CORPUSNIYT PODХOD PRI EЗIKOVITE IZSHEDAVANIA I
ПРИЛОЖЕНИЕТо В МУ В ПРЕПОДАВАНЕТО НА ТЕХНИЧЕСКИ
АНГЛИЙСКИ ЕЗИК

Билияна Овчарова, Лесотехнически университет – София
София Папаефтимиоу-Литра, Василис Хартзолакис - Атински университет, Гърция

In the area of teaching English for Specific Purpose (ESP) the domain-specific terms are of primary interest. These linguistic items used for the purposes of scientific accuracy and disambiguation pose the following challenges to the ESP teacher:

A. What is the correct translation of the terms?

In many narrow scientific fields there are no adequate bilingual dictionaries, and the foreign language teachers are left “to guess” what the correct translation might be. This paper will provide an example of terminology extraction which might be very useful for the purposes of teaching/translating of scientific texts (See Section 2).

B. Teaching the terms alone is not sufficient for effective communication. The paper will demonstrate how the “environment” in which the terms occur can be helpful to identify collocation patterns (See Section 5).

In this paper the two challenges above will be tackled through the methods of corpus-based study of language; therefore, it is important to establish first some of the basic definitions and notions related to corpus linguistics.

1. DEFINITION OF A CORPUS AND TYPE OF CORPORA.

The definition of a corpus has evolved over the years as the methods and tools for linguistic analysis changed. Nowadays many linguists admit that the factor that has influenced the study of language more than any other is the computer revolution which has made it possible to process large amounts of linguistic data. According to Hunston 2002: 2 [a] corpus [is] a collection of naturally occurring examples of language [...], a set of written texts or tape recordings, which have been collected for linguistic study [and] are stored and accessed electronically. Each corpus is built for a specific purpose and has a specific form. In this sense it is a planned selection of texts.

The notion of planning is not new to the foreign language teacher: teachers plan syllabuses, teachers design lesson plans and find learning materials to fit those syllabuses and lesson plans. Over time any ESP teacher accumulates specialised texts, some of which they particularly like and use “daily”; others they use from time to time for a special purpose; still others they discard as “useless” (often the reason being that they are too technical and difficult to comprehend). All these texts can easily be converted into a specialised corpus for language study. Two things have to be done:

- the first is very simple: all the texts should be made available in an electronic format (which means to use a scanner, if necessary);
- the second involves an important issue of corpus design – the issue of corpus representativeness. The ESP teacher “will have to define what it is that [the] corpus is to represent” (Halliday, M. A. K. et. al 2004: 116).

Corpus representativeness is a complex issue and linguists approach it in different ways; but it is not the purpose of this paper to discuss these approaches. Representativeness may involve a number of criteria: selection of texts, size of samples, inclusion of written/oral texts to name but a few. Although these are significant in designing a corpus, it seems
that for the ESP teacher there is one other understanding of representativeness that, to the greatest degree, concerns language studies: representativeness delimits the scope of investigation (ibid). To translate this in practical terms, it means that the ESP teacher will find it meaningful to focus on analysing the lexical items identified as terms and the collocations they form, and, having in mind these aims of analysis, will evaluate the representativeness of the corpora they use. The example in Section 5 is an illustration of terminology-focused analysis; we also believe that it may serve as a basis for developing an analytical corpus-based assignment for students to perform.

Having now established what a corpus is, it will be beneficial to broaden the perspective by looking at the types of corpora that exist. One type of corpora have already been mentioned: these are (1) specialised or domain-specific corpora which focus on studying a particular kind of language restricted to a particular register, time frame, social setting, topic, etc. There are also (2) general or reference corpora. These are multi-purpose corpora and contain a broad range of text types from as many genres as possible. The British National Corpus (BNC) has 90 million words of written and 10 million words of spoken data; the Bank of English Corpus contains over 400 million words and increases in size progressively. (3) Translation or parallel corpora consist of texts in one language and their translations into one or several other languages, and (4) comparable corpora consist of texts in different languages but as far as possible similar in text types and size. More details about types of corpora can be found in Altenberg and Grange 2002 and Hunston 2002.

As this paper is intended to illustrate what the implications of using corpus-based techniques for applied linguistics are, an example of terminology extraction is presented below.

2. USING PARALLEL SPECIALISED CORPORA TO MAKE A GLOSSARY OF TERMS: AN EXAMPLE.

Christine Chodkiewicz, Didier Bourigault and John Humbley 2002: 249-267 used parallel English-French corpora to extract terminology. The two corpora contained texts of legal documents of the European Commission: the complete texts of the European Convention on Human Rights, plus its 11 protocols and 36 court decisions. The texts were first aligned using the structured framework of the documents (number of decision, section of decision, number of section and etc.). After that the respective parts of speech were identified with markers i.e. the corpora received part-of-speech tagging. The term extractor, LEXTER, was used which performed a morphosyntactic analysis of the French text and produced a list of noun phrases which are likely to be terminological units (candidate terms). These were evaluated by a terminologist, and then using the parallel English corpora were translated into English. Thus a glossary of terms was obtained through computer-assisted term extraction.

This example is important to show how corpora were used to compile a terminological list for a restricted domain. Though expert knowledge was required to analyse the multiple correspondences, the machine-extraction provided a glossary that can be used by professional translators and legal professionals whose native language is not English or French. This was only one instance how corpus-based techniques can help the ESP teacher; however, another type of analysis is in the centre of the computer-assisted language research, and it will be presented and discussed in the forthcoming sections.

3. THE FOCUS OF THE CORPUS-BASED APPROACH TO THE STUDY OF LANGUAGE.

As suggested by Hunston 2002: 3 it is through corpora that the researcher can obtain a new perspective on the familiar by studying the frequency, phraseology and collocation of the linguistic items. Similarly, Biber et. al 1998: 6 state that the corpora can provide many additional kinds of information about language use; they call these patterns of use linguistic associations:

Linguistic associations fall into two major categories:

1. lexical associations – investigating how the linguistic feature is systematically associated with particular words. For example we say strong tea and powerful car, and not *powerful tea or *strong car;

2. grammatical associations – investigating how the linguistic feature is systematically associated with grammatical features in the immediate context. An example are adjectives denoting (un)appropriateness or (im)possibility which are used in impersonal constructions as in It is advisable to be on time.

Biber et. al 1998: 6

Consequently, it is not surprising that corpora provide valuable collocational data that has long been used in compiling dictionaries (for a detailed description see Meyer 2002: 14). As to grammar books, an outcome of the corpus-based approach is the Longman Grammar of Spoken and Written English in which Biber et. al 1999 which presents the grammatical features in terms of a comparison of frequency between four broadly defined registers: fiction, academic prose, newspaper reportage and conversation. (If you are interested to get a glimpse of it, a sample page can easily be downloaded from: http://www.longman.com/dictionaries/which_dict/ lgswe.html.) To analyse data special software
programmes are used which, fortunately, are not very
difficult to download from the Internet; more details
about their capabilities are discussed in the next sec-
tion.

4. TOOLS FOR CORPUS ANALYSIS.

The basic tool for corpus analysis is a software
program called concordancer. Such a programme is

WordSmith (http://www.lexically.net/wordsmith/) and below some of its features will be briefly demonstr-
ated.

Frequency

Very easily a concordancer package can generate frequency lists. Below you can see how the list

of most frequent words and the alphabetical list of

words in a text look like:

Key word in context (KWIC)

This is the most common and useful feature of

a concordancer programme. It presents the selected-
for-search word in the middle with the words that
come before and after it to the left and right. It is

through the concordance lines that meaning distinc-
tions and patterns of usage can be observed. Below
is the word “tree” as a node word in its context:

Naturally, the software programme cannot do
the analysis for us, but it is invaluable in providing
systematised data for it. An example of this will be
presented below by showing how a specialized word
in the field of forestry can be allocated through cor-
pus research.

5. THE EXAMPLE OF STOCKING.

Preliminary note

The analysis below will try to investigate the
meaning of the term *stocking* in forestry texts. Other senses of the word will not be the aim of this study. The ESP teachers might regard the steps described below as a procedure for analysis of terminology that learners can be assigned to perform themselves. To start off, let us first look at some dictionary definitions that will give a direction to our analysis.

**Dictionary definitions**

It seems appropriate to look at the senses of the lexeme STOCK in both its noun and verb form as the noun *stocking* in non-terminological dictionaries is defined only as “a covering for the foot and leg” (The American Heritage Dictionary, 2000).

The noun *stock* means among other things (ibid.):

1. A supply accumulated for future use; a store.
2. The total merchandise kept on hand by a merchant, commercial establishment, warehouse, or manufacturer. 3. All the animals kept or raised on a farm; livestock.

The verb to stock has two senses (‘b’ and ‘c’) which might give us a clue with reference to living things though not to trees (ibid.):

1. To provide or furnish with a stock of something, especially: a. To supply (a shop) with merchandise. b. To supply (a farm) with livestock. c. To fill (a stream, for example) with fish.

As the above dictionary definitions do not provide much information as to the meaning of *stocking* as a forestry term let us also look it up in Forestry Glossaries:

A. Glossary of Forestry Terms (online): *stocking*: a measure of the area occupied by trees, usually measured in terms of well-spaced trees per hectare, or basal area per hectare, relative to an optimum or desired level.

B. Glossary for Agroforestry (online) *stocking*: an indication of the number of trees in a stand, compared with the desirable number for best growth and management, such as ‘well stocked’, ‘overstocked’, ‘partially stocked’ or ‘understocked’. Applies also to domestic animals.

**Corpus design and collocations obtained**

Having looked at the dictionary definitions, the analysis will proceed with studying the term *stocking* in its actual context of use. The analysis of the term will be based on online resources, but still we believe that it can be useful for the purposes of language learning and training. Considering the fact that general dictionaries do not provide any information about the meaning and use of terms, as the definitions presented above prove, there is a need for making use of the available resources with the awareness of the diversity of the sources. In an ideal study the corpus for investigation would contain texts of established authors and publishers in the respective technical field.

The corpus for this particular study was built using the search engine Google to generate results for the query *forest stocking* which came up with 1,990,000 search results.

Of all the texts containing the term under investigation, upon common sense judgements about the text source and appropriateness, “targeted” texts were selected and processed with the help of Wordsmith Concordance. The programme generated seventy instances of the term *stocking* in its immediate context. After analysing qualitatively these instances, the last step was to identify the patterns of use: the desired outcome of the study. The identified patterns of use are presented below:

6. **CONCLUSIONS.**

To summarise, the following conclusions can be drawn about the advantages of the corpus-based approach to the study of language:

- it analyses the actual patterns of use in natural texts by using a collection of texts known as a corpus;
- it makes extensive use of computers for analysis employing automatic techniques;
- it depends on both quantitative and qualitative analytical techniques;
- it can be the basis for preparing assignments for learners that enhance awareness of language usage patterns.

**References**

CORPUS-BASED APPROACH TO THE STUDY OF LANGUAGE AND ITS APPLICATION IN TEACHING ENGLISH FOR SPECIFIC PURPOSES

Bilyana Ovcharova, University of Forestry, Sofia, Bulgaria
Sophia Papaefthymiou-Lytra, Vassilis Hartzoulakis - University of Athens, Athens, Greece

ABSTRACT

Teaching English for Specific Purposes (ESP) in tertiary education focuses on terminology acquisition as target competence. At the University of Forestry in Sofia, Bulgaria translation of specialised texts in the respective field of study (forestry, wood industry, landscape architecture, agriculture, etc.) is the major type of learning activity. Due to this practice, difficulties, for both teacher and learner, usually arise because of the lack or incompleteness of bilingual terminology dictionaries for the specific field.

This paper will present how through the methods of the corpus-based approach to the study of language the ESP teacher or/and the learner can analyse specific terminology in its context of use. Prior to that, some basic definitions and issues of corpus research will be presented and discussed.

Tell me and I’ll forget. Show me and I’ll remember. Involve me and I’ll understand

Confucius