Building specialised corpora for translation studies

Sattar Izwaini
Centre for Computational Linguistics, UMIST, PO Box 88,
Manchester M60 1QD, UK
Sattar.Izwaini@student.umist.ac.uk

Key words: Corpus Linguistics, Translation Studies, English, Arabic, Swedish.

Abstract

To study the strategies adopted by translators to render the vocabulary of information technology into two languages, namely Arabic and Swedish, three specialised corpora have been built: a corpus of Information Technology in English (ITE corpus) and two corpora of translational IT: Information Technology Translational Arabic corpus (ITTA corpus) and Information Technology Translational Swedish corpus (ITTS corpus). This paper discusses design issues, practical procedures, and what considerations we need to have in mind when specialised multilingual corpora are built for translation studies. [1]

1. Background

Using corpora has a potential impact and great significance in the empirical investigation of translation strategies. In recent years, there has been a great interest in using corpora for the investigation of translations. This led to the emergence of corpus-based translation studies as “a major paradigm in the field” (Baker 1999: 287).

There are different terms used for corpora whether monolingual, bilingual/multilingual or translation oriented. These terms are not established and lead to confusion (McEnery 1997:1; Baker 1995:230). For translation studies, Baker (ibid) proposes three types of corpora: parallel, multilingual and comparable corpora. The first consists of original SL texts and their translations, for example the Lancaster-Oslo parallel corpora (Johansson and Hofland 1998). This is also called translation corpora (McEnery 1997; McEnery & Wilson 1996; Johannson 1998). A multilingual corpus incorporates two or more monolingual corpora in different languages, such as the European Corpus Initiative Multilingual Corpus I (ECI/MCI) (ELSNET web site), and the Edinburgh Multilingual Corpora for Cooperation (MCC) (Armstrong et al. 1998). Comparable corpus refers to a corpus of two collections of texts: the first consists of texts translated into one language, whilst the second consists of original texts in the same language. For example, TEC (Translational English Corpus) which is housed at the Centre of Translation and Intercultural Studies at UMIST (Baker 1996; Laviosa-Braithwaite 1997).

Although the terminology he uses is partly different, McEnry (1997) reports on some parallel corpora in which Lancaster University has been involved, such as CRATER and MULTEXT. The first involves *inter alia* adding a third language, namely Spanish, to a 1.5 million word corpus of French/English parallel technical text. The second includes developing a bilingual English/French corpus of 200,000 words. Other examples of parallel corpora include the IJS-ELAN Slovene/English parallel corpus (Erjavec 2002), and the Croatian-English parallel corpus (Tadić 2001) as well as the Scania corpus which is held at Uppsala University. It is a specilised corpus with Swedish as the source language and eight target languages (Uppsala University web site).

Corpora have different profiles according to their composition and aims. With regards to information technology, to my knowledge, there is one parallel ‘translation’ corpus which was built to conduct a study on texts related to this field. Scarpa (1999) carried out a study based on an English/Italian specialised parallel corpus (150,000 words) of technical documentation of Microsoft Office 97. The aim was to provide trainee translators with a description of the strategies used to render genre-specific translation difficulties.

2. Aim

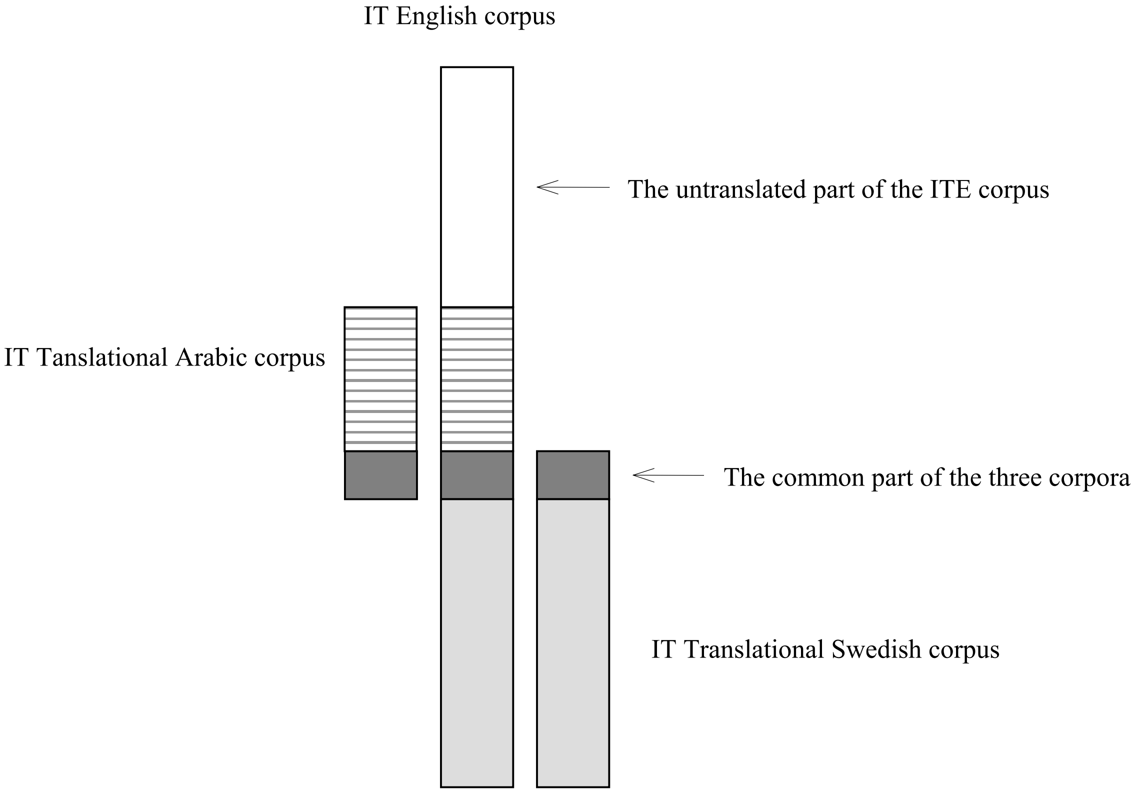
The variety of English that is used within the field of information technology to describe, explain and transfer knowledge and expertise can be called the language of information technology (LIT). It

[1] All trademarks referred to herein are acknowledged.

addresses specifically the use of the computer, its systems, software, hardware and peripherals, and networks as well as the internet. It is a special language with special terminology. This variety of language can be found in online help of computer systems and software, hardware and software manuals, user interface, tutorials, books and research papers closely related to IT. Needless to say, information technology, its language and terminology were born in an English-speaking environment, American in particular. Thus IT terminology has emerged in English and has been transferred into other languages.

Here we present a detailed account of design issues, practical steps taken and factors that played a role in constructing the three corpora built for the study: a corpus of Information Technology in English (ITE corpus) and two translational corpora IT: Information Technology Translational Arabic corpus (ITTA corpus) and Information Technology Translational Swedish corpus (ITTS corpus); see Figure 1 below. The corpora reported here are built to investigate how the lexis of information technology, in general, and lexical collocations in particular, are rendered into Arabic and Swedish.

In section 3, the structure of the material and collection procedure as well as issues of representativeness, scope and size are discussed. Section 4 describes the methodology of processing the corpora. In section 5, we give an outline of how to detect translation equivalents and their lexical collocations.


IT English corpus
The untranslated part of the ITE corpus
IT Tanslational Arabic corpus
The common part of the three corpora
IT Translational Swedish corpus

ITE corpus	ITTA corpus	ITTS corpus
7 million tokens	1 million tokens	2.7 million tokens

Figure 1: Size and distribution of corpora

3. Structure of material

English texts in the field of information technology and translations in Arabic and Swedish are collected. Texts are mainly online help for both computer systems and software, hardware and software manuals as well as from multilingual IT-specialised web sites. The ITE corpus is designed to produce a key word list of this variety of language. Availability of material has made the translational corpora much smaller (see figure 1). The original texts of the translational corpora are incorporated in the ITE corpus. There are three textual components (one of them is a major one: online help of Windows 98 and Office 2000) that are common originals for both translational corpora. This can help in looking at how IT is rendered into two languages of considerably different distances from the source language (SL).

The three corpora are not 100% parallel. Firstly the ITE corpus is designed to produce a key word list in the field of IT in the SL and thus has to include texts from different IT sub-fields whether they have translated versions or not. The untranslated part of ITE is those texts that give some balance in terms of representativeness; to include other platforms such as UNIX or to include material that has been produced to cater for different kinds of hardware and software. Secondly, because the originals of the translational corpora are only partly identical. It is worth mentioning here that the PC platform and Windows system, especially Microsoft web sites, have more material, monolingual and multilingual, available than other systems. This makes its share much more than other operating systems, especially with Windows 98 and Office 2000 being major components of the corpus.

ITE corpus has two purposes: to identify key words and their lexical collocation in IT, and to serve as an SL corpus. The two translational corpora serve to account for the translation equivalents of key words and their lexical collocations. Translated texts in both TLs had to be found first so that their originals can be included in the SL corpus. Size and range of all three corpora are subject to methodology, availability and obtaining permission from copyright holders. These are discussed below.

3.1 Representativeness, scope, and size

The aim of study, representativeness, availability of material, and obtaining permission dictates the size and scope of the corpora. To cater for the first purpose of ITE – to give an account of key words in English IT and their lexical collocation - it is of great significance to have a corpus that covers a wide range of systems and software (Apple, UNIX, Windows), and as many as possible of the sub-areas of this field: online help, manuals, readme files, books, research papers, tutorials, IT news items, IT journalistic articles, and web sites (specialised in software industry, web design, web hosting and internet service). As mentioned above, the lexical aspect is the prime focus of research. Thus the ITE corpus was built be as big as possible. As Johansson (1995: 246) puts it:

> Any Corpus must be seen against its intended uses. Corpora compiled for lexical studies must be large, to provide sufficient evidence for less frequent words and for collocational patterns.

Full texts are included in the corpus following a holistic view of including all the text rather than samples. Pioneering studies in corpus linguistics included samples of comparable size. Corpora such as the Brown corpus, Lancaster-Oslo-Bergen corpus and Macquarie University corpus consist of 500 samples of ca 2000 words each (Taylor et al., 1991). However, there is another view that advocates including as much material as possible. With less availability of machine-readable texts and technical limitations in the early stages of corpus linguistics, it is understandable that size and presentation of language could not go as far as we can now. At the same time an even size of samples from different fields is justifiable in order to have reliable results. It has been deemed necessary to have full length texts rather than samples of pre-determined size (Baker 1995: 240; Sinclair 1995:18, 27). If data samples of defined size are the choice, the question is what are the border lines of the SL text and the corresponding translated text? It is impractical to have a sample-text translation corpus due to the different word order, syntax and punctuation rules of the languages involved. Moreover, translations often have a different length from their originals. (See *5. Translation Equivalence* below regarding alignment of corpora).

The diversity and balance of presentation are to be taken into consideration in the process of collecting the data. A balance between the sub-types of texts representing IT in terms of size is by no mean attainable due to other factors such as availability and permission for use. For example, Arabic material of the Mac OS9 online help could not be found, whereas Arabic Windows online help is easily obtainable. Furthermore it constitutes a substantial part of the ITTA corpus.

Permission to use texts is an important factor that has an impact on the structure of the corpora. Acquiring permissions from a copyright holder takes long time and sometimes a lot of correspondence. In some cases, the copyright holder asks for details of the research to be carried out on the texts. In one case, the present author was asked about his nationality, and in another case he was asked about his political and religious beliefs! Waiting for permission can stretch to 6 months or more, for example, at the time of writing, no permission to use the *Answer Book* of *Solaris 8* (a UNIX system) could be obtained although the first contact with the copyright holder was made in February 2002.

3.2 Collection procedure

The collection process started by visiting the websites of IT companies and checking whether they have localized versions of their English websites, at the same time looking for other not well-known companies by checking web directories. Search engines and web directories are used extensively to look for IT specific web sites. At the same time IT web sites are checked for documents and archives that might include monolingual or multilingual material. Even if they do not have such material, free or demo versions of software and tools - whether they have an English version only or localized versions - usually include textual material in the form of online help, readme files and/or software guide.

Textual material is collected mainly from online help files and from the Web. Online help files are in two formats: *chm* and *hlp* files. The first are compiled html and xml files that need to be decompiled in order to be able to process them later on. The program KeyTools (of KeyWorks Software and Work Write, Inc.) is used to decompile the chm files. The hlp files are browsable help files for which the Microsoft Help Workshop tool is used to retrieve their texts. Decompiling chm files results in having other non-text files, which need to be removed. Names of files (whether chm, hlp or html files) are the same in the three languages, which makes it easier in terms of corpora organization and parallel processing. Multilingual files that are created by collecting texts from websites are given corresponding names for the same purpose.

Many files in the localized online help are left untranslated. These are removed so that they would not be accounted for in the target language (TL). This will be discussed in *5. Translation equivalents* below.

In the case of web sites, the procedure of including the material is ‘copy and paste’ and in the case of books and research papers, it is downloading and saving them in text format. Although saving web pages in HTML format or as text would be straightforward, it causes difficulty in opening them and deleting irrelevant material in the case of HTML format, or in including undesirable or irrelevant sections that have to be removed in a later stage in the case of text format. So ‘copy and paste’ is better in this respect, especially for alignment purposes.

As some link buttons are in a picture format or other format, they are lost along with the accompanying text or words when copied on-line. In order to include the text, they have to be keyed in their appropriate places. Some web pages have such features that include texts. These are only shown when the mouse is clicked or have drop menus that can be seen when moving the mouse, where the cursor goes over the main menu. These texts are traced and typed in.

In selecting translated material from web sites, a number of points are taken into consideration. Some companies have a main web site for the US market with a facility to choose web sites of the same company in other countries. In such cases the UK web site is chosen for the English version (SL) as it is more or less the same as the American web site. As the SL version is included in the ITE corpus, the SL site has to be the UK home page even if the TL home page has its own English version. This is to make sure that the ITE corpus includes material written by native speakers. Some Arabic IT web sites have their headquarters in the UK or have American or British staff and thus their English version is written either by a native speaker or by someone whose language of habitual use is English. In other cases where the web site has an English version but no UK home page, it is presumed that the writer has a specialised knowledge of the English language and the terminology of the field, so that his/her text can be included in the ITE corpus. Without this possibility, the translational corpus would be very small in size and would not yield reliable results.

The criterion to select texts from web sites to be included in the corpora is that texts have to be closely related to IT and to have translated versions in the TLs. Some web sites have links in the TL version that go back to the SL web site, so the SL links are not included in the parallel corpus. Some headings in the TL sites are linked to the SL version. In some instances the content is the same but it is not highlighted and presented to the reader in the same place or by the same link. This brings us to the question of whether or not to include the untranslated links. As every website is saved in one file, the decision was to include those links as part of the SL text with zero TL corresponding text. Some IT news stories that occupy the home page of a web site are translated only partly or they are summarized. The SL text usually has a link for more information or to read the story in full. The TL version of the site has a translated version of the introduction but the read-more link takes you back to the SL full story. Those links are not included.

Web sites are usually updated from time to time, particularly in computing and Internet electronic magazines. This has its effect on collecting the data especially if upgrading and adding new material is carried out during the time of retrieval, as one version of the web site, most probably the TL page, might not be updated then. Therefore the collection of the SL and its corresponding TL data has been carried out at the same time or within a very short period to avoid discrepancy. The SL and TL versions of web sites are opened in two windows of the web browser at the same time and their contents are compared against each other to exclude material of a product or service found on only one page.

In some cases, design preferences and layout of web pages make the order of the textual material different in both SL and TL versions. Even within one page, not all information found in the localized version might have corresponding information in the SL page, but it might have in another page. This has to be catered for and thus corresponding pages are put in the sequence in both the SL and TL texts. According to their order in the SL, the links might not correspond to the links in the TL. The sequence of the texts (in order of pages) then has to be maintained for technical reasons. Every web site has one file in the SL and a corresponding one in the TL that include the web pages. Thus these pages have to be put in the same order within each file. When processing the text these files have to be corresponding fully to each other.

The relevant IT sections are selected and incorporated in the corpora. Only texts that are specifically related to IT and computing are included. Such texts are those which describe, explain and discuss different aspects of the field of IT. Some web sites have links to specific fields that are beyond the scope of the research. Hence legal sections and notes as well as irrelevant sections, such as bibliographies and indices are excluded. For example, disclaimers and copyright notes as well as warranty pages are deemed to be unrelated to the field, as they manifest a language variety other than IT. They are legal documents and include legal language and terminology that would find their way into the frequency list and eventually to the key word list of ITE if they were included. It is worth mentioning that these irrelevant sections are separate independent documents within the website or a set of online help files that cannot affect the integrity of other files included.

Abstracts of research papers are kept as they are deemed to represent an integral part of the text. Repeated menus in different pages of the same web site have been removed to avoid over-representation of key words. Links and web pages that include information related to aspects of life that are not IT specific, such as introducing executives and the role of the company in the community, have not been included.

Cross-reference is not included in order to avoid multiple presentation, because the same topic is stated elsewhere in the text and at the same time to keep consistency of the steps within the overall outline of the text. Links of cross-reference create a problem by over-representing some lexical items as they are mentioned in the heading list and as a sub-heading as well as at the end of other sub-sections. This applies to 'help' links as they are found in the main list as well as at the end of every sub-heading.

Texts are then looked at and edited for blemishes resulting from the converting process. For example some PDF files lose some lines and specific letters when converted into text format.

4. Processing the corpora

Two programs are used for the processing and study of corpora in a Windows environment: Wordsmith Tools (Scott 1997), and ParaConc (Barlow 1995). The first is used to produce word lists and a key word list of the ITE corpus as well as concordances of key words. The second is used to find out the translational equivalents and their lexical collocations in the translational corpora.

The key words tool in Wordsmith works by comparing the frequency word list of the corpus under study with the frequency word list of a general language reference corpus which is the BNC in this case. Key words are those whose frequency is unusual in comparison with the reference corpus.

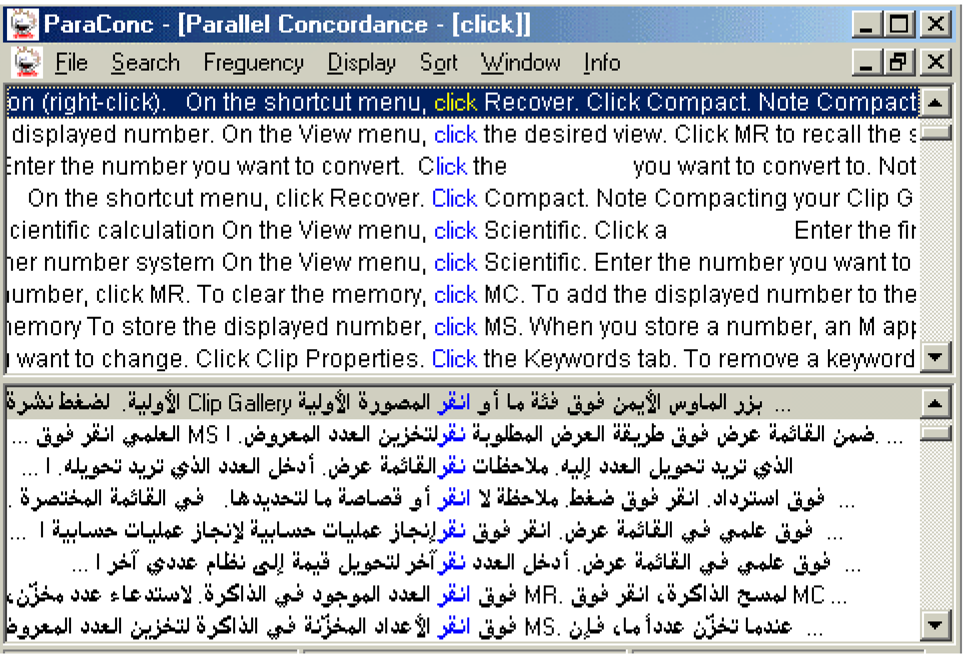
Writing system rules and traditions affect the results obtained. At the same time, choosing different settings according to which the corpus is processed has its bearing on the results as well, for example, selecting hyphenation as a word marker. The key word *click* can have different values of frequency and keyness depending on the settings chosen. This is because it is found in the corpus to be written with or without a hyphen, e.g. *double click*, *double-click* etc.

Span and frequency of co-occurrence are important factors in shaping the collocational pattern of a node. To account for the collocational patterns of the SL and TL key words, we need to have these two values interacting to have a good perspective on the lexical collocations. We should not look only at the closest slots on both sides of the key word node, although adjacent collocates are far more interesting than those which occur further away in the string. At the same time, highly frequent collocates within different distances from the key word are more interesting than those with less frequency even if they occur just next to the node.

Alignment is crucial for a translation-oriented parallel corpus. Generally speaking, html files are more manageable than text files. Html files have a higher percentage of alignment than text files that are converted from help files. Due to the conversion process, the latter tend to have sentences broken into two different paragraphs and even broken words. In most cases, text files have to be 'adjusted' manually to have the whole paragraph aligned to get the best results. This is due to the fact that languages have different punctuation pattern and word order. For example, Arabic tends to have long sentences in that it uses many conjunctions and coordinating particles before the sentence comes to a full stop, which is considered by the concordance program as the sentence boundary. This leaves all other sentences in the SL paragraph without corresponding sentences. Hence, alignment sometimes has to be within paragraph boundaries so that TL corresponding words can be traced. Other elements that affect the alignment of texts are layout and punctuation. Translated texts sometimes have different layouts than their originals or have picture captions in different places and thus they do not correspond fully to their SL counterparts. Periods that are used in numeration, at the end of acronyms and abbreviation as well as within file names and web sites addresses (URLs) are considered as the end of the sentence by the software used and thus may lead to confusion in the identification of sentence boundaries. This needs special attention, because languages do not have the same punctuation rules nor the same acronyms and abbreviations, e.g. in Swedish we have *t. ex.* (for example), *p. g. a.* (because of/due to).

The Arabic corpus is processed in an Arabic-enabled Windows system. Thus texts are looked at in their own original orthographic form. Although the corpus is untagged, processing Arabic in this environment helps in solving many problems associated with Arabic such as polysemy and ambiguity due to non-vocalization of words in their written forms. However, words that have diacritics showing short vowels can be problematic in the search process. Such words have to be written exactly the same in the searching window of the concordancer so that they can be found.

Concordances of Arabic words appear to be 'distorted' sentences with the key word in the middle and its next word at the far right (see the figure 2 below). When looking at KWIC (Key word in context) in Arabic concordances, the L1 collocate is either in the R1 slot or at the far right. The sentence usually starts in the L1 slot, interrupted by the node and continues at the far right. At the same time, the concordance tool treats the text as if running from left to right. Thus the collocates that are actually to the right are labeled as left and vice versa. This becomes even more problematic if a sentence includes English words in Latin script. Only when the concordance of a key word is saved in a text file does the concordance line become tidy and make sense.



on (right-click). On the shortcut menu, click Recover. Click Compact. Note Compact
displayed number. On the View menu, click the desired view. Click MR to recall the s
nter the number you want to convert. Click the you want to convert to. Not
On the shortcut menu, click Recover. Click Compact. Note Compacting your Clip G
cientific calculation On the View menu, click Scientific. Click a Enter the fir
ner number system On the View menu, click Scientific. Enter the number you want to
umber, click MR. To clear the memory, click MC. To add the displayed number to the
nemory To store the displayed number, click MS. When you store a number, an M ap
want to change. Click Clip Properties. Click the Keywords tab. To remove a keyword

... بزر الماوس الأيمن فوق فئة ما أو انقر المصورة الأولية Clip Gallery الأولية. لضغط نشرة
... ضمن القائمة عرض فوق طريقة العرض المطلوبة نقرلتخزين العدد المعروض. ا MS العلمي انقر فوق ...
الذي تريد تحويل العدد إليه. ملاحظات نقرالقائمة عرض. أدخل العدد الذي تريد تحويله. ا ...
... فوق استرداد. انقر فوق ضغط. ملاحظة لا انقر أو قصاصة ما لتحديدها. في القائمة المختصرة .
... فوق علمي في القائمة عرض. انقر فوق نقرإنجاز عمليات حسابية لإنجاز عمليات حسابية ا ...
... فوق علمي في القائمة عرض. أدخل العدد نقرآخر لتحويل قيمة إلى نظام عددي آخر ا ...
... MC لمسح الذاكرة، انقر فوق MR. فوق انقر العدد الموجود في الذاكرة. لاستدعاء عدد مخزّن
... عندما تخزّن عدداً ما، فإن MS. فوق انقر الأعداد المخزّنة في الذاكرة لتخزين العدد المعروض

Figure 2: Concordances of *click* and its corresponding Arabic *'unqur* [2]

ParaConc treats Arabic text as if running from left to right. Thus it allocates collocation slots in a reverse way. The linguistic knowledge of the research helps in knowing the real syntagmatic order. At the same time, as the Arabic words are envisaged and presented to non-Arabic speaking readers from left to right, the left hand slots (L1, L2 …) and right hand slots (R1, R2…) are to be looked at as locations of words occurring before and after the node respectively. However, because of the difference

[2] As Arabic has no capital letters, upper case is used here to denote the long vowel ا (alif) in order to differentiate it from the short vowel represented by *a*. For 'ayn ع, ^{c} is used and for the glottal stop, *hamza*, an apostrophe (') is used.

in the word order and translation techniques the corresponding collocates occupy different slots, for example, taking *computer* as the node:[3]

1. L1 Node R1
Marketing computer programs
L2 L1 Node
taswiiqu barAmiji al-kombiyutar
(marketing programs the-computer)

We can see that the SL L1 node corresponds to the TL L2 and the SL R1 corresponds to the TL L1. In reality, however, TL L1 is R1 and L2 is R2. When it comes to the two kinds of Arabic sentences (nominal and verbal) the corresponding slots are even more difficult to identify, e.g.

2. L1 Node R1 R2
The computer stopped working

2. A verbal sentence:
L1 Node R1 R2
tawaqafa al-kombiyutar ^{C}an al-^{C}amali
(stopped the-computer from the-work)

2. B nominal sentence:
Node R1 R2 R3
al-kombiyutar tawaqafa ^{C}an al-^{C}amali
(the-computer stopped from the-work)

As can be seen, the verb collocate in 2A occurs in L1 whereas it occurs in R1 in 2B. Only when the sentence has a verb in the imperative is there some correspondence between the two patterns as illustrated by the TL verb *'unqur* (click) in the following example:

3. Click on the file name
'unqur ^{C}alA 'ismi al-milafi
(click on name the-file)

This is less crucial in English/Swedish translation as the word order has less difference than that of Arabic. Hence in making a cross-linguistic account of collocates, their position whether to the left or to the right is not of much significance, but what counts is how close or far they are from the node.

Recognition of words and counting the size of a corpus makes comparison difficult. In the case of Swedish, compounds tend to have their constituent parts as one orthographic word, e.g. *systembuss* (system bus) and *nätverksoperativsystem* (network operating system). In the case of Arabic, an orthographic word is often more than one linguistic word. Some Arabic orthographic words can be made of up to five linguistic words. The Arabic definite article, some conjunctions and coordinators, pronouns in the accusative are normally attached to other words such as nouns, verbs, and prepositions, e.g. *'afasatuqAbilahum*? (so, are you going to meet them?) is composed of five attached words. In the ITTA corpus many words have three linguistic words, e.g. *litashGeelahu* (to operate it).

One Arabic letter, the glottal stop *hamza*, is written with different letters according to the vowel that precedes or follows it. If it is initial, it is written with a silent *alif* (A). Initial *hamza* is not always written, but rather the *alif*. It can be written in different ways: إ أ ا. For example, *click* is translated into Arabic as *'unqur*. This word can be found in two different forms depending on whether the initial *hamza* is written or not. Thus the same word with the *hamza* being written or not can be recognized as two different words by the concordance program.

5. Translation equivalents

Translation equivalents are looked at as being the corresponding TL lexical items that translators used to render SL key words and their lexical collocations. Translators opt for different strategies to translate the language used in the field of information technology. As one of the prominent features, if not the

[3] As the definite article, some prepositions as well as some pronouns are attached in Arabic writing and do not stand alone, the concordance tool does not consider them as collocates.

most prominent one, lexis is the main concern and focus of the translation process of a special language. The main aim here is how IT vocabulary is translated; how translators go about looking for and creating TL vocabulary and collocates for the SL vocabulary and collocates in this field.

Tracing corresponding TL lexical items gives insight into the way translators deal with IT vocabulary such as novel terms, e.g. *byte*, lexical items that have been generated by figurative use and compounding, e.g. *desktop* and *search* engine (Izwaini, forthcoming), as well as acronyms. On the other hand, there are different strategies adopted by translators to render lexical collocations. TL lexical items that are found to be renditions of SL key words have collocational patterns that may or may not correspond to the original patterns (Izwaini 2003). It is important to find out how translators make these choices.

To investigate how lexis is translated, SL and TL texts are preferably aligned at the sentence level, but this is very difficult. Although concordance programs can give a good percentage of alignment, human intervention is needed. Length of sentences, punctuation pattern, word order and syntax of the languages investigated play a major role in having non-corresponding sentences. Hence, a paragraph-correspondence alignment has to be adopted. The linguistic knowledge of the researcher is a valuable asset in tracing equivalents within the number of hits produced by the concordance program. ParaConc does not give the precise corresponding TL lexical items, but rather shows the sentences where potential equivalents can be. When these are asked for, they can be highlighted and put in a KWIC concordance (see figure 2 above). Having the exact correspondence of texts in both the SL and TL can contribute greatly to the process of tracing equivalents and having reliable results. Such results are very much affected by the size, choice, and correspondence of parallel texts. Initial results in the first stage of the study are considerably different from those when the corpora have been expanded in terms of size and scope.

As mentioned above, the ITE corpus is used to account for key words of this special language in English. The lexical collocations of key words are accounted for only in the parallel part of the corpus. The collocational patterns have to be established in the English originals of the translational corpora so that a cross-linguistic comparison can be made. Once key words are identified, their equivalents are detected using the ParaConc program. To make an account of TL collocates, the distance within which they occur and their frequency are more important than their position whether to the left or to the right of the node (see the discussion in *4. Processing the corpora* above).

Tracing corresponding words in the TL is not an easy task. Having different written forms of the same word can result in having different collocational patterns that need to be put together to have the final result (see the discussion on Arabic *hamza* in *4. Processing the corpora* above) In the case of Arabic, transliterated borrowed words can have more than one form in the TL, e.g. *computer* and *internet*. This is because of the realization of the vowels of the SL word in Arabic.

SL lexical items, especially terms and acronyms, which are usually found in Latin script in both Arabic and Swedish texts, help in finding their equivalents which are usually, but not always, found next to them. Such SL lexical items are present in their original form either due to lack of a corresponding word, for the sake of brevity in the case of acronyms, or to clarify the reference of the TL technical vocabulary.

Having special design and layout preferences of translated texts and localized web sites, translations tend to have a different profile in terms of lexical collocations. Some collocates have picture counterparts in the TL text which disappear when saving it in text format. This leads to the collocational pattern being disturbed. Moreover, many online help files are left untranslated and thus not included in the translational corpora. Although this can be considered as zero translation from a theoretical point of view, such files are not processed within the parallel corpora because they have no translation equivalents, but rather the same SL words.

A distinct feature of translated IT texts is what can be called a ‘skeleton’ translation. We find only headings and links titles in the TL but not the content. Online help files have their headings translated but not the whole text. Although some web sites have parallel sites in two languages, the two sites are not fully translated. Only the main menu, titles of links and features are translated. Many topics under the same headings or links are not the TL version of those in the original. Such ‘skeleton’ translation seems to be triggered by the consideration of the TL market as well as by the delay of carrying out translations from SL into TL. In the case of online help files, however, there seems to be no reason as why the heading or the title is translated but not the content.

6. Conclusion

This is a general overview of the methodology adopted in creating a multilingual corpus of the special field of Information technology with the aim of studying translation strategies. There are many considerations to be taken into account when compiling such a corpus especially if it is a specialised one with the aim of investigation translation equivalence on the lexical level. These include scope, size and range of texts included. Other factors that play a significant role in the profile of a corpus include availability and copyright permission.

References

Armstrong S, Kempen M, McKelvie D, Petitpierre D, Rapp R, Thompson H, 1998 *Multilingual Corpora for Cooperation (MCC)*. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp. 975-980.

Atkins S, Clear J and Ostler N 1992 Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1):1-15.

Baker M 1993 Corpus Linguistics and Translation Studies, Implications and Applications. In Baker M, Francis G and Tognini-Bonelli E (eds), *Text and Technology, in Honour of John Sinclair*. Amsterdam, John Benjamins, pp 233 -250.

Baker M 1995 Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7(2): 223-243.

Baker M 1996 Corpus-based translation studies: The challenges that lie ahead. In Somers H (ed.), *Terminology, LSP and Translation*. Amsterdam, John Benjamins, pp 175 -186.

Baker M 1999 The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. *International Journal of Corpus Linguistics* 4(2): 281 -298.

Barlow M 1995 *A Guide to ParaConc*. Houston, Athelstan.

Biber D 1993 Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4): 243 - 257.

Biber D, Conrad S and Reppen R 1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, Cambridge University Press.

ELSNET web site: http://www.elsnet.org/resources/eciCorpus.htm visited on 27 February 2003.

Erjavec T 2002 The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics* 7(1): 1 -20.

Izwaini S 2003 Cross-linguistic Study of Lexical Collocations in the Language of Information Technology. In *CLUK 2003*, *Proceedings of the 6th CLUK Colloquium, 6-7 January 2003,* Edinburgh. pp 135-139.

Izwaini S (forthcoming) A Corpus-based Study of Metaphor in Information Technology. A paper to be presented at the Workshop on Corpus-Based Approaches to Figurative Language, 27 March 2003, and to appear in the proceedings of Corpus Linguistics 2003, Lancaster, 28 - 31 March 2003.

Johansson S 1995 ICAME – Quo Vadis? Reflections on the Use of Computer Corpora in Linguistics. *Computers and the Humanities*, 28:243-252.

Johansson S 1998 On the role of corpora in cross-linguistic research. In Johansson, S and Oksefjell S (eds), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam, Rodopi. pp 3-24.

Johannsson S and Hofland K 2000 The English-Norwegian Parallel Corpus - Current Work and New Directions. In McEnery A M, Botley S P and Wilson A (eds), *Multilingual Corpora in Teaching and Research*. Amsterdam, Rodopi. pp 134 -147.

Laviosa-Braithwaite S 1997 How Comparable Can 'Comparable Corpra' Be?. *Target*, 9(2): 289-319.

McEnry, A M 1997 Multilingual Corpora – Current Practice and Future Trends. In *Proceedings of the 19th ASLIB Machine Translation Conference*, London. pp 71 -83.

Scarpa F 1999 Corpus Evidence of the Translation of Genre-Specific Structures. *Textus*, 12: 315-332.

Scott M 1997 *WordSmith Tools, version 2.0*. Oxford, Oxford University Press.

Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.

Sinclair J 1995 Corpus Typology – A framework for classification. In Melchers, G. and Warren, B. (eds), *Acta Universitates Stockholmiensta, Stockholm Studies in English LXXXV*. Almqvist & Wiksell, Stockholm. pp 17 -33.

Sinclair J 1997 Corpus Linguistics at the Millenium. MS.

Tadić M 2001 Procedures in Building the Croatian-English Parallel Corpus. *International Journal of Corpus Linguistics* 6(special issue): 107 -123.

Taylor L, Leech G and Fligstone S 1991 A Survey of English Machine-readable Corpora. In Johansson S and Stenström B (eds), *English Computer Corpora: selected papers and research guide*. Berlin and NY, Mouton de Guyter, pp 319 -353.

Uppsala University web site: http://stp.ling.uu.se/~corpora/#research visited on 1 march 2003.