Multilingual corpora in cross-linguistic research. Focus on the compilation of a Dutch-Swedish parallel corpus

Gudrun Rawoens

Department of Nordic Studies– Ghent University – B-9000 Ghent – Belgium

Abstract

The topic of this paper is the use of corpora in general and in cross-linguistic research in particular. Special attention will be given to multilingual corpora and to the compilation of the Dutch-Swedish parallel corpus. In the first part of the paper a few reasons for using a variety of corpora will be adduced. The use of empirical data for linguistic research has gained momentum since large text corpora have become accessible. These corpora can both be used as underlying data on the basis of which hypotheses can be formulated and as a tool for verification of these hypotheses. A discussion will follow of a few theoretical issues concerning cross-linguistic research and corpus use with particular emphasis on the intrinsic value of multilingual corpora in the detailed study of one particular language. The second part of the paper will focus on the compilation and use of the SALT Dut-Swe corpus, a Dutch-Swedish parallel corpus containing fiction and non-fiction texts, comprising a total of three million words.

Keywords: multilingual corpora, cross-linguistic research

1. Introduction: Corpus linguistics

By way of introduction a definition of corpus linguistics is given, followed by a discussion of its applicability and a summary of the widely discussed pros and cons of corpus use.

Corpus linguistics is commonly defined as the study of language based on text corpora representing authentic language use (e.g. Aijmer and Altenberg, 1996:1; McEnery and Wilson, 2001: 1). Most scholars seem to agree upon the fact that corpus linguistics does not refer to «a domain of study, but rather to a methodological basis for pursuing linguistic research» (Leech, 1992: 105). In other words corpus linguistics refers to a heuristic method used in linguistic studies.

In defining what a corpus really is, McEnery and Wilson (2001: 29) mention four criteria that a corpus should fulfil: a corpus should have representativeness, its size should be finite, it should come in machine-readable form and it should serve as a standard reference. If these conditions are not fulfilled one should rather speak of a collection of texts than of a corpus (McEnery and Wilson, 2001: 103).

Since their emergence, corpora have been used in a wide range of domains. For instance, corpora have been widely used within lexicography. A corpus can illustrate how a word or a phrase is being used in a certain context or illustrate its syntactic properties in a given sentence (*e.g.* Atkins et al., 1992: 13). The use of empirical data in linguistic research has gained momentum since large text corpora have become accessible. Corpora can be used in purely empirical or even in more theoretical studies (*e.g.* Atkins et al., 1992: 14). McEnery and Wilson (2001:

103-132) mention a number of areas of linguistic studies where corpora can be used, such as semantic studies, psycholinguistics and cultural studies.

The linguist can use a corpus as an empirical base for illustrating and verifying hypotheses, an approach which is commonly denoted as *corpus-based*. Conversely, a corpus can also be used as the starting-point for a certain study or hypothesis, an approach denoted as *corpus-driven* (Ooi, 1998: 51; Tognini-Bonelli, 2001: 65-98; Altenberg and Granger, 2002: 15). The former approach is deductive or top-down because the linguist starts from a theory or hypothesis that is applied to the corpus data in order to validate or refute it. If the approach is corpus-driven it is inductive (bottom-up) since the linguist aims at arriving at a theory starting from the empirical data. In practice, though, it is often necessary to combine both approaches (Ooi, 1998: 52).

The pros and cons of the use of corpora in linguistics have been discussed extensively.

One of the apparent advantages is that corpora allow the linguist to pursue a more objective way of working than if he or she would simply rely on his or her own linguistic competence. The interest in, and a reliance on corpora, have come about from the insight that introspection alone is not sufficient (*e.g*. Sinclair, 1991: 39). Sinclair (1991: 5) is very negative about language examples that are invented by the linguist himself or herself since what can represent language in a better way than real examples? What is more, the linguistic intuitions of just one person are simply unreliable (Ooi, 1998: 48). In other words, there are obvious limitations to the linguist's subjective – not necessarily correct – language intuition (cf. Stubbs, 1996: 28). This criticism goes back to Chomsky's view that the object of linguistic studies should be competence and not performance. According to Chomsky, the only way to study competence is through introspection, the linguist's own (native) language intuition. Performance as an expression of real language use is influenced by a great many external (*e.g*. situational) factors and can therefore not reflect competence (see *e.g*. Cook, 1988; McEnery and Wilson, 2001: 5-10).

In case the linguist is studying a language that is not his or her own, a corpus offers the clear advantage of providing the linguist with a great amount of authentic language material (Leech, 1991: 74).

On the other hand, many (*e.g*. Svartvik, 1992: 10) have warned against blind reliance on corpora. First of all, a corpus is not necessarily perfect or without any mistakes, just like real language use, which it represents. Therefore, the language study should be corpus-based rather than corpus-bound (Summers, 1996: 266). Besides, one should be aware of merely quantifying studies that only confirm what one already knows (Leech, 1966: 73). Also, it is rather naive to focus merely on the size of a corpus (Leech, 1996: 10). A bigger corpus does not necessarily mean a better corpus (Svartvik, 1992: 10). There are limitations even to a huge corpus, because no matter how big a corpus is, it can never represent the language as a whole (Stubbs, 1995: 50).

Furthermore, the discussion about size cannot be seen independently of the frequency of the linguistic phenomenon that is being studied. In case of rather rare linguistic phenomena a bigger corpus is needed in order to obtain statistically reliable results (Sinclair, 1991: 18-19; Atkins et al., 1992: 5). If the occurrence of a specific linguistic phenomenon is not as high as expected one should not be tempted to conclude that it is a rare phenomenon in language as a whole. Church et al. (1991: 124) warn for this so-called failure-to-find fallacy. As McEnery and Wilson (2001: 30) say: «more common utterances might be excluded simply by chance».

If a corpus analysis does not render enough hits or if the quality of these results is insufficient the corpus analysis can always be supplemented by introspection or elicitation (De Mönnink, 1997: 227; Lauridsen, 1989: 118). Moreover, it is good to diversify and to consider all different methods as complementary. A critical view and some subjective reflection upon corpus results is always a good idea (Svartvik, 1992: 8; Leech, 1991: 74; Sinclair, 1997: 32).

More important than size is perhaps quality. It is an advantage to have a diversified corpus, with material taken from many different sources and speakers, and containing texts that have a high degree of representativeness of a certain sublanguage they aim to represent (*e.g*. Sinclair, 1991: 18; Atkins et al., 1992: 6; McEnery and Wilson, 2001: 29). Many corpora, however, are not very diversified: the language material they contain often originates from similar sources or genres, such as corpora containing exclusively newspaper texts or literary texts. The risk with this kind of corpora is that the language can be biased, *e.g*. influenced by the genre of by the writer's personal style (Sinclair, 1991: 17). If a diversified corpus is not available to the linguist a possible solution can be to match the results from one corpus study to another corpus (Stubbs, 1995: 50). Also Woods et al. (1986) stress the importance of checking hypotheses in many different ways.

To summarize, we could say that the linguist should try to use a corpus that is «so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study». (Atkins et al., 1992: 6). Of course, a corpus can come in many shapes and the choice of the 'right' corpus is also linked to the kind of research envisaged. Finally, it is equally important to relate the results from the corpus analysis to the nature of the corpus.

2. Typology of monolingual and bilingual corpora

This section provides a typology of different corpus types that can be used in linguistic studies with particular emphasis on corpora used in cross-linguistic studies.

There are several types of corpus which serve different purposes to choose from.

For the study of one particular language there is a great variety of monolingual corpora that can be relied on. Diachronic corpora contain texts from different time periods of a language, whereas synchronic corpora contain texts taken from one particular era in a language. A further distinction can be made between written and oral corpora. Amongst monolingual corpora there is also a great variety of corpora reflecting different registers or text genres *e.g*. juridical texts.

For contrastive studies the linguist can make use of bilingual or multilingual corpora, containing language material from two or more different languages.

Contrastive studies can be conducted within a wide range of language research fields, such as language teaching, cross-linguistic studies and lexicographical studies (Lauridsen, 1996: 64). The aim of a contrastive linguistic analysis is not only to highlight the differences between the languages that are being studied, but also to trace details about one specific language that might not have been discovered had this language not been studied in contrast with another language (Johansson, 2000: 4; Aarts, 1998; Aijmer and Altenberg, 1996: 12). Furthermore, a contrastive perspective is especially interesting in the case of two closely related languages, such as Dutch and Swedish, where subtle differences can only be revealed by comparing the languages in detail thereby using a great amount of language data.

There are various kinds of multilingual corpora. There exists some confusion as to the terminology used to denote the different types of multilingual corpora (see Johansson, 1998: 4-5 for this discussion). The terminology used here is the one that seems to have gained the upper hand (*e.g*. Johansson, 1998; Altenberg and Granger, 2002). Following the classification drawn up in Altenberg and Granger (2002: 8) multilingual corpora come in two different shapes, comparable corpora and translation corpora. Following the definition used here comparable corpora do not contain any translations, but texts written in different original languages that are comparable as to genre or specific specialized fields. Conversely, translation corpora do contain translated texts, or rather, both original texts and their translations. Translation corpora can be unidirectional or bidirectional depending on whether the translations go in one direction only or in two directions.

The choice of a particular type of multilingual corpus depends on the aims of the study.

The advantage of comparable corpora is that they contain authentic original language material only, i.e. no translations. This type of corpus is mainly suitable for specific terminology studies (*e.g.* Lauridsen, 1996). The disadvantage may be that only few comparable corpora are well balanced since they are difficult to compile due to the fact that it is often difficult to find relatable pairs that are comparable in function and style in the languages involved (Aijmer and Altenberg, 1996: 13; Granger, 1996: 38; Johansson, 1998: 5; Altenberg and Granger, 2002: 8-9, 13). As a result, it may often be difficult to trace cross-linguistic equivalents not making them a suitable tool to compare language pairs with.

If the linguist is interested in studying translation equivalents, translation corpora obviously offer a much wider spectrum of possibilities (Chesterman, 1998: 90; James, 1980: 67, 178; Johansson, 2000: 4). This line of thought can also be found with Aarts (1998) who states that «[f]ull comparability can only be achieved in translation corpora».

Translation corpora, however, also have a number of disadvantages. One disadvantage that has been discussed by many linguists is that the target text can be influenced by the source text – these signs of influence are denoted by the term *translationese* (Gellerstam, 1996: 53-54) – or that the target text shows clear signs that are typical of translated text (James, 1980: 117-118; Aijmer and Altenberg, 1996: 13; Granger, 1996: 48-49; Johansson, 1998: 5; Aarts, 1998). What is more, translations seldom have a hundred percent mutual correspondence (Altenberg and Granger, 2002: 19). If the outcome of the calculation of the mutual correspondence is low, it can be explained by a number of factors such as diverging polysemy, different pragmatic systems in both languages, lexical gaps in either language or system interchange (Altenberg and Granger, 2002: 19 ff.). Another disadvantage of translation corpora is that they are seldom well balanced due to the fact that certain genres, text types or translators are often overrepresented (Johansson, 1998: 6; Altenberg and Granger, 2002: 9), which has repercussions on the validity of the comparison (Johansson, 2000: 4).

In summary, both types of multilingual corpora should be seen as complementary sources in a cross-linguistic study thereby benefiting from the advantages of either type (Altenberg and Granger, 2002: 9). Furthermore, the corpus results from translation corpora can always be checked against monolingual corpora (Johansson, 1998: 5).

3. The compilation of a Dutch-Swedish parallel corpus

Following these general remarks and echoes from the literature discussed in the previous section, this section will be devoted to a discussion of the compilation of the bilingual Dutch-Swedish corpus.

3.1. Background

The Dutch-Swedish parallel corpus was compiled in collaboration with colleagues from Gothenburg University, as part of the SALT project (SALT stands for *Språkbankens Arkiv för*

Länkade Texter 'The Archive of Aligned Texts of the Swedish Language Bank') coordinated by Språkbanken. The aim of the SALT project is to compile a number of parallel corpora with Swedish as the central language, and for each of the corpora linked to another European language.

3.2. Structure

The different subcorpora contain at least one component with fiction texts (The Dutch-Swedish subcorpus also contains non-fiction texts) consisting of Swedish originals plus their translations into the other language (*e.g*. Russian, Dutch), and originals in the foreign language and their translations into Swedish (cf. ESPC, Aijmer et al., 1996: 79-80 and ENPC, Johansson, 1998: 7-8). As such, the corpora are bidirectional. The SALT corpora are actually a combination of comparable and translation corpora, and are denoted as parallel corpora. The bidirectional structure in the SALT corpora has the advantage of making possible various kinds of analyses. It is not only possible to compare original texts and their translations, but also to compare texts in both of the original languages, or even the originals in one language and the translated texts in that same language as illustrated by the structure in the SALT Dut-Swe corpus in Fig. 1.

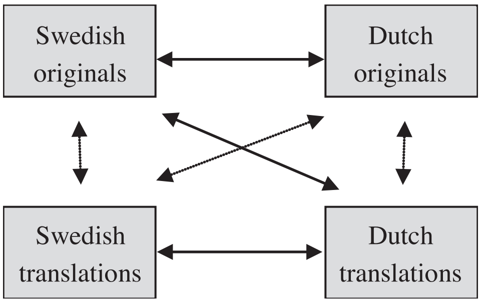

Figure 1: The structure in the Dutch-Swedish corpus

The SALT corpora contain entire works instead of smaller text parts. A drawback is that the language in the corpus as a whole is certainly not as diversified as if the corpus had consisted of smaller text parts since only a limited number of writer and translation styles are represented. An advantage is that parts of the corpus, *e.g*. one particular novel, can be used for smaller linguistic studies. The homogeneity of the language and style throughout a particular work can also be seen as an advantage (Sinclair, 1991: 19).

In the following subsections focus will be on the compilation of the Dutch-Swedish parallel corpus. The work is done in accordance with Språkbanken's guidelines and following certain templates (*e.g*. for the annotation of the material). In accordance with the guidelines different types of texts from the second half of the 20th century are represented in the corpus. The fiction component contains twenty-six novels in total, that is thirteen originals (seven Swedish and six Dutch) and their translations, amongst these for instance Bergman, I., *Laterna Magica*, Claus, H., *De Geruchten*, Krabbé, T., *Het Gouden Ei*. The non-fiction component includes a diversified language sample from a number of different areas such as psychology, scientific history, physics, geography and social history. This component contains twelve works (four Swedish and two Dutch originals plus their translations) and are, amongst others, Cullberg, J., *Dynamisk psykiatri i teori och praxis*, Dekker, R. and van de Pol, L., *Vrouwen in mannenkleren*.

3.3. Copyright, scanning and correction

Språkbanken took care of the copyright matters in asking the different publishers for permission to use the material for the purpose of linguistic research (on copyright matters see also Atkins et al., 1992: 4 and Sinclair, 1991: 15). Once permission was obtained, the material was scanned. The digitalized image files (*.tif*-files) were subsequently loaded into an OCR-program (Omnipage Pro 10) in order to transfer the image files into text files (*.opd*-file). The next stage involving the correction of the scanned text files turned out to be very time-consuming. Many mistakes had occurred when the image files had been transferred to text files (see also Mason, 2000: 31-32 on this issue). Typical mistakes were small spots on the page that had been interpreted as punctuation marks or as accents *e.g*. the Swedish word *f.ar* (with a small spot on the page) denoting *far* 'father' was misinterpreted as *fàr* (with an accent). Other common misinterpretations were letter combinations such as *rn* that had been misinterpreted as *m* (*e.g*. the Swedish word *barn* 'child' was misinterpreted as *bam*, a non-existing word). Apart from the correction of apparent mistakes, split words were written as a single word in order to reduce the risk of misinterpretations during the alignment phase.

3.4. Annotation

The correction phase was followed by the annotation phase which was carried out according to predefined schemes designed by Språkbanken concerning both text-external and text-internal tags. The texts were saved in ASCII-format so that tags could be added easily. The annotation in the Dutch-Swedish corpus is actually a simplified mark-up format that is turned into a mark-up format following the xml-version of CES (Corpus Encoding Standard, also called XCES) at a later stage. The text-internal SGML tags the texts were annotated with are for instance chapter tags (e.g. <chapter><title>JULIA</title></chapter>), page number tags (*e.g*. <p15>) and tags indicating a few typographical peculiarities such as bold style or italics (e.g. <it>det</it>). All tags were inserted manually except for paragraph tags which were inserted automatically prior to the alignment process.

3.5. Alignment

Following the tagging phase, the texts were aligned at sentence level by a computer linguist at Språkbanken using the Unix-based alignment program Vanilla Aligner [1]. Subsequently, the automatic alignment was checked manually, with special attention for cases where the outcome revealed unbalanced instances, that is all instances not reflecting a 1-1-relation (i.e. one sentence in L1 aligned to one sentence in L2). In some cases a 1-2-relation indicated a wrong alignment or a case where one original sentence had been translated by two sentences. A 1-0-link could be a wrong alignment or a case where the original sentence simply had been omitted by the translator.

The SALT-corpora can be searched by means of the bilingual concordancer ParaConc (Barlow, 1999). The aligned files are uploaded into the program and can be viewed and searched on the screen simultaneously. When searching something in one original language the program shows the corresponding equivalent in the target language. An online concordance is currently being developed by Språkbanken.

[1] See e.g. http://spraakdata.gu.se/lb/tools.html. This alignment program written by Danielsson and Ridings at the Department for Swedish at Gothenburg University is actually an implementation of the alignment algorithm presented by Gale and Church (1993). The basic mechanism of the alignment program is based on sentence boundaries. In preparation of the alignment a program is run to insert sentence boundaries into the texts.

4. Conclusion

This paper has dealt with corpus linguistics in general and with multilingual corpora in particular. Following a discussion of the pros and cons of corpus use it has presented a typology of corpora with particular focus on corpora that can be used in cross-linguistic studies. As a case in point, the paper has discussed the compilation of the Dutch-Swedish parallel corpus which can be used for a wide variety of purposes.

References

Aarts J. (1998). Introduction. In Johansson, S. and Oksefjell, S., editors, *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Amsterdam: Rodopi.

Aijmer K. and Altenberg B. (1996). Introduction. In Aijmer, K., Altenberg, B. and Johansson, M., editors, *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies in Lund*, 4-5 March. Lund: Lund University Press, pp. 11-16.

Aijmer K., Altenberg B. and Johansson M. (1996). Text-based contrastive studies in English. Presentation of a project. In Aijmer, K., Altenberg, B. and Johansson, M., editors, *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies in Lund*, 4-5 March. Lund: Lund University Press, pp. 73-85.

Altenberg B. and Granger S. (2002). Recent trends in cross-linguistic lexical studies. In Altenberg, B. and Granger, S., editors, *Lexis in Contrast*, Amsterdam: John Benjamins, pp. 1-48.

Atkins S., Clear J. and Ostler N. (1992). Corpus design criteria. *Literary and computing*, 7: 1-16.

Barlow M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics*, 4 (1): 319-327.

Chesterman, A. (1998). *Contrastive Functional Analysis*. Amsterdam-Philadelphia: John Benjamins.

Church K., Gale W., Hanks P., Hindle D., Bell Laboratories and Oxford University Press (1991). Using statistics in lexical analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale (New Jersey): Erlbaum, pp. 115-164.

Cook V.J. (1988). *Chomsky's Universal Grammar: An Introduction*. Oxford: Blackwell.

De Mönnink I. (1997). Using corpus and experimental data: a multi-method approach. In Ljung, M., editor, *Studies in English Corpus Linguistics. Papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*, Amsterdam: Rodopi, pp. 227-244.

Gale W.A. and Kenneth W.C. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19 (1): 75-101.

Gellerstam M. (1996). Translations as a source for cross-linguistic studies. In Aijmer, K., Altenberg, B. and Johansson M., editors, *Languages in Contrast*, Lund: Lund University Press, pp. 53-62.

Granger S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer, K., Altenberg, B. and Johansson, M., editors, *Languages in Contrast Papers from a symposium on text-based cross-linguistic studies in Lund*, 4-5 March, Lund: Lund University Press, pp. 37-51.

James C. (1980). *Contrastive analysis*. London: Longman.

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

Johansson S. (2000). Contrastive Linguistics and Corpora. *SPRIK Reports* http://www.hf.uio.no/german/sprik/.

Lauridsen K.M. (1989). Tekstkorpora. Ny forskningsaktivitet ved Handelshøjskolen. In *Handelshøjskolen 50 år. Festskrift udgivet i anledning af Handelshøjskolens 50-års jubilæum* 31. august, Aarhus: the Aarhus School of Business, pp. 118-125.

Lauridsen K.M. (1996). Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In Aijmer, K., Altenberg, B. and Johansson, M., editors, *Languages in Contrast*, Lund: Lund University Press, pp. 63-71.

Leech G. (1966). *English in Advertising: A Linguistic Study of Advertising in Great Britain.* London: Longman.

Leech G. (1991). Corpora. In Malmkjaer, K., editor, *The linguistics encyclopedia*, London: Routledge, pp. 73-80.

Leech G. (1992). Corpora and theories of linguistic performance. In Svartvik, J., editor, *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, ["The State of the Art in Corpus Linguistics"]*, Stockholm, Sweden, August 4-8, NY: Mouton de Gruyter, pp. 105-122.

Leech G. (1996). The state of the art in corpus linguistics. In Aijmer, K. and Altenberg, B., editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman: pp. 8-29.

Mason, O. (2000). *Programming for corpus linguistics. How to do text analysis with Java*. Edinburgh: Edinburgh University Press.

McEnery T. and Wilson A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi V.B.Y. (1998). *Computer corpus lexicography.* Edinburgh: Edinburgh University Press.

Sinclair J. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair J. (1997). Corpus Evidence in Language Description. In Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G., editors, *Teaching and Language Corpora*, London-New York: Longman, pp. 27-39.

Stubbs M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of language*, 2 (1): 23-55.

Stubbs M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell.

Summers D. (1996). Computer lexicography: the importance of representativeness in relation to frequency. In Thomas, J. and Short, M., editors, *Using corpora for language research: studies in the honour of Geoffrey Leech*, London: Longman, pp. 260-266.

Svartvik J. (1992). Corpus linguistics comes of age. In Svartvik, J., editor, *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82 ["The State of the Art in Corpus Linguistics"]*, Stockholm, Sweden, August 4-8, New York: Mouton de Gruyter, pp. 7-13.

Tognini-Bonelli E. (2001). *Corpus linguistics at work*. Amsterdam-Philadelphia: John Benjamins.

Woods A., Fletcher P. and Hughes A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.