Bilingual Corpora and contrastive Language Studies. A corpus-based study of causative constructions in a Dutch-Swedish contrastive perspective.

Gudrun Rawoens[1]

1. Introduction

As shown in (Aijmer and Altenberg, 1996: 12, Aarts, 1998: Introduction, Johansson, 2000: 4), the value of a contrastive linguistic analysis not only resides in the fact that allowance is made for cross-linguistic comparison, but also in that it allows us to trace details about one specific language that might not have been discovered without the comparative angle in the study of the research object.

Furthermore, a contrastive perspective is especially interesting in the case of two closely related languages, such as Swedish and Dutch, where subtle differences can only be revealed by comparing the languages in detail thereby using a great amount of language data involving both languages.

2. Corpora in cross-linguistic Research

It is a well-known fact that the use of empirical data for linguistic research has experienced a considerable increase since large text corpora became accessible (e.g. McEnery and Wilson, 2001). Corpora can be used in a great number of research fields and in a number of different ways. They 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. The former approach is commonly referred to as the corpus-driven approach, whereas the latter is known as the corpus-based method (Ooi, 1998: 52).

The pros and cons of the use of corpora in linguistics have been widely discussed. One of the apparent advantages is that corpora provide the linguist with a great amount of authentic language material, which in its turn allows 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. On the other hand, many have warned for a far too positive approach of corpus use. The linguist should for instance never blindly trust the corpus. The language study should be corpus-based, not corpus-bound (Summers, 1996: 266). The ideal approach in solid linguistic research is really to combine a number of approaches along with the corpus-based method, such as introspection and elicitation.

The use of multilingual corpora in contrastive linguistics has gained momentum over the past decades and their value for various types of semantic and syntactic linguistic studies as well as a great number of terminological studies has been repeatedly spelled out. As noted by Altenberg and Granger (2002: 14) the apparent advantages of multilingual corpora in contrastive linguistics are that

[1] Department of Nordic Studies, Ghent University, Belgium
e-mail : Gudrun.Rawoens@Ugent.be

- they give new insights into the languages compared – insights that are likely to be unnoticed in studies of monolingual corpora;
- they can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as of universal features;
- they illuminate differences between source texts and translations, and between native and non-native texts;
- they can be used for a number of practical applications, e.g. in lexicography, language teaching and translation.

Various types of multilingual corpora can be distinguished. There is some inconsistency in the terminology used (see e.g. Johansson, 1998: 4–5 footnote for this discussion). In the following typology drawn up by Altenberg and Granger (2002: 8), a distinction can be made between translation corpora and comparable corpora (Figure 1). Translation corpora consist of texts in one original language and their respective translations in one or more other languages. They can be either unidirectional or bidirectional, depending on whether the translations go in one direction only (L1 > L2) or in two directions (L1 > L2 and L2 > L1). Comparable or parallel corpora do not contain any translations, but texts written in two or more original languages that are comparable as to genre or specific specialized fields.

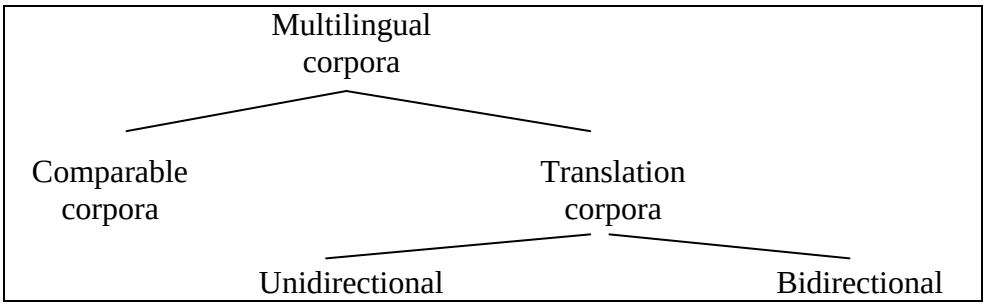

Figure 1 A typology of multilingual corpora (Altenberg and Granger, 2002: 7)

As such, corpora can come in many shapes and each corpus type has its own advantages and disadvantages. Also, the choice of the 'right' corpus is linked to the kind of research envisaged. Moreover, it is also important to relate the results from the corpus analysis to the nature of the corpus used.

The advantage of comparable corpora is that they only contain authentic original language material, i.e. no translations. This corpus type is most suitable for specific terminology studies (Lauridsen, 1996). However, such corpora are difficult to compile and, as a result, not many of them are well balanced. In many cases, it is difficult to find relatable pairs that are comparable in function and style in the languages involved (Aijmer and Altenberg, 1996: 13, Johansson, 1998: 5, Altenberg and Granger, 2002: 8–9, 13). In other words, it is very difficult, if not impossible, to trace cross-linguistic equivalents from a comparable corpus which explains why this type of corpus is not 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). Translations have to do with similarities or equivalents, with how specific elements from one language are being

transferred to another language without loss of meaning. This is what Chesterman (1998: 17) calls ”the translation identity assumption”. Chesterman (1998: 18) also stresses that ”[…] translation equivalence […] is always equivalence-in-context”. Studying translations is studying parole rather than langue. In a translation process the translator does not simply try to find an equivalent linguistic utterance in the other language, but also takes into account context related aspects and other pragmatics related aspects. The linguist in his or her turn should be careful with deriving conclusions from the translation material. Similarities or equivalents in two languages do not necessarily imply identical usage.

Translation corpora, however, also have a number of disadvantages. One well-known disadvantage for cross-linguistic studies 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). The target text can also show clear signs that are typical of translated text (James, 1980: 117–118, Aijmer and Altenberg, 1996: 13, Aarts, 1998: Introduction, Johansson, 1998: 5). Altenberg and Granger (2002:19) also point out that “translation equivalents seldom have 100% correspondence in translation corpora”. What is more, ”translation equivalents in two languages seldom have the same distribution” (Altenberg and Granger 2002:18). A way for trying to find out if there is a high or low degree of correspondence between the compared languages is to calculate the so-called mutual correspondence in the bilingual translation corpus. A low degree of mutual correspondence 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: 19ff). Another disadvantage with translation corpora is that they are seldom well balanced since certain genres or text types are often overrepresented (Johansson, 1998: 6, Altenberg and Granger, 2002: 9). Therefore, the aim should be to compile multilingual corpora where various text types, authors and translators are represented in order to raise the ”the validity and reliability of the comparison” (Johansson, 2000: 4). A limitation of translation corpora is also that they by definition only contain written language (Johansson, 1998: 6; Altenberg and Granger, 2002: 9).

Altenberg and Granger (2002: 9) conclude by saying that both types of multilingual corpora should be seen as complementary sources of cross-linguistic data. In other words, one could use both translation and parallel corpora simultaneously thereby benefiting from the advantages of both types.

3. The SALT dut-swe Corpus

The material used for this study is taken from the SALT dut-swe corpus. This Dutch-Swedish corpus has been compiled by colleagues of Gothenburg University and myself within the SALT-project. SALT stands for Språkbankens Arkiv för Länkade Texter ’Språkbanken’s Archive for Aligned Texts’ and is a project initiated by Språkbanken ‘the Bank of Swedish’ and financed by the Faculty of Arts and Humanities at Gothenburg University and Riksbankens Jubileumfond ‘The Bank of Sweden Tercentenary Foundation’. The Bank of Swedish can rely on a long tradition and expertise as for the compilation of monolingual Swedish corpora dating back to the sixties when the compilation of the Swedish corpus Press 65 containing one million words was initiated by Sture Allén. In 1975 Språkbanken was established as a national center aiming at collecting Swedish corpus databases. To date, the Språkbanken corpus databases are still growing, already amounting to more than one

hundred million words. The SALT project was set up in 2001 and originates in the need to create a number of bilingual corpora with Swedish as the central language and to make them available to researchers within linguistics, language pedagogy and language technology applications such as CAT.

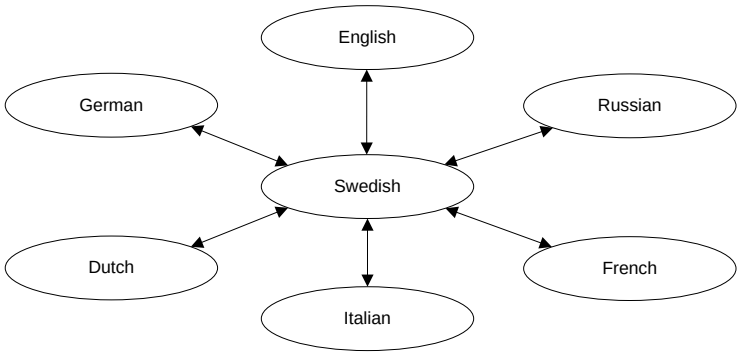

Figure 2 The SALT-corpora

The SALT-corpora contain Swedish original texts and their translations into the other language (e.g. Russian or Dutch), plus original texts in the foreign language and their Swedish translations (Figure 2). As such, these corpora are actually a combination of translation and comparable corpora, but are usually being denoted as parallel corpora.

A variety of text styles and genres is represented in the Dutch-Swedish corpus. In contrast with the other SALT corpora that are limited to fiction material, even non-fiction texts are included in the Dutch-Swedish corpus in order to diversify the language even more. Furthermore, twenty different authors and thirteen different translators are represented in the corpus. The corpus contains three million words in total and the texts are aligned at sentence level. The bi-derectional structure of the corpus allows for various kinds of analyses. Not only is it possible to compare original texts and their translations, it also allows the comparison of texts in both of the original languages, or even the originals in one language and the translated texts in that same language (Figure 3).

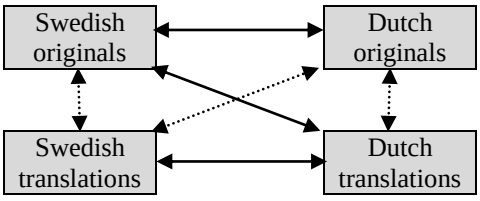

Figure 3 Structure of the SALT dut-swe corpus

Furthermore, this bilingual corpus can be used in a whole range of contrastive lexical, syntactic or pragmatic studies. It can also be used as an educational tool or for purely quantitative purposes.

4. Causative Constructions in Dutch and Swedish

The last section of this paper is dedicated to the presentation of a number of results from an analysis of analytical causative constructions in a Dutch-Swedish contrastive perspective based on the SALT dut-swe corpus. The study presented here is part of a more extensive investigation on causative constructions in Dutch and Swedish (Rawoens 2007), where a bi-directional approach is applied. In this paper, only one dimension, viz. the study of the translations departing from the originals, will be highlighted.

Analytical causative constructions consist of a causal predicate and an effected predicate (e.g. Comrie, 1985, Verhagen and Kemmer 1997). The patterns of these constructions can vary according to the type of causal predicate and the type of complement. In this paper, analytical causative constructions where the causal predicate is a grammaticalized verb followed by an infinitival complement are considered.

In Dutch the verbs *doen* (cognate with 'do') and *laten* (cognate with 'let') are the central causative verbs in these constructions, followed by a bare infinitive in the complement. The syntactic patterns are as follows illustrated with the examples taken from the parallel corpus:

- *doen* + NP + V_{inf}
  Er klonk een aarzelend applaus dat hem verlegen naar zijn stoel terug *deed* lopen. (WolN1)
  'The sound of hesitant applause embarrassed him and made him go back to his chair.'

- *laten* (+ NP) + V_{inf}
  Bij 8-7 voor Frankrijk *liet* ze hen van speelhelft wisselen. (KraN1)
  'At 8-7 to France she let them switch play sides.'

An important semantic difference between these two verbs is that *doen* is used to express relatively simple and direct causal relations. *Laten* on the other hand expresses rather indirect causal relations (e.g. Verhagen and Kemmer, 1997: 70). The subject referent of *doen* can either be human or non-human, the subject referent of *laten* is usually human. In some cases the differences between these two verbs are very clear, in other cases the differences are rather subtle and depend on style or register. In just a few cases the two verbs can even be freely exchanged.

The immediate equivalents of *doen* and *laten* in Swedish are the causative verbs *få* 'get' and *låta* 'let', respectively. The verbs *komma* 'come', *ha* 'have' and *förmå* 'induce, persuade' can also appear in this kind of construction. The syntactic patterns of these Swedish constructions are as follows:

- *få* + NP + *(till)* + *att* + V_{inf}
  Jag visste […] att jag *fick* mänskor att göra som jag ville, […]. (BerS1)
  'I knew […] that I made people do as I wanted them to do […]'

- *låta* + NP + V_{inf}
  Jag *lät* skriva ut materialet. (BerS1)
  'I had the material typed out.'

- *komma* + NP + *(till)* + *att* + V_{inf}
  Där fanns särskilt en bild som *kom* mig att yla av sorg. (BerS1)
  'There was one picture in particular that made my cry with grief.'

- *ha* + NP + *(till)* + *att* + V_{inf}
  Hon *hade* honom att bygga ett nytt garage. (SAG)
  'She made him build a new garage.'

- *förmå* + NP + *(till)* + *att* + V_{inf}
  Jag *förmådde* honom att söka psykiatrisk vård – ingenting hjälpte. (BerS1)
  'I made him look for psychiatric help – nothing could help.'

The differences between the Swedish causatives are of another kind than the differences between the Dutch causatives. The verbs *få* and *låta* can never be exchanged in the same way as *doen* and *laten*. On the other hand the verbs *komma*, *ha* and *förmå* can occur as an alternative to the verb *få*. However, they are much less frequent than *få* due to a number of semantic and stylistic restrictions (SAG, 1999, Rawoens, 2007): they can express different degrees of coercion and modality and they are often more formal. They also typically combine with different kinds of subject referents. The subject referent of *få* is either human or non-human. *Komma* takes non-human subject referents, whereas the subject referents of *låta, ha* and *förmå* are usually human.

In the following I will not go into the semantic or stylistic differences between the constructions mentioned. The main aim of the contrastive study presented here is to analyse the Swedish translations of the Dutch analytical causative constructions and vice versa starting from the following translation patterns represented by three main categories CAT1, CAT2 and CAT3.

CAT1 contains so-called one-to-one translations, i.e. instances where an analytical causative construction with an infinitival complement in the source language is matched by the same type of analytical causative construction in the target language.

CAT2 contains instances where an analytical causative construction with an infinitival complement in the source language is translated either by another type of analytical causative construction than the one in CAT1, e.g. with a finite complement (CAT2ana), or by a synthetic lexical verb (CAT2syn).

CAT3 comprises instances where an analytical causative construction is either translated by a non-verbal causal construction (CAT3cau) or where it does not get a translation equivalent (CAT3Ø).

In the source text material there are 797 sentences containing analytical causative constructions, 374 sentences in the Dutch material and 423 in the Swedish material. As indicated by the relative frequencies in Table 1, analytical causative constructions occur much more frequently in the Dutch material than in the Swedish material. One explanation is the high frequency of the Dutch constructions with *laten* (301 occurrences) which outnumber the constructions with *doen* (73). In the Swedish material causative constructions with *få* appear most frequently (215 occurrences) followed by constructions with *låta* (165). The verbs *komma* (31), *förmå* (10) and *ha* (2) are much less common. The proportions in the Dutch and Swedish material are

similar to those found in other monolingual analyses of Dutch and Swedish analytical causative constructions, viz. Verhagen and Kemmer (1997) and Rawoens (2007).

	Absolute frequencies	Relative frequencies (per 100,000 words)
Swedish originals (1,026,003 words)	423	41.23
Dutch originals (429,410 words)	374	87.10
Total (1,455,413 words)	797	54.76

Table 1 Absolute and relative frequencies of the analytical causative constructions in the Swedish and Dutch original texts

The results of the analysis of the Swedish (S) translations of the Dutch (D) analytical causative constructions with *doen* and *laten* as represented in Table 2 reveal that a Dutch analytical causative construction with *doen* or *laten* corresponds to a Swedish analytical causative construction with *få, komma, ha, förmå* or *låta* in forty-four percent of the cases (CAT1). A similar proportion of the Swedish constructions is translated as an alternative verbal causative construction (CAT2), whereas twelve percent of the total number of translations are instances where a non-verbal construction is used or where the translation is absent (CAT3).

CAT1	CAT2		CAT3		Total D > S
	CAT2ana	CAT2syn	CAT3cau	CAT3Ø	
165	5	159	3	42	
165	164		45		374
44.1%	43.9%		12%		100%

Table 2 The D > S translations

Examples (1) and (2) illustrate the "default" translation patterns belonging to CAT1: in the first sentence a Dutch analytical causative construction with *doen* corresponds to a Swedish analytical causative construction with *få* and in the second sentence a construction with *laten* is equivalent to an analytical construction with *låta*:

(1) Ik heb iets bij me dat uw hartje sneller zal *doen* kloppen. (ClaN1)
Jag har med mig något som ska få ditt hjärta att slå snabbare. (ClaS2)
'I have something with me that will make your hart beat faster.'

(2) Om uit de voeten te kunnen met stok of looprek heeft zij alle overtollige stoelen en prullen door Oscar naar boven *laten* brengen. (EnqN1)

För att hon ska kunna ta sig fram med käpp eller rollator har hon låtit Oscar bära upp alla överflödiga stolar och prylar på övervåningen. (EnqS2)
'In order to be able to get about with a stick or a walking frame he had Oscar bring all his superfluous chairs and rubbish upstairs.'

In CAT2 there is a strikingly uneven distribution between the subcategories CAT2ana and CAT2syn (Table 2). The former contains only five occurrences, whereas the latter contains 159 instances, which stand for almost forty-three percent of the total number of the D > S translations.

Interestingly, the majority of the translations in subcategory CAT2syn are translations of constructions with the verb *laten*, as illustrated in example (3) where the original Dutch construction *laat…opstellen* 'let…draw up' is translated by the synthetic causative *sätter upp* 'draw up':

(3) Ik ben blij als jullie er komen wonen, ik *laat* een huurcontract opstellen zodat je zekerheid hebt. (EnqN1)
Jag blir glad om ni flyttar in, jag sätter upp ett hyreskontrakt så att du har någon säkerhet. (EnqS2)
'I will be glad if you move in, I will draw up a tenancy agreement so that you have some certainty.'

One explanation for the high number of occurrences in this subcategory seems to be that a synthetic causative in many cases expresses the causative meaning in a more tangible way than an analytical causative construction. This goes especially for constructions with *laten*, which could be related to the fact that *laten* can also express permission. By translating a construction with *laten* by means of a synthetic causative, the translator accentuates its causative meaning. Besides, more literal translations by means of an analytical causative construction, though possible, can in many cases be considered as rather rigid in Swedish. This is for instance true for the Dutch construction *laten zien* 'let see' in example (4) which could indeed be translated into Swedish by *låta se* 'let see'. However, this translation would either be considered as unnatural or be misinterpreted as conveying a permissive meaning. Instead the translator has chosen an appropriate lexical verb, viz. *visa* 'show' which is the most natural expression in the target language.

(4) Hij *laat* haar het kasteel van binnen zien, zij kijkt haar ogen uit. (ClaN1)
Han visar henne slottet från insidan, hon tror inte sina ögon. (ClaS2)
'He shows her the castle from the inside, she cannot believe her own eyes.'

As for the subcategories in CAT3, CAT3cau contains only three instances, whereas CAT3Ø contains forty-two instances. Among these, nine are instances where a translation is absent. The remaining instances are cases where the syntactic and semantic relations have been altered in the translation. This kind of alteration is known as a syntactic transformation (e.g. Blum-Kulka, 1989) and is the result of a translation strategy rather than a language-typical phenomenon. A syntactic transformation is illustrated in example (5) where the direct object in the second main clause in the Dutch sentence *het presenteerblad* 'the tray' becomes the subject *brickan* 'the tray' in the corresponding main clause in the Swedish translation. The

translation of the Dutch second main clause reads literally as '… and then the tray with the empty glasses was left to drift about':

(5) Voordat de wijn kon verwateren dronken we hem op en *lieten* toen het presenteerblad met de lege glazen ronddobberen. (WolN1)
Innan vinet hunnit späs ut drack vi upp det, och sen fick brickan med de tomma glasen flyta omkring. (WolS2)
'Before the wine could be diluted, we drank it, and then made the tray with the empty glasses drift about.'

At first glance, the Dutch translations of the Swedish analytical causative constructions display a similar distribution pattern in comparison to the D > S translations (Table 3). However, when the subcategories are considered as well, some more striking differences turn up.

CAT1	CAT2		CAT3		Total S > D
	CAT2ana	CAT2syn	CAT3cau	CAT3Ø	
	60	79	43	41	
200	139		84		423
47.2%	32.9%		19,9%		100%

Table 3 The S > D translations

Within CAT1 there is a notable majority of instances where a Swedish analytical causative construction with *låta* is translated with a Dutch analytical causative construction with *laten*, as in example (6). Interestingly, this subgroup represents almost thirty percent of the total number of translations.

(6) Jag *lät* skriva ut materialet. (BerS1)
Ik liet het materiaal uittikken. (BerN2)
'I had the material typed out.'

Within CAT2 there is a fairly even internal distribution between the two subcategories: CAT2ana contains sixty instances and CAT2syn contains seventy-nine. In other words, the number of cases where an analytical causative construction corresponds to a synthetic causative in the target language is remarkably low compared to the same category in the D > S translations.

CAT3 represents about a fifth of all S > D translations, which makes this category fairly more extensive than the corresponding category in the D > S translations. It further appears that more than half of the occurrences in CAT3 are classified in CAT3cau as in example (7), where the Dutch translation features the causal preposition *van* 'from'. The literal English translation of the Dutch sentence is 'He became nauseous from the light and sound'.

(7) Ljus och ljud *fick* honom att må illa. (EkmS1)
Van licht en geluid werd hij misselijk. (EkmN2)
'Sounds and light made him feel nauseous.'

What is more, the subcategory CAT3cau represents about a tenth of all S > D translations, which is a lot more than its proportion in the D > S material where it represents less than one percent of the total number of translations. One explanation for the high frequency of non-verbal causal constructions in the Dutch translation texts could be that causal conjunctions and adverbials are considered to be the most common ways of expressing causality in Dutch in general, as argued by Degand (2001).

5. Conclusion

After having discussed a number of reasons for choosing a contrastive and corpus-based approach in linguistic research, this paper has presented the results of a contrastive analysis of analytical causative constructions in Dutch and Swedish based on the SALT dut-swe corpus.

This study has shown that the relation between the analytical causative constructions and their translations is asymmetrical. Whereas synthetic causatives appear to be more common in Swedish, Dutch uses more analytical causative constructions – especially with *laten* – and more non-verbal causal constructions such as prepositions and adverbials. The analysis has also revealed that the translators often use translations that imply a high degree of explicitness and disambiguation or that they adhere to language-specific features. These tendencies are confirmed by the results of the whole corpus analysis where the other dimensions in the corpus material were taken into account as well (Rawoens, 2007).

References

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

Aijmer, K. and B. Altenberg (1996) Introduction, in: K. Aijmer, B. Altenberg and M. Johansson (eds), Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994. Lund: Lund University Press, pp. 11–16.

Altenberg, B. and S. Granger (2002) Recent trends in cross-linguistic lexical studies, in: B. Altenberg and S. Granger (eds), Lexis in Contrast. Corpus-based approaches. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 1–48.

Blum-Kulka, S. (1989) Shifts of cohesion and coherence in translation, in: J. House and S. Blum-Kulka (eds), Interlingual and Intercultural communication. Tübingen: Günter Narr Verlag, pp. 17–35.

Chesterman, A. (1998) Contrastive functional analysis. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Comrie, B. (1985) Causative verb formation and other verb-deriving morphology, in: Shopen, T. (ed.), Language typology and syntactic description. Cambridge: Cambridge University Press, pp. 309–348.

Degand, L. (2001) Form and Function of Causation. A Theoretical and Empirical Investigation of Causal Constructions in Dutch. Leuven: Peeters.

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

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

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

Johansson, S. (2000) Contrastive Linguistics and Corpora no.3, *SPRIKreports*. Available on-line from http://www.hf.uio.no/german/sprik/ (accessed: 22 June 2007).

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

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

Ooi, V. B.Y. (1998) Computer Corpus Lexicography. Edinburgh: Edinburgh University Press.

Rawoens, G. (2007) Kausativa verbkonstruktioner i svenskan och nederländskan. En korpusbaserad syntaktisk-semantisk undersökning. PhD dissertation, Ghent University.

SAG I–IV = Teleman, U., S. Hellberg and E. Andersson (1999) Svenska Akademiens grammatik. Stockholm: Svenska Akademien.

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

Verhagen, A. and S. Kemmer (1997) ‘Interaction and causation: Causative constructions in modern standard Dutch’. *Journal of Pragmatics* 27. 61–82.