English for Specific Purposes World
Online Journal for Teachers

ESP W RLD

ISSN 1682-3257
Indexed in
Genamics JournalSeek

There is more to knowing a language than knowing its words: Using parallel corpora in the bilingual classroom

Pernilla Danielsson & Michaela Mahlberg

1. Introduction

A 'corpus' is a large collection of naturally occurring texts and corpus linguistics is concerned with the compilation and investigation of such corpora. Corpus linguistics is a relatively new discipline which originates from the second half of the Twentieth Century when the first machine-readable corpora were compiled. Large text collections had existed before, such as bible concordances, and many of the early lexicographers used language samples for their work. One major difference was the introduction of computers, the evolution of this technology has enabled linguists to not only store huge amounts of text, but analyse these natural language samples in a way which had not before been possible. However, corpus linguistics is responsible for triggering far more innovation in linguistics than just methodological developments. Corpora provide new insights into the way language operates and calls for new approaches to linguistic theory. Furthermore, it should be stressed that corpus linguistic applications can prove useful in many other areas. In human interaction language plays a major role and the evidence provided by texts (stored in corpora) can be used to investigate many kinds of questions, such as those of a cultural, historical and political nature.

In the present article we will focus on the use of corpus linguistics for teaching English as a foreign language. It could be regarded as a characteristic feature of corpus linguistics in ELT (English Language Teaching) that it may provide a way of bridging the gap between theory and practice, as its relevance for the language classroom can be seen on various levels (cf. Hunston 2002, Mahlberg in preparation). On the one hand corpus linguistic approaches are of theoretical importance, as they aim to describe language and open new perspectives on grammar and vocabulary (cf. for instance, *The Longman Grammar of Spoken and Written English*, *The Collins Cobuild English Dictionary*). Thus, corpus linguistics can have an impact on the content of language classes, i.e. what is to be taught. On the other hand corpus linguistic methods can be used within teaching to both determine how the content is presented and how the students get involved into experiencing language.

In this article we will deal with theoretical and practical questions that concern corpus linguistics

and ELT. The article is laid out as follows: This section, section 1 (above), has given a general introduction to the subject, section 2 will discuss some general aspects of the information found in corpora and section 3 will focus on the ways that corpora may be used in teaching. Section 4 will introduce relevant software for searching parallel corpora, section 5 will detail a case study of the Swedish word *stor* and its corresponding items in English and section 6 will sum up the main points of the article.

2. The evidence from corpora

The new perspective to language, that is taken by corpus linguistics, draws on the evidence provided by natural language. This evidence can bring to light phenomena that has so far gone unnoticed, or that runs contrary to the picture created by more traditional theories. A simple rule which generations of pupils have been taught may serve as an example: traditional grammars state that *any*, in contrast to *some*, is used in questions, negative or conditional contexts. Corpus findings, however, show that in roughly 50 per cent of its occurrences *any* is found in positive contexts, see for example Tesch (1988: 62f.), Tognini-Bonelli (2001: 17). Though there are different approaches within corpus linguistics and corpus linguistic theories can follow different paths it may be seen as one of the major assets of this disciplines that it stresses the context dependency of meaning. In particular, it has been shown that the traditional separation between lexis and grammar cannot be upheld. It is an oversimplification to just draw a dividing line between the words of a language and the rules according to which these words are combined. Language cannot simply be described in terms of a slot-and-filler model, where text is created by the interplay of grammatical rules and lexical choices, enabling a series of slots to be filled from a lexicon (cf. Sinclair 1991: 109ff.). Linguistic choices are often characterised by 'co-selection', i.e. certain combinations of words are selected as groups. This feature of language is considered by corpus linguistics in the description of phrases and units of meaning larger than a single word form.

Turning to language teaching, we find the idea of phrases is no new innovation (cf. for example, Nattinger & DeCarrico 1992, Cowie 1988) but in the classroom phrases are often treated as special cases. For instance, when phrases are presented as short-cuts for the language learner in specific communicative situations. For example, in conversation we find phrases such as *Would you like some..?, You?re welcome, I beg your pardon,...* whereas in essay writing the following phrases may be useful *on the one hand...on the other hand, in contrast to this, in conclusion.* The marginal status of phrases gains particular emphasis when textbooks talk of 'fixed phrases' or 'idioms' when the meaning of individual words and purely grammatical rules fail to explain combinations, such as *a piece of cake, put up with, go off,* or *drives me nuts*. From a corpus linguistic point of view it is made clear that phraseology "encompasses all aspects of preferred sequencing as well as the occurrence of so-called 'fixed' phrases" (Hunston 2002: 138). Thus, phraseology does not only account for expressions like *a piece of cake*, it also plays a part in explaining why we say *afraid of* and not *afraid in*, and why an accident *happens* while a meeting *takes place* and problems *occur*. However, for the human observer this pervasiveness of co-selection in language is hard to notice.

One of the computer tools that are helpful to discover the behaviour of words is a 'concordancer'. A 'concordancer' finds all the occurrences of a word in a corpus and presents them in the form of a list which is called a 'concordance'; Concordancing is one of the oldest ways of browsing through a corpus. Early in text computing the KWIC (KeyWord In Context) model was established and it is still the standard way of presenting concordance information. The 'keyword' is the word under investigation and the 'context' is provided by the words to its left and to its right. The concordancer allows us to specify the extent of the context, e.g. five words either side and we can select the number of examples to be listed. For example, if a word occurs 10,000 times in a corpus a more digestible number of examples may be selected. Furthermore, concordance software has various options for displaying and sorting the selected examples. Thus it is possible to sort a concordance according to the first, second, etc. word to the left or right of the keyword. Below is a subset of a concordance for the word *afraid* sorted by the first word to the right.

s flagship Odeon cinema said: `I am afraid Ally's been subbed." <p> Rangers

notions in ten years. I'm afraid I have to tell you that I believe

company operations you own; I'm afraid I've got to get some detail." <p>

ou know me?" he said grumpily. `I'm afraid I don't know you. I've shaved

of that band of loonies? I'm afraid it's just no contest at all. <hl>

hellip; we cocoon because we are afraid of saying the wrong thing or not

a Pullman, and partly because he was afraid of being seen there by some

unspeakable practices; Gregory was afraid of him because once, under a

attacks. They think she is afraid of upsetting the politicians linked

Into the lens, but equally do not be afraid of photographing against the light.

The British forces at Belsen: `I am afraid that you may think I am

because health professionals were afraid to cross lines marking religious

willowy decadence, blase, stupid, afraid to be seen sober or with a book,

t say yea <p> or nay, and he was afraid to ask. <p> When the new field lay

with their eyes what they were afraid to put into words. Is it possible?

Monica) Lewinsky report, you are afraid to put a cigar in your mouth. The

Table 1: Concordance for *afraid*.

This concordance reveals a number of phrases and uses for the word *afraid*. In the above concordance we find, for instance, evidence for the combination *afraid of*. However, it is not only structural information that is revealed by the concordance. Other types of information may be exemplified by the use of *I'm afraid* where an important feature of this phrase is that it is used in spoken English to introduce information that might be unwelcome or unpleasant.

With the help of the computer we can take a more objective view towards language as our observations are not restricted to what our intuitions allow us to notice. Moreover, the computer opens the possibility of providing quantitative information that can reveal what is frequently and typically used in language. Such information has had great impact on new approaches in lexicography that characterise dictionaries based on corpora (e.g. *Collins Cobuild English Dictionary for Advanced Learners* 2001 and *Macmillan English Dictionary for Advanced Learners* 2002). These reference works pay special attention to frequent words and typical uses that can be useful to learners. Similarly, when corpora are used in the language classroom, it is important to find ways of directing the students' attention towards what is useful for them and what will help them find their way through the vast information that corpora contains. Although the computer can process enormous amounts of natural language data the human observer still has the important task of interpreting the data.

In the course of this article we will discuss some of the relevant factors, one of the points to consider is that information from corpora is never fully objective, but depends upon the texts that make up the corpus. Generalisation about language has to be seen in relation to the data upon which they are based. The context dependency of words is related to aspects of variation, such as

register and genre, which need to be considered in the compilation of corpora and the interpretation of corpus data. The question of the appropriateness and representativeness of corpora is a difficult one and within the scope of this article we will have insufficient space to go into great detail. For a discussion of this problem see, for example, Atkins et al. (1992), Biber (1993). In section 3 we will only briefly hint at some problems of particular relevance for ELT.

Finally, it has to be stressed that the questions addressed by a corpus analysis are closely related to the methodology that is used. In this article we focus on the investigation of contextual relations of words by using a concordancer, but there are various other possibilities for analysing corpora, see for instance Biber et al. (1998), or Scott's webpage (http://www.lexically.net/wordsmith/) for information on corpus software.

3. Corpora and language teaching

Corpus linguistics can enter the classroom in various ways. Teachers, who do not have the necessary technical equipment to work with corpora, may still use reference works or teaching materials that are based on evidence from corpora. Corpus-based dictionaries lend themselves readily to the familiarisation of students with ideas of phraseology, however, there are also other reference works (e.g. Sinclair 1990, Mindt 2000, Ungerer 2000) that can bring in various types of corpus linguistic approaches. When teachers have the opportunity to use corpora in the classroom they have a variety of applications to chose from. Corpora may merely be sources for example and illustration, but they may also allow students the possibility for individual activity, for an overview see Mahlberg (in preparation) .

The focus of the present article will be on the use of corpora for 'discovery learning'. Discovery learning is not a particular feature of corpus linguistics but plays a part in teaching in general. In language teaching corpora offers new possibilties for discovery learning when students are exposed to natural language data. New resources ensure that they do not simply discover textbook rules (that might not even represent a true picture of the language) but the characteristics of real language. In this way, corpus linguistics can contribute to bridging the gap between theory and practice: corpus linguistic theory builds on evidence from natural language and corpus linguistic methods enable students to experience natural language. In corpus linguistics discovery learning is typically connected with the concept of ‘data-driven learning (DDL)’, which was developed by Tim Johns for teaching international students at the University of Birmingham (cf. e.g. Johns 1991). The characteristic feature of DDL is that learners take an active part in discovering the foreign language by acting as ‘language detectives’ (Johns 1997: 101).

When students make their own observations about language it is desirable to have a corpus that is as relevant as possible to the age group in question. A corpus compiled for young second language learners may include Harry Potter novels and other books aimed at this age group. Thus, pupils may recognise parts of the texts and the contents, which can help to increase their motivation and raise their interest in the work. Literature for children may also ensure that the vocabulary is not too difficult for the pupils to understand, although there are other factors to be taken into account.

At present there are no freely available corpora aimed at children or young language learners, therefore the corpus we are using for our case study is not directly aimed at young learners. Still, as we will show in section 5, the language examples speak for themselves and have a lot to offer the students. An important point, however, is the role of the teacher. It is the task of the teacher to make the potential of natural language examples accessible to the students. The teacher needs to be aware of the impact that the design of a corpus has on the results of a corpus study and has to introduce and explain the methods of analysis. The huge amount of information in a corpus can have the adverse effect of distracting a students' attention from the problem under investigation. Lessons with corpora need careful planning and the teacher will require thorough background

knowledge[1] so that he or she will know how to make any necessary simplifications for students and teach them how to read concordance lines. While a text is normally read horizontally, it is one of the features of a concordance that it needs to be read vertically (cf. Tognini-Bonelli 2001: 3). Thus, the fact that concordance lines are generally not displayed as full sentences is not purely a technical matter. Full sentences might encourage students to focus too much on each line and overlook the patterns that only become evident in relation to the other concordance lines. Sentence segments focus the view on the more general patterns and the possibilities of sorting on the surrounding words. On the whole, the generalisations drawn from corpus examples and the simplifications that become necessary in this process vary according to the level of the students. The teacher has an important task in setting the scene for the activity of the students, however, at the same time his or her role becomes more complex. Once the students have entered the process of discovering facts about language the teacher's role shifts towards being that of a mentor on a self-exploratory journey, rather than a finger-wagging tutor.

So far we have focused on 'monolingual' corpora, i.e. corpora containing only texts of just one language. However, the corpus used in our case study is a 'parallel' corpus, i.e. it consists of original and translated texts. Our parallel corpus contains 4 Swedish novels and their English equivalents, all in all 500,000 words per language. The corpus was not compiled with teaching primarily in mind, originally, it was used to extract translation units from authentic data (Danielsson 2001). Unfortunately, the rather small size of the corpus will bear visible consequences on the results and a larger corpus would have been preferable. At present however, no other parallel corpora consisting of novels is available to us, making this corpus the most appropriate. The development of parallel corpora lags behind the progress made for monolingual corpora, since the compilation of parallel texts is more complicated. Still, there are currently several projects developing parallel corpora and the situation appears set for improvement in the near future.

Using parallel corpora for second language teaching is still in its infancy. Experiences from using parallel corpora in higher education have been reported by, among others, Peters et al (2000), demonstrating how to use their Italian-English corpus for foreign language scholars and Danielsson & Ridings (2000) in a study of the use of multilingual corpora for translator's training programme. However, parallel concordancing also offers attractive possibilities for application in secondary schools. The language phenomena of interest may vary depending upon the level of the learner, but the type of exercises may remain the same.

The reasons for using parallel corpora may vary: parallel corpora can be used to support a contrastive approach to language teaching, they can also be used to deal with problems that occur in students' writing that arise due to the influence of their native language. Parallel texts can also serve as a form of internal differentiation within a group of learners, where some students may better understand certain phenomena when they work with comparison. Such considerations fall within the scope of the overall organisation of a teaching plan for a particular class, in the following we will focus on methodological aspects of parallel concondancing and a concrete language example.

4. Concordance tools for parallel texts

The case study is performed using a software entitled 'ParaConc' (Michael Barlow 1999, the software is available from the internet on the website www.athelstan.com). If you are familiar with using 'Microsoft Word' (MSWord), you will find this software easy to use. Just as with MSWord, you have menus up at the top of the window and simply clicking "ok" can close all pop-up windows. ParaConc accepts any text file as long as it is saved as 'plain text'. Before using the concordancer, the original texts have to be aligned with the translated versions. This means each corresponding sentence is linked from the original text to the source text. If you use an existing parallel text or corpus, rather than compiling a corpus yourself, the alignment information should already be included.

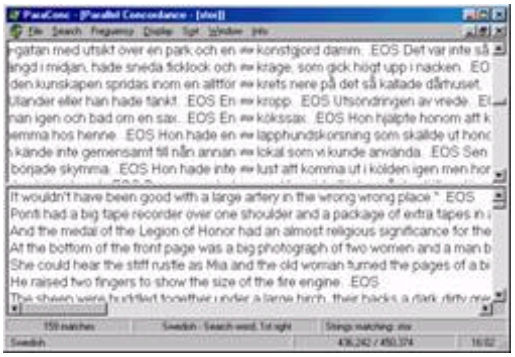

Figure 1: The interface of ParaConc

The example of *afraid* in section 2 illustrates a monolingual KWIC concordance. While monolingual KWIC concordance lines can give a good picture of how to use a word in English, they purvey only half the story to a second language learner. Making concordance lines for parallel texts (in two or more languages) will mean not only displaying the lines but also finding a way to show the link between the source texts and their equivalent translations. In ParaConc each language is displayed in a separate window where the line number refers to the corresponding sentence or sentences (see figure 1 above). A simple click on a sentence in one window will highlight its corresponding sentence in the other window.

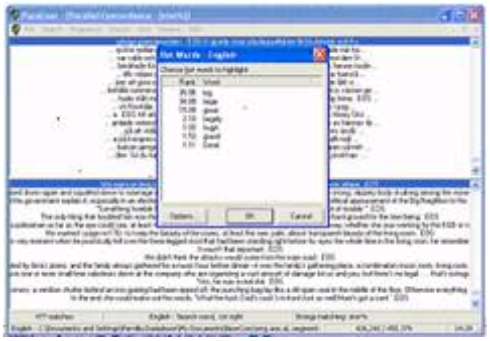

Figure 2: Displaying 'HotWords' in ParaConc

In addition to retrieving all lines that include the keyword in question, ParaConc provides a very valuable feature that calculates significant words in the target language using a statistical method. That means you may obtain a quick overview of the words that are likely candidates for translation equivalents of the keyword. This feature is referred to as 'HotWords' and the significant words are displayed in a smaller pop-up window (see figure 2 above). The 'HotWord function' retrieves words that occur more frequently in the translated concordance lines than expected. Not all 'HotWords' occur as translations of the keyword, very often the 'HotWord' list includes at least some of the equivalents. On the other hand, in some cases all the words in the 'HotWord' list can be used as translation equivalents of the search word. That is the case in our example of *stor* below. We will be using the 'HotWord' function to guide us to the relevant corresponding items in our study below.

Our main interest will be to find 'patterns' around a keyword in question. 'Patterns' are understood here as recurrent combinations of words, or word groups illustrating a common way of using the word. For example, in the concordance, table 1, for the word *afraid* we found patterns such as *afraid of* and *I'm afraid*. By studying concordance lines it becomes apperant that language regularly arranges itself into larger units of structure and meaning. This is an observation that may be made by anyone working with real language data. When a pattern forms a meaningful unit, such as *keep an eye on someone*, we will refer to it as being a 'multi-word unit' or 'unit of meaning'.

5. The case study: *stor, stort* and *stora*

This case study will illustrate how young second language learners may use the corpus in the classroom; we will investigate how to translate the Swedish word *stor* into English. The intention is to illustrate a method for young pupils to explore a second language in a teacher-guided environment, such as the classroom. It should be seen as an awareness-raising activity where the pupils are guided towards identifying patterns in language as well as similarities and differences between their native and second languages. Thus, they can learn how to get access to more information than is normally available from traditional language resources, such as bilingual dictionaries. If you search for *stor* in a bilingual English-Swedish dictionary you will find the following information:

stor *adj* ***1*** large; i betydelse rymlig, big; lang tall; abstrakt great ***2*** vuxen grown up

("**1** *large*; meaning spacious, *big*; lengthwise *tall*, abstract *great* **2** adult grown up")

Example 1: The entry for *stor* in Norstedts Lilla Engelska Ordbok ("Norstedt's compact two-way

English-Swedish Dictionary")

In addition to this information there is also a short listing of multi-word units such as *stor bokstav* translating into 'capital letter' and *det ar stora pengar* translating into 'that's a lot of money'.

The dictionary entry above covers a whole lemma, i.e. all word forms belonging to the same headword. In the case of *stor,* the entry also covers the Swedish word forms *stort* and *stora*. Which form of this adjective to use depends upon number ('singular' or 'plural') and gender. In Present-day Swedish two genders exist: 'n-gender' and 't-gender'. They are referred to in this way to indicate the final letter of the definite form. Take for example the word *bil* ('car') which becomes *bilen* ('car_the') in the definite form, which tells us that the word *bil* belongs to the group of 'n-gender' words. Similarly, the word *hus ('house')* becomes *huset* ('house_the') in the definite form and therefore belongs to the group of 't-gender'. When you use an adjective in conjunction with a noun the adjective and the noun agree in gender. Therefore you say *en stor bil* ('a big car') but *ett stort hus* ('a big house'). Note that the form of the word *stor* changes to *stort* when used with a word that belongs to the group of t-gender. The table below illustrates the usage of the different forms.

Gender/Number	Singular	Plural
n-gender (such as *bilen 'the car'* and *hunden 'the car'*)	**Stor**	**Stora**
t-gender (such as *huset 'the house'* and *bordet 'the table'*)	**Stort**	

Table 2: The usage of *stor, stort, stora*

In order to retrieve all three forms when searching our corpus, we will make use of a 'wild card'. Rather than searching for three words, *stor, stort* and *stora*, we will ask for '*stor%'*. This should be interpreted as searching for all words that begin with *'stor'* and have zero or one additional character at the end. As such, it will match the form *stor, stort, stora*, but unfortunately it will also match forms such as *storm ('storm')* and *stork* ('*crane'*). Fortunately, in our corpus none of these additional words exist to garble our findings. In cases where such a wild card lists forms that are not relevant, it is the task of the software user to edit the concordance lines accordingly. Despite the fact that corpus linguistics is computer-dependent, human intervention is still very important on all levels.

While a dictionary may be the best general description of a language that currently

exists, it is only a compressed version of a language description. Due to space limitations (and other problems) it cannot cover all the possibilities that a language holds, nor does it provide the relevant contextual information. In the case of a young language learner much of the information in a dictionary entry, such as the one above, is cryptic: 'what do they mean with spacious?', 'which words can be described as *lengthwise*?' and 'when is a word denoting something abstract?' This is where parallel texts can provide useful information. As Barlow (2001) states: "What a parallel text provides is, effectively, an on-line contextualised dictionary which language learners can exploit. The equivalences are present, but importantly, so is the context". Compared to dictionaries parallel texts offer the advantage of allowing pupils to explore language themselves. The texts can be used more creatively and allow them to make their own observations about language.

In this study we do not restrict our view to only one language, but search for patterns and tendencies in two languages. The starting point is the source language since we are interested in possible translations of *stor*. However, differences between using one translation unit, rather than another, are found in the target language. For Swedish native speakers it may be difficult to see the differences between when to use *big* and when to use *large* or *great* as all words correspond to the word '*stor'* and information given in dictionaries and textbooks may often be confusing or even misleading. As we go through the examples here, we will see that the picture is complicated and that there are several other translation equivalents to *stor*. The function HotWords in ParaConc suggests the following significant English words as corresponding units to the Swedish word *stor*: *big, large, great, largely, huge,* and *grand* (see figure 3 below).

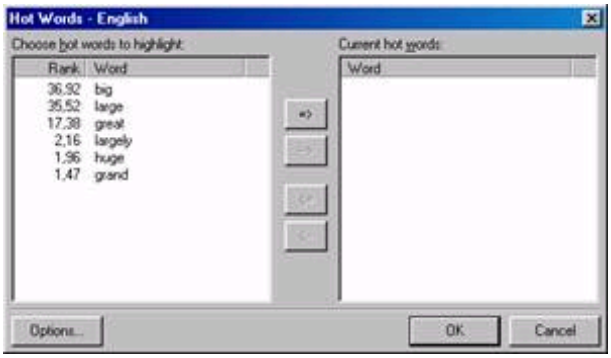

Figure 3 The 'HotWords' around 'stor%'

At a first glance these words seem to belong to a common semantic set sharing roughly the same meaning, although they may belong to different word classes (adjectives and adverbs). However, as will be shown below, pupils need to learn how to differentiate between when to use one in preference to another. What is worth pointing out is the discrepancy between the HotWord list and the information from the bilingual dictionary; *tall* is not found as a significant translation candidate in the parallel texts although present in the dictionary, while *largely*, *huge*, and *grand* are significant in the texts but not in the dictionary.

The first translation candidate on the HotWord list is the English word *big*. This is not surprising since *big* is usually seen as the prototypical translation equivalent of the Swedish word '*stor'*, although *large* was listed first in the bilingual dictionary. With *prototypical* referring to the answer most people would offer with given the question 'what does *stor* mean in English?' As *big* is a very frequent word the concordance lines cannot give us simply one dominant pattern or multi-word units. Instead, they show a tendency of *big* to occur in combination with a concrete object, such as *bike*, *camera*, and *book* (see concordances below). Even without the knowledge of linguistic categorizations young learners of a second language may be able to pick up this, 'big + concrete noun', as being characteristic of *big* when it is used as a translation for *stor*. If the pupils have problems identifying the pattern the teacher may steer them in the right direction by discussing various types of nouns.

... Sen tog hon fram en [[stor]] bok, jattetung, det var en av broderna ...

... Then she took out a [[big]] book, awfully heavy, by one of the von ...

... Bredvid ficklampan lag en [[stor]] kamera. Carl grubblade en stund o ...

... Near the flashlight lay a [[big]] camera. Carl aimed the remote con ...

... ut tidningar pa den tiden, pa en [[stor]] gammal damcykel som var hans...

... used to take the papers out on a [[big]] old bike of his mother's. ...

Example 4: The HotWord *big*

Continuing the exploration of the translation equivalents for the Swedish word *stor*, the next significant correspondence is the English word *large*. Amongst our concordance lines the following recurrent pattern appears: "*a large* X *of*", where X can be substituted by words such as *number*, *part*, *amount*, and *proportion*. This may be an indication to the pupils that *large* is often used for quantification. If the pupil has the time, a closer study would show that each of the multi-word units has a different usage: *a large* ***number*** *of* is mainly used to talk about *people*, *pupils*, *members* or even *companies* and *countries*, i.e. groups of people. On the other hand, *a large* ***part*** *of* is more often used to refer to difficulties, such as *a large part of the problem*, or even in the phrase *to shoulder a large part of the blame*. Observations such as these could be more problematic for the pupils to make on their own and the teacher may need to inform them of the larger surrounding context. Furthermore, it can be helpful to discuss semantic sets with the pupils. Discussions on semantic sets may be introduced with the help of a thesaurus.

... I sa fall skulle en [[stor]] del av ett vasterla ...

... Otherwise a [[large]] part of a partly Sw ...

... oligen skulle fa en [[stor]] del av skulden. ...

... e would truly get a [[large]] part of the blame. ...

... "Jag har agnat en [[stor]] del av dagen at att ...

... "I've spent a [[large]] part of the day try ...

Example 5: The HotWord *large* in the phrase *a large part of*

The third HotWord, *great*, offers a complementary use to that of the word *big* as it tends to be used in conjunction with abstract words, such as *difference*, *respect*, *effort*, *enthusiasm*, (see examples below). Furthermore, we find the multi-word unit *a great deal (of)*, which is, by mere coincidence, often used to express that *someone can learn a great deal*. The word *Great* (with a capital G) most often occurs in conjunction with *Britain*, forming *Great Britain*.

The translation equivalent *largely* is an example of a clear patterning in the source language, i.e. Swedish. This English word only occurs in two Swedish contexts in our corpus, either as a corresponding unit to *i stort sett* (six occurrences) or as a correspondence to *till stor del* (occurring twice). Although the frequencies are low, they are still convincing as no other corresponding units occur within the texts. Completely clean data, such as these illustrated in figure 4 below, cannot be expected to occur often. Language must still be viewed as a set of possibilities rather than as a list of rules. Yet, when learning a new language a few simple rules are

always easier to remember. Corpus data, as a reflection of real language, will not always provide this but when it does, as in this case, it is most convincing. If the picture is more complex, the teacher's role will be to steer the pupils towards more general patterns, which is easily done in the cases below.

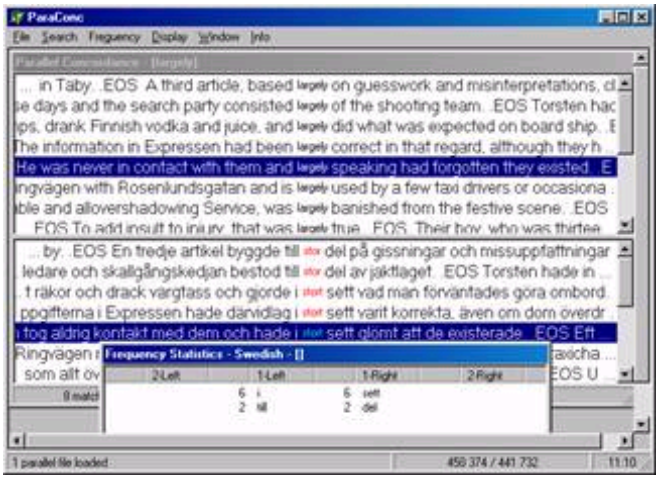

Figure 4: ParaConc.

In the window above the clear pattern can is illustrated in a separate 'Frequency Statistics Window', which here only has two lines (normally many more). This statistical information is retrieved by using the function 'Collocate Frequency Data', under the drop-down menu 'Frequency'. The first column, '1-Left', says *i* and *till* while the second column, '1-Right', shows *sett* and *del* respectively. In between we should add the keyword (stor%), yielding the two units mentioned above: '*i stort sett*' and '*till stor del*'.

The last two HotWords *huge* and *grand*, have too few occurrences to allow for any general observation. The HotWord *huge* seems to be used together with concrete nouns, similar to the use of *big*, but with only seven occurrences, compared to 207 in the case of *big* and no repetitive words, this may not yield the complete picture.

To sum up, our main findings are displayed in the table below.

SWEDISH	**ENGLISH**	**When to use it**
stor, stort, stora	*big*	When followed by a concrete noun, such as *bike, house* or *book*
stor	*great*	When followed by an abstract noun, such as *desire, respect* or *enthusiasm*
En/ett stor/stort X	*a large X of* *a large number of X* *a large part of X*	Where X denotes a quantifier such as part or number Where X denotes a groups of people, such as students, members, countries etc Where X often refers to *the problem* or *the blame*
Ta pa sig en stor	*shoulder a large part*	Fixed expression

del av skulden	*of the blame*	
i stort sett or *till stor del*	*largely*	Fixed expression

Table 2: Translation equivalents for *stor*.

By comparing the information from the corpus with that available from the dictionary we find several notable differences. The most obvious is the absence of the word *tall* in the translated texts. Though the word *tall* occurs in the English texts it does not occur as a translation of *stor*. Instead, this word corresponds to *lang* or *hog*. On the other hand, the word *largely* is not given in the dictionary even though the unit was found to be a stable translation to the units i *stort sett* and *till stor del*. Neither does the dictionary mention constructions such as *a large X of* or *a big X of*. This may be used as an illustration that natural language is much more complex than suggested by the simplified account in any dictionary or textbook. Thus, exploratory learning will inevitably offer pupils greater variety and possibilities than traditional resources can. It gives pupils the opportunity to experience real language and to complement the picture created by existing reference works.

6. Conclusion

In this article we could only give a brief introduction to the use of parallel corpora in the classroom. The opportunities that corpus linguistics opens for language teaching are vast, but a teacher will need to find the appropriate approach for his or her course. Any topic, ranging from vocabulary learning to text analysis, may be dealt with from a corpus linguistic perspective by making use of a variety of methods. The advantage of corpus linguistics is that it allows teachers to establish a direct link between theories about language and the facts revealed by natural language. In this article, we wanted to offer some brief ideas that may encourage teachers to start looking into the possibilities corpora can offer them.

References

Atkins, S.; Clear, J. & Ostler, N. (1992). Corpus design criteria. Literary and Linguistic Computing, 7, 1-16.

Barlow, M. (1995). ParaConc: A concordancer for parallel texts. Computers and Texts 10. (CTI Textual Studies)

Barlow, M. (2000). Parallel texts in language teaching. In S.P. Botley, T. McEnery, A. Wilson (eds), Multilingual Corpora in Teaching and Research (pp. 106-115). Amsterdam: Rodopi.

Biber, Douglas (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8, 4, 243-257.

Macmillan English Dictionary for Advanced Learners (2002). M. Rundell (ed.), Oxford: Macmillan.

Mahlberg, M. (in preparation). Corpus linguistics and English language teaching: bridging the gap between theory and practice.

Mindt, D. (2000). An Empirical Grammar of the English Verb System. Berlin: Cornelsen.

Nattinger, J. R. N. & DeCarrico, J. S. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.

Norstedts lilla engelska ordbok : engelsk-svensk, svensk-engelsk (1991). Vincent Petti & Kerstin Petti (eds.), Stockholm : Norstedt.

Peter, C., Picchi, E. and Biagini, L. (2000). Parallel and comparable bilingual corpora in language teaching and learning. In S.P. Botley, T. McEnery, A. Wilson (eds): Multilingual Corpora in Teaching and Research (pp 73-85). Amsterdam: Rodopi.

Sinclair, J. (1990). *Collins Cobuild English Grammar*. London: Collins.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: OUP.

Sinclair, J. (1998). The lexical item. In E. Weigand (ed.), Contrastive Lexical Semantics (pp. 1-24). Amsterdam: Benjamins.

Stubbs, M. (2001). Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Tesch, F. (1988). 'Some' und 'any' in affirmativen und negativen Kontexten. In D. Mindt (ed.), EDV in

der Angewandten Linguistik: Ziele, Methoden, Ergebnisse (pp 59-68). Frankfurt: Diesterweg.

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

Ungerer, F. (2000). Englische Grammatik heute. Stuttgart: Klett.

---

[1] Helpful introductions are, for instance, Kennedy (1998), Stubbs (2001), Hunston (2002).