishers since 1980. Since it is a monitor corpus, it is very up-to-date, with the majority of texts entered after 1990.

The most powerful feature of Collins COBUILD is the Bank of English (BOE). The Bank of English consists of over 450 million words from a variety of sources and different registers – written and spoken – and it is accessible only to subscribers, with a limited version, COBUILDirect, being available online for free. On entering the BOE, users are requested to select corpora sources. This facility has special interest for statistical analysis, as users can observe differences in the use of the same word when this is examined in various subcorpora.

2.3 Corpora and medical terminology

In the present study a general corpus is used for specific vocabulary, instead of a corpus specifically aimed at medical students – ESP learners. Thus, widely-used medical vocabulary such as vaccine, cervix, tumour – not to mention items such as infectious diseases, prevention, treatment – are commonly encountered in the corpus, whereas very specialized ones such as polyp, adenoma¹ and many more are usually rarely found in the general corpus.

¹ For adenoma, the BOE produced only one entry. Polyp also seems to be rarely encountered (15 instances).
2.4 Implications for ESP teaching

However, since ESP instruction is a branch of applied linguistics, in general, and of TEFL, in particular, a resourceful instructor is able to enjoy the advantages of corpora applications in EFL when teaching ESP as well. Besides, Stevens claims that “concordancing has been used effectively in ESP” and refers to “a vocational training lesson in which banking students concordanced words like *share, debenture, unit* etc. and then compiled the information in text files. Students reported results to the class at large, and found the concordance program “fascinating”” (Stevens 1995: 2-10).

One of corpora’s main functions is the users’ familiarisation with unknown vocabulary encountered in articles, or any materials used in the classroom. Most ESP instructors prefer using corpora for the usual gap filling, comprehension and similar vocabulary activities. Moreover, corpora are very effectively used when the teacher wishes to introduce new material.

The type of activities allocated by teachers depends on the degree of students’ direct involvement with the corpus. Mallikamas (1999) states that teachers can either let students explore the corpus directly or they can search the corpus themselves and produce corpus-based class activities. There is also the possibility that advanced students might be offered the opportunity to use the corpus in a research-oriented way. However, the activities mentioned in this paper are mostly teacher-generated corpus-based tasks where students do not have to process raw corpus output.

3. Corpus-based ESP teaching

A resourceful instructor can make the most of electronic corpora, by assigning different kinds of tasks to students. The extent to which the teacher will exploit the wealth of language produced by the corpus software depends on his/her imagination and creativity.

The general outline of corpus-based teaching, as summarized by Mallikamas (1999: 1-17) is that corpora provide an “inductive reasoning approach (that) consists of three phases: observation, classification and generalization”. These three steps can be applied on grammar or vocabulary tasks, provided that the teacher has prepared appropriate material for his/her students.

3.1 Analysis-exploitation of data

The below table displays some concordance lines for *uterus* which the teacher has processed for linguistic purposes. For instance, the collocates – words that
Table 1. Concordance lines processed by the ESP instructor for the word uterus

1) The blastocyst undergoes further growth inside the uterus, receiving oxygen and nourishment from secretions of...

2) Uterine contractions open the cervix or neck of the uterus, dilating it wide enough for the baby’s head to move

3) And also produces changes in the lining of the uterus that prevent implantation even if an egg should...

4) Disintegrating, it implants itself in the wall of the uterus. One of the surprising things about this very early

5) Hysterectomy, or surgical removal of the uterus is one of the most common operations performed in

6) The progesterone also affects the lining of the uterus, so implantation is unlikely even if fertilization...

7) Of implantation and transplanted into the wife’s uterus. If it transplants successfully, a normal pregnancy...

8) Involvement includes infections of the lining of the uterus (chlamydial endometritis) which occurs in about half

9) From the ovary down the fallopian tube and towards the uterus, where it gradually disintegrates and is expelled as

10) The embryo itself. When it touches the wall of the uterus, the outer shell of cells in the blastocyst breaks

11) Was suffering from an unexplainable bleeding of the uterus and was forced to return to New York to consult

12) Whether the anatomical absence of the uterus alone is responsible or whether there may be other...

13) Mass of cells that lies against the wall of the uterus. The placenta is fully developed by about four weeks

14) Flow patterns, chemical substances secreted in the uterus, etc.) is unclear at present. Aging does not appear

15) The risk of ovarian cysts and fibroid tumours of the uterus. In addition, women using oral contraception

appear most frequently with our key word – are in bold, so that learners can make observations about certain language structures. Also, some interesting verb-preposition combinations are in italics (i.e. suffering from), whereas special terminology words are underlined for emphasis. Consequently, by analysing the corpus output, teachers can draw learners’ attention to specific targets.

The data set above presents only a small part of the 164 lines found for uterus. The original BOE window is presented in figure 1 below. One can see at the bottom part of the window the number of lines found for the key word.
Considering that the BOE is a general corpus and **uterus** is a specific terminology item – thus somewhat rare – 164 instances are quite a lot; in this way one can see the size superiority of the corpus. That is, there is so much data in the corpus that users can look up any word, even rare ones. Users also have the possibility to access longer extracts and the source of each concordance line. Viewing the source of the key word and the full paragraph of the concordance line – the context – is an important facility for the users.

### 3.2 Gap filling

A simple activity that can be assigned would require students to fill in gaps that the teacher has blanked out. Providing 10-15 lines where the word in question has been blanked out, the ESP instructor can check learners’ understanding of the term, especially when the lines are presented together with those of potentially confusable terms – here **cervix** and **ovary**. These three parts of the female reproductive system are likely to be confused, especially by beginners not only in ESP, but also in medical studies. Experienced doctors and nurses should be able to know the differences, merely by means of context.

Possible gaps can be words or phrases of medical terminology – underlined in the data set – or anything the teacher wants to emphasise. It is up to the
teacher to decide which part of grammar or vocabulary to focus on. S/he can draw attention to special verbs – *secrete, disintegrate* – and others that can be used both in special and in general language contexts. Students can be asked to distinguish between these two types of verbs in the given lines, or to focus on structure instead and comment on forms such as *lies against* (line 13) and *implants itself* (line 4).  

The word *uterus* for instance, in combination with concordances of terms likely to be confused with – *cervix* (2), *ovary* (9), *placenta* (13) and *fallopian tube* (9) – can form separate gap filling or "guess the definition" exercises. The teacher can exploit the above set of examples by blanking out key terms in the specific field of medicine, such as *hysterectomy* (5), *transplants* (7), *cervix* (2), *secretions* (1) and many more, thus having students involved in examples taken from authentic texts.

Such use of the corpus can be applied to the part of the lesson where new material is introduced. Instead of presenting students with the chapters of grammar and vocabulary that are planned for the lesson of the day, the instructor can present students with authentic examples and elicit the new terms or structures from them. By being asked to observe the prepositions preceding *uterus*, students should come up with the *in/into the uterus* as the most common. Moreover, in the same context, they could be asked to fill prepositions in the gaps, after having studied some of the typical verb-preposition combinations – *secreted in, developed by, suffering from* (lines 14, 13 and 11 accordingly).

### 3.3 Collocates: Observation, classification, generalisation

An activity that can derive from the set of concordances is the deduction of conclusions about collocates. Teachers can provide students with all the results of the query and ask them to form a list of the ten most frequently encountered phrases – collocates. In the above case of *uterus* they should definitely come up with phrases such as *wall/lining of the uterus* (3/4), *removal* (5), *neck* (2) and *bleeding of the uterus* (11). They should be able to compare and comment on the words occurring more or less often with the item of study; *prolapse* and *contraction* for instance are among the rarest. *Uterus* features in a variety of collocations, as opposed to other medical vocabulary items that appear almost exclusively in fixed phrases.  

In particular, as can be seen in figure 2, the

---

2 The line numbers refer to the concordance lines in table 1, p. 3.

3 Search for *bone marrow* for instance, produced 105 entries that lack variety in context.
overwhelming majority of bone marrow instances occur with transplant — and less with donor.⁴

3.4 Guessing

3.4.1 The meaning

Students could be required to guess the meaning of the phrases, which may seem simple as they are used in the above data set, but in some cases combinations of words can really form collocates with surprising meanings. If specialized terms — such as chlamydial endometritis — were entered, guessing their meaning would not be a simple task; learners should be encouraged to elicit their meaning by resorting to the use of etymology, after having identified the root of the word.

⁴ On the same screenshot one can see the facility of viewing the context of the selected line, which can prove very useful. See also the source of the line — newspaper Sun- on the bottom part of the window.
3.4.2 The process

Assuming that the majority of ESP learners nowadays are in the first years of their vocational careers, often lacking real-world experience in their professions as well, we suggest that they might find helpful a task that helps them distinguish the different registers and the different processes described in the output. In the above data set, most of the lines describe processes – and in particular the pregnancy stages – whereas bone marrow data displays a limited variety of functions within text. Most bone marrow instances are references to transplantations, possibly taken from articles in newspapers, magazines, or television shows that deal with general issues and current affairs. At the instructor's prompt to guess the process, ESP learners should be able to make the distinction between lines 9, 10, 13 – in the data set – which describe the changes in the woman's genitals during pregnancy and 11 which is just a news report, that may be originated from any non-scientific text in the corpus. Overall, uterus is a term encountered in various texts – both scientific and simplified, aimed at wider readerships.

3.4.3 The context

When a corpus such as COBUILD's BOE is used as an aid to classroom-based teaching, an important tool that should not be neglected is the option to view the context of every entry. Part of the guessing activities should be dedicated to a “guess-the-context” task. In ESP more than in any other EFL environment, learners should be able to distinguish between the different registers, as not only accuracy but also appropriacy is their aim. When prospective doctors, nurses or midwives are found in an English-speaking environment they should be able to respond to the needs of the circumstances, i.e. they should be in a position to know when to use sophisticated or simplified language. In order to practice context-significant teaching, ESP instructors can ask learners to guess the context of selected lines and possibly the source, thus helping learners to understand special terminology. Teubert (1999: 4-19) states: “most texts (in corpora) occur as communication acts, that is, as the interactions between members of a language community”. Jordan (1997: 231) also stresses the importance of genre in ESP.

3.5 Techniques for vocabulary expansion - elicitation

3.5.1 Derivatives

Medical terminology is ideal for corpus research, providing ideal material for enhancement of students' awareness of derivatives. Provided that learners have
some experience on derivative-based tasks, they can be asked to practice on the special terminology: secret-secretion, implant-implantation, either by locating it in the corpus or by producing the parts of speech themselves. Learners can rarely appreciate the importance of derivatives in EFL, but in reality, producing derivatives provides a foundation for vocabulary expansion with the least possible effort. Especially in the field of medicine, where the unfamiliar words can hamper the comprehension of a text, derivatives can prove an important help – a source for deduction of meaning.

3.5.2 Etymology

Etymology can also help learners to cope with unfamiliar vocabulary. It is an ideal field to work on, it exploits features of both medical terminology and corpus study, and, after some practice, one can arrive at spectacular conclusions about the roots of lexical items. ESP instructors can make students realize through personal experience and research that some roots appear often in many different forms. This is evident even in this small sample of 15 lines, where one can detect two roots that appear in two different forms: cysts-blastocysts and (plant)-implant-transplant. When students deal with a larger variety of lines, they are bound to discover that certain roots are repeated regularly.

The teacher’s role is important in this task; s/he must be prepared to answer some Latin-related queries, since the majority of terms in medical English derive from Latin and Greek. If the ESP environment is in Greece, then learners will come across far fewer vocabulary problems, as they are expected to be familiar with several terms that derive from Modern Greek. The importance of derivatives and etymology tasks is considered more important in environments where students are not familiar with Greek or Latin.

3.6 Grammar

Through the use of appropriate activities, a corpus can prove a useful tool not only for vocabulary building but also for grammar. Since a corpus is a collection of samples of running texts, Aarts (1991: 47) claims that “ideally, the intuition-based grammar, through its confrontation with corpus data, becomes an observation-based grammar, i.e. one that also accounts for the facts of language in use”. Shei (1999: 1) also indicates a tendency shift, from the more “structural” character of grammar – as a set of rules according to Chomsky – towards the “performance grammar”, viewed as a flexible set of tendencies formed by real life data.
3.7 Learners' direct access to a corpus

When the instructor decides to allow students to explore the concordance output by themselves, s/he must expect problems that may arise due to students' lack of familiarity with the software. A possible scenario is that the raw output confuses students, probably because they have entered too general a query. Stevens notes "formulation of productive queries is particularly difficult for language learners, who may need assistance until they have become familiar with the technique. Misspellings which spoil productive searches are common". Also, "the relationship between raw data and output is not obvious to all, and the very existence of the text base, its particular bias, and its relevance to the students must all be explained and emphasized" (Stevens 1995: 2-10).

After the necessary introduction to the benefits of corpus linguistics, the instructor should assign to learners the task of doing their own corpus research, having a specific goal. This specific or general vocabulary task can be based on authentic texts, such as medical journals – either online or printed. Learners

![Figure 3. List of collocates for 'ovary']. Users can see the words appearing most frequently in contexts of 'ovary'

---

5 Note that medical terms are unlikely to produce as many results as do items of general everyday language. Even the phrasal verb put off has an output of 373 instances in the Bank of English, not to mention verbs like get which have thousands of instances.
could be asked to work in groups, a paragraph assigned to each, and asked to arrive at conclusions after research on concordances. The tasks can be of either special terminology or grammatical nature. When asked to explore the concordances of the term *ovary*, for instance, apart from the raw output of concordances, they can explore the collocates. Collocates in the Bank of English can be displayed both in the form of a list and as a ‘picture’, where users view collocates in boxes, before or after the node – query.

By using interesting keywords for search, teachers introduce learners smoothly to the notion of polysemy. After having first-hand experience, they will be able to deal with complex polysemous words and phrases at a more advanced stage of their English courses. Alternatively, as a form of testing and auto-correction, learners can be asked to write down their own examples and then compare them with the corpus output. Actually this task can be performed with any part of speech – both grammatical and lexical items. In this way, learners arrive at conclusions about language themselves, by processing data and thus they start taking initiatives in their learning. It seems, then, that corpora-aided teaching is learner-centred and develops learners’ observation skills as well as their creativity. That’s why Stevens wrote: “DDL is distinct from other inductive models of learning in that the teacher facilitates student research into the language without knowing in advance what rules or patterns the learners will discover” (Stevens 1995: 2-10).

4. Evaluation - advantages

Summing up, the use of corpora in language learning yields many advantages which Stevens summarizes as such: “first, it interjects authenticity (of text, purpose, and activity) into the learning process. Second, learners assume control of that process. And third, the predominant metaphor for learning becomes the research metaphor, as embodied in the concept of data-driven learning (DDL)” (Stevens 1995: 2-10). In short, language teaching follows basic patterns and principles of the communicative approach, where learners and their needs are at the centre of the learning process, which is structured according to their preferences. This is the point where ESP coincides with corpora-aided EFL, because, as Robinson has stated (1980: 10), ESP courses should be learner-centred.

Learners should appreciate the fact that examples are authentic, collected from a variety of texts, and not made-up for a particular coursebook. Take, for instance, the introduction of the phrasal verbs in the learning process. The average course book presents a set of phrasal verbs and then asks students to
practice them in the usual gap-filling, multiple-choice or matching exercises. A corpus, on the other hand, can provide a whole list of examples in use, illustrating automatically all the nuances of meaning, since most phrasal verbs are polysemous. After such exposure to the concordance lines, learners should be able to understand all the subtle differences in the meanings and forms of the verb.

Overall, corpus-enhanced teaching will not only enhance the prestige of the activity in the minds of even the most doubtful students, but it will bridge the well-known gap between classroom environment and the real world, which can severely influence future professional activity of ESP learners. They will feel much more confident if they know that what they are learning is derived from real-life situations. Besides, they are likely to take an original interest in the activities originated from corpora since they realise that these involve authentic uses of language. The boost of confidence and enhanced interest might also change learners' attitudes to learning.

5. Conclusion

Dictionaries have always been a valuable tool to foreign language learners. Nowadays, corpora aided by concordancing software prove so useful tools to foreign language learning, that they could surpass them in popularity. Instructors benefit from corpora by using authentic material, in order to expose learners to authentic real-life contexts. According to Stevens (1995: 2-10), corpus linguistics is based on a fundamentally social phenomenon: language; and is experienced through the empirical data of the software.

This paper offers the EFL instructor a new perspective in technology-assisted instruction. ESP teachers, in particular, often look for inspiration and authenticity in an area that is in need of communicative tasks and authentic materials. However, apart from ESP teachers, many EFL instructors are likely to find the use of corpora, as suggested above, extremely appealing, since corpora are invariably among the best sources for context-based teaching. The paper concludes by impelling language instructors to venture in the area of corpora, since corpora-assisted language learning is a step forward to integrating technology in language learning.

References


http://www.dai.ed.ac.uk/homes/shei/collocation/node6.html


http://tractor.bham.ac.uk/ijcl/past.html