Teaching terms: a corpus-based approach to terminology in ESP classes

Maria João Cotter
Lisbon School of Accountancy and Administration (ISCAL) (Portugal)

Abstract

This paper will build up on corpus linguistic theories and methodologies to analyze terminological data, which will later be applied to the actual context of Business English classes. Our aim is to show that simple corpus mining techniques impact greatly on students’ performance, since they will engage in real-life tasks, based on authentic material. The first part of the paper concentrates on theoretical issues and starts with the assumption that teaching materials very often do not reflect authentic language use.

In the second part of this paper, we will deal with practical applications to the didactics of ESP. With a tailored corpus, the teacher will be able to analyze texts linguistically in order to extract the lexical units that will be more relevant to the current or future professional activity of the students.

Key words: collocation, corpus linguistics, ESP, terminology.

Introduction

Given the specificity of ESP, many will argue that readily available materials in the form of textbooks are usually too broad and not easily adaptable to real learner’s needs (Harwood, 2005). Building on this assumption and considering that the lexis plays an essential role in ESP courses, we will focus on terminological issues and endeavour to show that creating customised corpora is an easy task that will prove beneficial for both teachers and learners.

In the second part of this paper, we use the theoretical groundings explored and apply them to a Business English course which is roughly based on the Technical English course 3rd year students at Lisbon School of Accountancy and Administration are required to take. The course is structured around the following topics: Accounting, Banking, Taxation, Auditing, Economics and Ethics, which will be the basis for our terminological approach.

Corpus linguistics and ESP

Corpus linguistics uses real data to assess hypothesis about language allowing us to explore the relevant linguistic structures for the aims we set for our research. «A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria» (Bowker & Pearson, 2002: 9).

McEnery and Wilson (1996) argue that foreign language teachers usually produce simplified examples, which will raise difficulties for students when these are confronted with real, more complex language that sometimes they are incapable of processing. Corpus linguistics can thus contribute to rendering learning a foreign language more effective since students will be faced with «real language».
Sinclair’s position may sound too radical but given an ESP context the difficulty of producing examples that «seem real» is even greater. Since ESP teachers are rarely experts in the learners’ field of knowledge and sometimes they are not native speakers of the language, they will be faced with two types of problems: producing examples in a language that is foreign to them and in a field of knowledge that they do not master. However, the ESP teacher plays a fundamental role in teaching students to recognize technical vocabulary and understanding definitions (Chung & Nation, 2003).

If in terms of grammatical structures teachers with a good mastery of the language, and in spite of the aforementioned difficulties, are able to formulate statements that exemplify a given construction relying only on their intuition, lexically speaking, and in the case of ESP in particular, the situation is more complex since lexis is one of the most specific areas in ESP.

Both learners and teachers will then have to rely on the materials used for the ESP course as the main source of new vocabulary. When using textbooks, vocabulary choices were already done by the authors and do not take into account the specific needs of each group of learners (Bogaards, 1994). By producing additional materials based on authentic texts, however, teachers will greatly improve their learners’ proficiency in ESP. In order to produce these materials, teachers can resort to custom-made ESP corpora.

**Building an ESP corpus**

Several factors are significant when building a corpus. Sinclair (1991) defines some criteria for the design of general language corpora, recommending the inclusion of all LSP material in a separate corpus. But to start with one has to define the purpose of the corpus. In the second part of this paper, we focus our attention on the creation of an ESP corpus that can be used both in class and to prepare classes.

Pearson (1998) talks about several relevant factors that need to be taken into account when building special purpose corpora, such as size, text type and origin, authorship, factuality, technicality, audience, intended outcome, setting and topic. In an ESP corpus the selection of the topics depends on the learners’ field of knowledge. The criteria of authorship and audience are very important, since these will determine the technicality degree of the texts. Not all these aspects are equally significant, and ESP teachers have to make learner-oriented decisions since «[t]he optimal design of a corpus is highly dependent on the purpose for which it is intended to be used» (Kennedy, 1998: 70).
Collocations

It’s not enough to learn lexical units separately, learners also have to know how to combine those units. Therefore, ESP teachers have to draw learners’ attention to «chunks of language», i.e., collocations.

Collocation is the occurrence of two or more words within a short space of each other (…) Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated. (Sinclair, 1991: 170)

The easiest way to identify collocations is by checking the frequency of a given group of words in a corpus. Obviously a thorough analysis of the data should be done because that group may occur more often simply because it is composed of very common words (Biber et al., 1998).

A large part of our mental lexicon consists of combinations of words that customarily co-occur. The occurrence of one of the words in such a combination can be said to predict the occurrence of the other(s). (Kjellmer, 1991: 112)

Native speakers have a natural tendency to combine words in pre-defined chunks that are easily recognizable by their listeners. Kjellmer (1991) classifies those groups of words in relation to their internal degree of cohesion in a continuum from what is totally unknown to fixed expressions.

Conceptually restrained, technical vocabulary is formed by collocations that introduce specialised knowledge in ESP. The identification of this special language is made by inferring idiomatic constructions from concordance samples. The aim is to perceive the fixation of long compounds, and to appreciate the value of this lexical restriction in the subjects. (Curado Fuentes, 2001: 118)

By identifying and studying collocations learners will form their mental lexicon not only from isolated independent units, but also from pre-combined units, thus consolidating a conceptual system that will allow them to become more proficient at an initial stage of learning.

Retrieving terms from corpora

Distinguishing terms from common words is not an easy task. Pearson (1998) offers an interesting discussion on the subject and arrives at the conclusion that what constitutes a term depends on the communicative setting. However, we do know that terms label concepts and must be learnt by ESP students.

Compiling and exploring a thematic corpus

In order to demonstrate the effectiveness of a corpus-based approach to terminology in the ESP class, we compiled a set of articles on a common topic –the impact of the Sarbanes-Oxley Act– called BEC_audit. This corpus is composed of 32 texts collected from different online sources written over a year. All texts were stored in a .doc format, complete with date, URL address, graphs and illustrations. Subsequently, irrelevant items were removed (everything but the text itself) and a new .txt file was created, i.e. the corpus proper.
The BEC_audit corpus was explored with ConcApp, a freeware concordancer available at http://www.edict.com.hk/PUB/ConcApp/ (27.05.2006). Concordancers provide concordances, i.e., alphabetical lists of words in context, which can be searched according to various criteria.

We started by using the option Statistics > Unique Words to count the total number of words in the corpus (19,852), as well as the number of unique words (3,445).

**FIGURE 1.** The nine most frequent words in the BEC_audit corpus.

The most frequent units in a corpus are usually functional words, such as prepositions, articles and conjunctions. In Figure 1, we can see that the ninth most common word is *audit* with 178 instances and a usage rate that is higher than any form of the verb *to be*. A frequency this high confirms that the corpus was adequately compiled for the purpose of studying Auditing terminology.

**TABLE 1** The ten most common lexical units in the BEC_audit corpus.

<table>
<thead>
<tr>
<th>Position</th>
<th>Lexical Unit</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>9th</td>
<td>audit</td>
<td>178</td>
</tr>
<tr>
<td>12th</td>
<td>said</td>
<td>139</td>
</tr>
<tr>
<td>21st</td>
<td>auditors</td>
<td>98</td>
</tr>
<tr>
<td>22nd</td>
<td>companies</td>
<td>97</td>
</tr>
<tr>
<td>24th</td>
<td>accounting</td>
<td>91</td>
</tr>
<tr>
<td>26th</td>
<td>firms</td>
<td>83</td>
</tr>
<tr>
<td>29th</td>
<td>internal</td>
<td>76</td>
</tr>
<tr>
<td>32nd</td>
<td>Sarbanes-Oxley</td>
<td>70</td>
</tr>
<tr>
<td>36th</td>
<td>financial</td>
<td>63</td>
</tr>
<tr>
<td>38th</td>
<td>clients</td>
<td>61</td>
</tr>
</tbody>
</table>

Ignoring functional words, Table 1 shows the ten most common lexical units in the BEC_audit corpus. Except for the verb *said*, all other units are part of the terminology
of Auditing. Using the tool Concordance > Search, we analysed these units and
observed the contexts in which they occur in order to determine whether they are
actually terms in this LSP. This type of analysis will also make the observation of
colloocations easier and it is essential in the case of adjectives such as internal, for
example, which isolated is not a term, although intuitively we may think it makes part
of collocations that are relevant for the terminology we are dealing with.

**FIGURE 2. Associated word search.**

![Figure 2](image)

Figure 2 shows the first twenty instances of the term internal where we can see four
complex terminological units: internal accounting, internal audit, internal auditing and
internal auditors.

ConcApp also enables us to test our intuition regarding expressions that have a great
probability of co-occurring by using the option Associated Words. We can define the
item we intend to search, the other word that co-occurs with it, and the relation between
the two items in case it is relevant: the second word might occur on the right or on the
left of the first.

We tested the expression Big Four, since the lexical unit big occurs 47 times in the
corpus and the unit four occurs 46 times and these data might prove important given the
scope of the corpus. In effect, we saw that all 38 co-occurrences of both words are
capitalized, and are obviously a term (see Figure 3).

**FIGURE 3. Collocation Big Four in the BEC_audit corpus.**

![Figure 3](image)
The less frequent words in a corpus might also be interesting, particularly those that occur only once, called *hapax legomena*. In a small newspaper article with around 200 words, 150 may be *hapax legomena*; but in larger representative corpora the proportion of *hapaxes* is usually much lower (Kennedy, 1998). Sometimes these are simply mistakes in the original texts that need to be corrected, but they may also be relevant terms in the LSP we are analysing. In the BEC_audit corpus we found some interesting *hapaxes* that are clearly auditing terms: bribery, leveraging and stakeholders.

Using corpora to make exercises

In this section we used the corpus BEC_bank composed of 20 articles on the topic of banking downloaded from the Financial Times website. Our aim is to show that a body of readily-available sentences facilitates the ESP teacher’s task when the time comes to make both vocabulary and grammar exercises. Exercise 1 provides an example of sentences taken from the BEC_bank corpus that test the acquisition of Banking terms.

Exercise 1

| a. | They are swimming in __________ card and student loan debts. |
| b. | Halifax's new __________ account gives cash-back on debit card purchases, up to £100 a year. |
| c. | The big four __________ banks have little incentive to speed up the payments system because they make money from it. |

It is also possible to search for prefixes or suffixes using ConcApp and thus analyse certain grammar categories, such as adverbs of manner ending in –ly or verb constructions. Exercise 2 is aimed at testing the usage of Gerunds and Exercise 3 tests prepositions.

Exercise 2

| a. | Many lenders will charge you double the interest rate for __________ (use) a credit card. |
| b. | This type of deal would depend on __________ (get) a price from the broker’s agent in the country of origin. |
| c. | The group will start __________ (look) at whether to speed up cheque clearing in October. |

Exercise 3

| Disputes ______ cash rose ______ 347 ______ the year ending ______ March 31, ______ from 117 ______ the previous 12 months. |

One of the main advantages of using a corpus to collect material that can be used to make exercises is the fact that teachers will actually use terminology and syntax that are characteristic of the particular LSP they are teaching.
These special purpose corpora, however, are not only advantageous for teachers but also for learners. Several authors have shown that it is possible for learners to explore corpora individually or in group. Gavioli (1997) presents a series of activities that are aimed at providing the students with the necessary tools to be able to ask relevant questions and the capacity to interpret the answers given by the machine; Kennedy & Miceli (2001) describe a similar experiment.

Final Remarks

The usage of customized ESP corpora is extremely useful for the preparation of materials. The possibility of analysing specific terminology ensures that teachers are conveying information that will actually be valuable for learners in their professional life. The methodology we have described in this article is easy to follow and allows ESP teachers to adapt materials to the specific needs of learners (Aston, 1997).

The larger the corpus, the easier it will be to make generalizations and say that the selection of the terms was based on the usage criterion which Bogaards (1994) identifies as the most important one.

Much is still to be done in this area, such as ascertain whether learners will benefit more from direct or mediated access to corpora, defining strategies for both direct and mediated access, and how to integrate this experiences in the language learning environment (Chambers, 2005).

References


