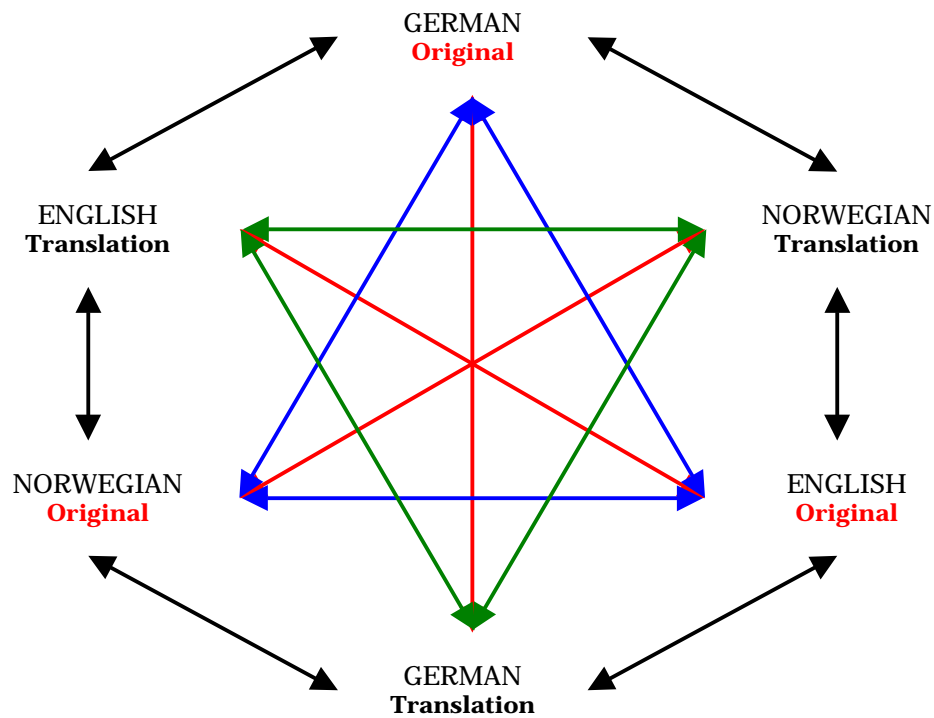


SPRIKreports

Reports from the project *Languages in Contrast* (Språk i kontrast)
<http://www.hf.uio.no/german/sprik>

no. 3, October 2000



Contrastive Linguistics and Corpora

Stig Johansson
Department of British and American Studies, University of Oslo

Contrastive linguistics and corpora¹

(draft, October 1999; please do not quote)

Stig Johansson
University of Oslo

1 Aim

In this paper I will give a brief introduction to contrastive linguistics, with the emphasis on more recent developments. The main aim is to consider the role of computer corpora in contrastive studies.

2 What is contrastive linguistics?

Contrastive linguistics is the systematic comparison of two or more languages, with the aim of describing their similarities and differences. The objective of the comparison may vary:

Language comparison is of great interest in a theoretical as well as an applied perspective. It reveals what is general and what is language specific and is therefore important both for the understanding of language in general and for the study of the individual languages compared. (Johansson and Hofland 1994: 25)

Contrastive linguistics is thus not a unified field of study. The focus may be on general or on language specific features. The study may be theoretical, without any immediate application, or it may be applied, i.e. carried out for a specific purpose.²

The term 'contrastive linguistics', or 'contrastive analysis',³ is especially associated with applied contrastive studies advocated as a means of predicting and/or explaining difficulties of second language learners with a particular mother tongue in learning a particular target language. In the Preface to his well-known book, Lado (1957) expresses the rationale of the approach as follows:

The plan of the book rests on the assumption that we can predict and describe the patterns which will cause difficulty in learning and those that will not cause difficulty.

It was thought that a comparison on different levels (phonology, morphology, syntax, lexis, culture) would identify points of difference/difficulty and provide results that would be important in language teaching:

The most efficient materials are those that are based upon a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner. (Fries 1945: 9)

The high hopes raised by applied contrastive linguistics were dashed. There are a number of problems with the approach, in particular the problem that learning cannot be understood by a purely linguistic study.⁴ So those who were concerned with language learning instead turned to the new disciplines error analysis, performance analysis or interlanguage studies, and contrastive analysis was rejected by many as an applied discipline.⁵

In spite of the criticism of applied contrastive linguistics, contrastive studies were continued, and their scope was broadened. This is the next point I will address.

3 New directions

Although Lado (1957) included a comparison of cultures, early contrastive studies focused on what has been described as microlinguistic contrastive analysis (James 1980: 61ff.): phonology, grammar, lexis. Examples of research questions:

- What are the consonant phonemes in languages X and Y? How do they differ in inventory, realization, and distribution?
- What is the tense system of languages X and Y?
- What are the verbs of saying in languages X and Y?

With the broadening of linguistic studies in general in the 1970s and 1980s, contrastive studies became increasingly concerned with macrolinguistic contrastive analysis (James 1980: 98ff.): text linguistics, discourse analysis. Examples of research questions:

- How is cohesion expressed in languages X and Y?
- How are the speech acts of apologizing and requesting expressed in languages X and Y?
- How are conversations opened and closed in languages X and Y?

When questions of this kind are raised, it becomes increasingly important to base the contrastive study on texts. This brings me to my next point.

4 The role of corpora

In the course of the last couple of decades we have seen a breakthrough in the use of computer corpora in linguistic research, i.e. collections of texts in machine-readable form. Computer corpora are used for a wide range of studies in grammar, lexis, discourse analysis, language variation, etc. They are used in both synchronic and diachronic studies - and increasingly also in cross-linguistic research (see the survey of contrastive projects in Danielsson and Ridings 1996).

Raphael Salkie, editor of the new journal *Languages in Contrast* (published by John Benjamins) goes as far as to say:

Parallel corpora [i.e. multilingual corpora] are a valuable source of data; indeed, they have been a principal reason for the revival of contrastive linguistics that has taken place in the 1990s. (Salkie 1999)

In the rest of this paper I will focus on the role of corpora in contrastive linguistics. As a starting-point, I will use the possibilities offered by bilingual corpora as listed by Aijmer and Altenberg (1996: 12):

- 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 understanding 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, eg in lexicography, language teaching, and translation.

I will take up each of these points in turn, in the order in which they are listed above. For ease of reference, I will refer to bilingual and multilingual corpora as multilingual corpora and to the paper by Aijmer and Altenberg (1996) simply as Aijmer & Altenberg.

5 Analytical comparison

Comparison is a good way of highlighting the characteristics of the things compared. This applies to language comparison as well as more generally, and it is notable that this is the first point in Aijmer & Altenberg's list. Vilém Mathesius, founder of the Linguistic Circle of Prague, spoke about analytical comparison or linguistic characterology as a way of determining the characteristics of each language and gaining a deeper insight into their specific features (Mathesius 1975). He used it in his comparison of the word order of English and Czech, and the study has been followed up by Jan Firbas in particular. In the opening chapter of his *Functional sentence perspective in written and spoken communication* (1992: 3ff.) Firbas compares an original text in French with its translation into English, German, and Czech, and he uses the same sort of comparison later in the book. Firbas says:

The contrastive method proves to be a useful heuristic tool capable of throwing valuable light on the characteristic features of the languages contrasted; ... (Firbas 1992: 13).

There is no difference in principle between the contrastive method of Firbas and the way we use multilingual corpora, except that the study can be extended by the use of computational techniques. As an example, consider Jarle Ebeling's study of presentative constructions in English and Norwegian, based on the English-Norwegian Parallel Corpus (Ebeling, in progress). Ebeling studies three constructions which are found in both languages, termed full presentatives (1), bare presentatives (2), and *have/ha*-presentatives (3):

- (1) There's a long trip ahead of us.
Det ligger en lang reise foran oss.
- (2) A long trip is ahead of us.
En lang reise ligger foran oss.
- (3) We have a long trip ahead of us.
Vi har en lang reise foran oss.

Although the constructions are similar in syntax, semantics, and discourse function, there are important differences. The contrastive study defines these differences and at the same time makes the description of the individual languages more precise.

6 Contrastive studies

Highlighting the characteristics of the individual languages and defining the relationship between languages are just differences in perspective. In a comparative study the focus may be on language-specific, typological or universal features, as Aijmer & Altenberg say in their second point. Here I am particularly concerned with contrastive studies focusing on a comparison of pairs of languages.

One of the most serious problems of contrastive studies is the problem of equivalence. How do we know what to compare? What is expressed in one language by, for example, modal auxiliaries could be expressed in other languages in quite different ways. Then we do not get very far by a comparison of modal auxiliaries.

Most contrastive linguists have either explicitly or implicitly made use of translation as a means of establishing cross-linguistic relationships, and in his book on contrastive analysis Carl James reaches the conclusion that translation is the best basis of comparison:

We conclude that translation equivalence, of this rather rigorously defined sort [including interpersonal and textual as well as ideational meaning] is the best available TC [tertium comparationis] for CA [contrastive analysis]. (James 1980: 178)

In his paper on 'the translation paradigm' Levenston suggests that contrastive statements

... may be derived from either (a) a bilingual's use of himself as his own informant for both languages, or (b) close comparison of a specific text with its translation. (Levenston 1965: 225)

The use of multilingual corpora, with a variety of texts and a range of translators represented, increases the validity and reliability of the comparison. It can indeed be regarded as the systematic exploitation of the bilingual intuition of translators, as it is reflected in the pairing of source and target language expressions in the corpus texts.

It is probably not very well known that a corpus of this kind was set up for the Serbo-Croatian - English Contrastive Project (Filipovic 1969). The reasoning was formulated in this way by Spalatin:

(1) similarity between languages is not necessarily limited to similarity between elements belonging to corresponding levels in the languages concerned, and (2) similarity between languages is not necessarily limited to similarity between elements belonging to corresponding classes or ranks in the languages concerned. (Spalatin 1969: 26)

The basis of comparison was to be a bidirectional corpus of English texts and their translations into Serbo-Croatian (half of the Brown Corpus was selected!) and a corresponding material consisting of texts in Serbo-Croatian and their translations into English. The translations were especially commissioned for the project and were made by 'reasonably competent professional translators' who were 'deliberately chosen outside the Project' (Filipovic 1971: 84). Apart from the fact that we have used published translations, this is exactly the model which we chose many years later for the English-Norwegian Parallel Corpus. We were unaware of the parallel when we started the project, and the matter came up only recently in connection with Jarle Ebeling's thesis work.

An example of a corpus-based contrastive study is Berit Løken's (1996, 1997) investigation of expressions of possibility in English and Norwegian, based on the English-Norwegian Parallel Corpus. One of her findings was that there are major differences in the expression of epistemic possibility, although the two languages have similar means at their disposal. English epistemic modals are rendered in Norwegian in approximately half the cases by an adverb (4, 5) or by a combination of a modal and an adverb (6):

- (4) You *may* not know about this one: it's a modern sin.
Du kjenner *kanskje* [lit. 'perhaps'] ikke til den, det er en moderne synd.
- (5) I had become frightened on the way home, thinking that my father *might* be waiting up for me.
På veien hjem var jeg blitt ganske redd da jeg tenkte på at faren min *kanskje* [lit. 'perhaps'] satt oppe og ventet på meg.
- (6) At moments ... he realized that he *might* be carrying things too far.
Iblant ... innså han at han *kanskje kunne* [lit. 'perhaps could'] drive det for vidt.

The opposite relationship, i.e. where Norwegian epistemic modals were translated by some other expression than an English modal, was far less frequent. Løken concludes:

Most of the differences between English and Norwegian found in the corpus material may be results of the differing degrees of grammaticalisation of the two sets of modals, the Norwegian modals being less grammaticalised than the English ones. (Løken 1997: 55f.)

Løken's observations on English vs. Norwegian have later been shown to apply for English vs. Swedish as well in a recent study by Aijmer (1999), who associates the results with the relative degree of grammaticalisation of Swedish *kan* and English *may/might*. These studies illustrate how hypotheses on more general cross-linguistic differences can be inspired by corpus findings.

In this connection, I would like to quote from Andrew Chesterman's recent book on *Contrastive functional analysis*:

Corpus studies are a good source of hypotheses. But they are above all a place where hypotheses are tested, albeit not the only place. The more stringently a given hypothesis is tested - against a corpus, other speakers' intuitions, in a controlled experiment... - the better corroborated it will be. (Chesterman 1998: 60f.)

The use of a corpus is not bound to any one linguistic theory. The investigator is free to choose whatever linguistic theory is appropriate to account for the data.

7 Translation studies

In their third point Aijmer & Altenberg mention the study of differences between source texts and translations, i.e. original and translated texts in the same language. The study of the nature of translated texts by means of corpora was advocated by Baker (1993), and a special issue of a periodical for translators was indeed recently devoted to the 'corpus-based approach' (*META*, December 1998). In her opening paper the editor writes:

... a growing number of scholars in translation studies have begun to seriously consider the corpus-based approach as a viable and fruitful perspective within which translation and translating can be studied in a novel and systematic way. (Laviosa 1998: 474)

A study of translated texts may focus on features induced by the source language (as in Gellerstam 1996) or on more general characteristics of translated texts (as suggested by Baker 1993).

In my paper on "Loving and hating in English and Norwegian: A corpus-based contrastive study" (Johansson 1998c), we find a good example of differences in distribution between original vs. translated texts; see Figure 1. The figure shows that the English verbs are about three times as common as their Norwegian counterparts in the original fiction texts of the English-Norwegian Parallel Corpus. In the translated texts, however, the frequencies for the Norwegian verbs go up while the figures for the English verbs go down, presumably induced by the source language. Examples of this kind can easily be multiplied.

An example of a study of more general features of translation is the investigation by Linn Øverås (1996, 1998) of explicitation in translated English and Norwegian, based on the English-Norwegian Parallel Corpus. The following examples illustrate a rise in explicitness in translation from English into Norwegian (7, 8), and vice versa (9, 10):

- (7) At least I haven't had to pin anything this time, *he said*.
Denne gangen slapp jeg i alle fall å bruke skruer, *sa ortoped*. [lit. 'said the orthopedist']
- (8) Her companion hesitated, *looked at her*, then leaned back and released the rear door.
Den andre kvinnen nølte og så på piken, så snudde hun seg og trakk opp låseknappen på døren bak.
[lit. 'looked at the girl']
- (9) *En av dem* får tak i øksa til tømmermannen.
[lit. 'one of them']
But then one of them got hold of an axe belonging to the carpenter ...
- (10) *Husk nå* at du ikke gir fra deg så mye som en bitteliten lyd.
[lit. 'now remember']
Now remember, she admonished, not a sound.

In (7) and (8) the translator has inserted a more specific referential expression, in (9) connectors are added, and in (10) there is a reporting clause without a corresponding form in the original text. Such changes were far more common, in both directions of translation, than the opposite type of shift (implicitation). The ultimate objective of the studies of Øverås is to reach conclusions on translation norms. For this she is investigating a small corpus of translations assembled especially for the project where some of the best and most experienced translators in Norway have been commissioned to translate the same texts, a short story and a scientific article.

Now, if it is the case that translated texts have particular features, how can we then use such material for contrastive studies? I will return to this question in Section 9.

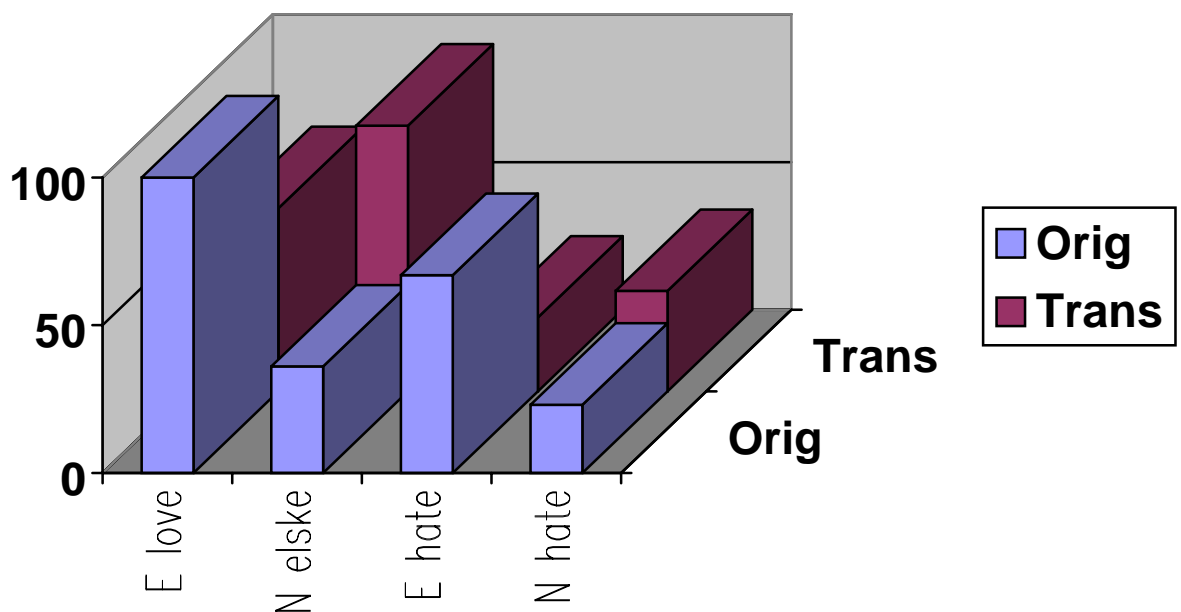


Figure 1 The distribution of English *love* and *hate*, and Norwegian *elske* and *hate* in original and translated fiction texts of the English-Norwegian Parallel Corpus (30 texts of each type)

8 Applications

In their last point Aijmer & Altenberg draw attention to a number of practical applications: in lexicography, language teaching, and translation. Those who know a bit about the history of linguistics are well aware of the danger of making exaggerated claims. In the decades after the Second World War there was a boom for applied contrastive linguistics. The hopes were dashed, however (see Section 2 above).

Now that there seems to be a new boom for contrastive linguistics, brought about by the new corpus methodology, it is important not to overstate the claims of this approach. No matter how good a multilingual corpus is, it will not allow us to make safe predictions of learners' difficulties. What can be done will depend upon the type of corpus. This brings me to my next point.

9 Which type of corpus for which type of study?

This question has been discussed in a number of papers, e.g. Lauridsen (1996), Granger (1996), Teubert (1996), and Johansson (1998a, 1998b), and therefore I will be as brief as possible. The nature of the corpus will vary depending upon the type of study. What is common for all the types of study we are concerned with here is that they require parallel corpora of some sort or other, in particular:

- multilingual corpora of original texts and their translations (for contrastive studies and translation studies)

- multilingual corpora of original texts which are matched by criteria such as genre, time of composition, etc. (for contrastive studies)
- monolingual corpora consisting of original and translated texts (for translation studies)

Rather than discussing the possibilities and limitations of each type, I will just mention that all three types can be combined within the same overall framework, as we have done in the English-Norwegian Parallel Corpus (Johansson 1998b; see Figure 2), and each type can then be used to control and supplement the other. In this way we can use the same corpus both for contrastive studies and translation studies and thus circumvent the problem raised in Section 7 above.

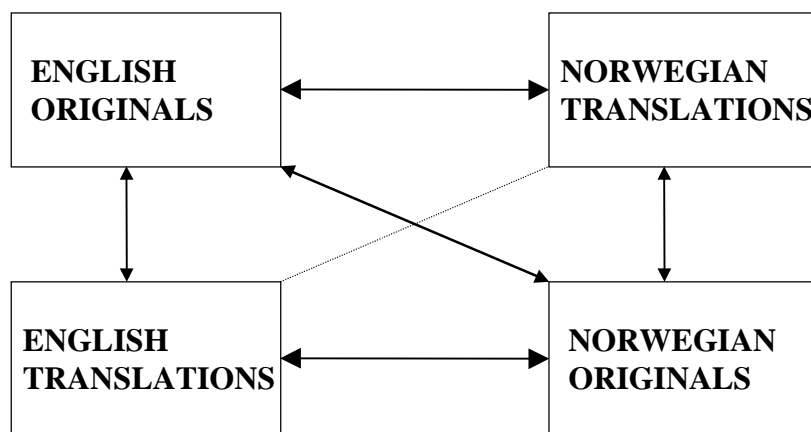


Figure 2 The structure of the English-Norwegian Parallel Corpus

A special type of corpus is required for the study of learner language, including differences between native and non-native texts (cf. Aijmer & Altenberg's third point) and between texts produced by learners with different mother-tongue backgrounds. A great deal of progress has been made recently in this area in connection with the International Corpus of Learner English (Granger 1998). The studies we get here are not contrastive in the narrow sense, but can be viewed as representing a corpus-based approach to error analysis and performance analysis.

10 Conclusion

In my survey I have taken contrastive linguistics in a very broad sense. Not everybody will agree with this broad definition, but I hope to have shown that corpora have an important role to play in all the areas I have taken up. The use of corpus-based methods is in many cases a follow-up and an extension of types of studies which were carried out by traditional methods in the past, e.g. the use of translation by Firbas and Levenston. But with the help of a corpus we get unprecedented opportunities to study and contrast languages in use, including frequency distributions and stylistic preferences. Corpora are absolutely essential for macrolinguistic studies, but they will also enrich studies of lexical and grammatical patterns.⁶

The development is only just beginning, however. I should like to draw attention to some particularly interesting challenges for the future:

- We need more work on multilingual corpora - and now I mean multilingual in a true sense, with a range of languages represented. The study of such corpora will increase our knowledge of language-specific, typological, and universal features.
- We need to carry on the work on corpora of translated texts and learner language, with systematic variation of the source and target language of the translations and of the mother tongue of the learners. In this way we can uncover general as well as language-specific features of translated texts and learner language.
- We need a new generation of grammars and dictionaries, based on the study of language in use. Ideally - whether we are talking about one or more languages - we need a new integrated language description, in electronic form, with links between grammar, dictionary, and corpus (Johansson 1998b). There are some beginnings, but we are still far from this goal.

To end this brief survey, I will quote the prediction from a paper we wrote at the beginning of our work on the English-Norwegian Parallel Corpus:

The importance of computer corpora in research on individual languages is now firmly established. If properly compiled and used, bilingual and multilingual corpora will similarly enrich the comparative study of languages. (Johansson and Hofland 1994: 36)

I hope the future will prove that we were right.

Notes

1. This is a revised and expanded version of a paper presented at the 20th ICAME Conference, University of Freiburg, May 1999.
2. As regards the different types of comparative linguistic studies, see Fisiak (1980), where a distinction is made between Comparative Historical Linguistics, Comparative Typological Linguistics, and Contrastive Linguistics. The latter can be theoretical or applied.
3. 'Contrastive linguistics' and 'contrastive analysis' are often used indiscriminately, but the former is the more general term and may be used to include developments from applied contrastive analysis (cf. note 5 below).
4. For a discussion of problems of applied contrastive linguistics, see Johansson (1975: 15), Ringbom (1994: 738-740), and Sajavaara (1996: 17-20).
5. Error analysis is concerned with the description and analysis of errors made by second language learners, whether they are induced by the mother tongue or derive from some other source. Performance analysis broadens the study to a description of the total performance of learners and is not just concerned with errors. In interlanguage studies the focus is on the development of learner language and on the strategies used by the learner. See the survey of developments originating from contrastive analysis in Ringbom (1994: 740-741).

6. The number of contrastive studies is growing rapidly. See the huge on-line bibliography on contrastive linguistics at:
<http://bank.rug.ac.be/contragram/biblio.html>

References

- Aijmer, Karin and Bengt Altenberg. 1996. Introduction. In Aijmer, et al (1996), 11-16.
- Aijmer, Karin. 1999. Epistemic possibility in an English-Swedish contrastive perspective. In H. Hasselgård and S. Oksefjell (eds.), *Out of corpora. Studies in honour of Stig Johansson*, 301-323. Amsterdam and Atlanta, GA: Rodopi.
- Aijmer, Karin, Bengt Altenberg, and Mats Johansson (eds.). 1996. *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*. Lund Studies in English 88. Lund: Lund University Press.
- Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, 233-250. Amsterdam and Philadelphia: John Benjamins.
- Chesterman, Andrew. 1998. *Contrastive functional analysis*. Amsterdam: Benjamins.
- Danielsson, Pernilla and Daniel Ridings. 1996. PEDANT: Parallel texts in Göteborg. Unpublished paper. Department of Swedish, University of Gothenburg.
- Ebeling, Jarle. In progress. Presentative constructions in English and Norwegian. Doctoral thesis. Department of British and American Studies, University of Oslo.
- Filipovic, Rudolf. 1969. The choice of the corpus for the contrastive analysis of Serbo-Croatian and English. In *The Yugoslav Serbo-Croatian - English Contrastive Project B. Studies 1*, 37-46. Institute of Linguistics, University of Zagreb.
- Filipovic, Rudolf (ed.). 1971. *Zagreb conference on English contrastive projects, 7-9 December 1970. Papers and discussion*. Institute of Linguistics, University of Zagreb.
- Firbas, Jan. 1992. *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.
- Fisiak, Jacek. 1980. Some notes on contrastive linguistics, *AILA Bulletin* 1 (27): 1-17.
- Fries, Charles C. 1945. *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press.
- Gellerstam, Martin. 1996. Translations as a source for cross-linguistic studies. In Aijmer et al (1996), 53-62.
- Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer et al (1996), 37-51.
- Granger, Sylviane. 1998. *Learner English on computer*. London: Longman.

- James, Carl. 1980. *Contrastive analysis*. London: Longman.
- Johansson, Stig. 1975. *Papers in contrastive linguistics and language testing*. Lund Studies in English 50. Lund: CWK Gleerup.
- Johansson, Stig. 1998a. On computer corpora in contrastive linguistics. In W.R. Cooper (ed.), *Compare or contrast? Current issues in cross-language research*. Tampere English Studies 6, 259-289. Tampere: University of Tampere.
- Johansson, Stig 1998b. 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*, 3-24. Amsterdam and Atlanta, GA: Rodopi.
- Johansson, Stig. 1998c. Loving and hating in English and Norwegian: A corpus-based contrastive study. In D. Albrechtsen, B. Henriksen, I. M. Mees, and E. Poulsen (eds.), *Perspectives on foreign and second language pedagogy. Essays presented to Kirsten Haastrup on the occasion of her sixtieth birthday*, 93-103. Odense: Odense University Press.
- Johansson, Stig and Knut Hofland. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie, and P. Schneider (eds.), *Creating and using English language corpora*, 25-37. Amsterdam and Atlanta, GA: Rodopi.
- Lado, Robert. 1957. *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.
- Lauridsen, Karen. 1996. Text corpora in contrastive linguistics: Which type of corpus for which type of analysis? In Aijmer et al (1996), 63-71.
- Laviosa, Sara. 1998. The corpus-based approach: A new paradigm in translation studies, *META* 43: 474-479.
- Levenston, E. A. 1965. The 'translation paradigm': A technique for contrastive syntax, *International Review of Applied Linguistics* 3: 221-225.
- Løken, Berit. 1996. Expressing possibility in English and Norwegian. Unpublished *hovedfag* thesis. Department of British and American Studies, University of Oslo.
- Løken, Berit. 1997. Expressing possibility in English and Norwegian, *ICAME Journal* 21: 43-59.
- Mathesius, Vilém. 1975. *A functional analysis of present-day English on a general linguistic basis*. Transl. L. Dusková, ed. J. Vachek. Prague: Academia.
- Ringbom, Håkan. 1994. Contrastive analysis. In R. E. Asher and J. M. Y. Simpson (eds.), *Encyclopedia of linguistics*, Vol. 2, 737-742. Oxford: Pergamon Press.
- Sajavaara, Kari. 1996. New challenges for contrastive linguistics. In Aijmer, et al (1996), 17-36.

Salkie, Raphael. 1999. How can linguists profit from parallel corpora? Paper given at the symposium on parallel corpora, 22-23 April 1999, University of Uppsala.

Spalatin, Leonardo. 1969. Approach to contrastive analysis. In *The Yugoslav Serbo-Croatian - English Contrastive Project B. Studies 1*, 26-36. Institute of Linguistics, University of Zagreb.

Teubert, Wolfgang. 1996. Comparable or parallel corpora? *International Journal of Lexicography* 9: 238-264.

Øverås, Linn. 1996. In search of the third code: An investigation of norms in literary translation. Unpublished *hovedfag* thesis. Department of British and American Studies, University of Oslo.

Øverås, Linn. 1998. In search of the third code: An investigation of norms in literary translation, *META* 43: 571-588.