Corpus-based contrastive studies: Beginnings, developments and directions

Hilde Hasselgård

University of Oslo

This article outlines the beginnings of corpus-based contrastive studies with special reference to the development of parallel corpora that took place in Scandinavia in the early 1990s under the direction of Stig Johansson. It then discusses multilingual corpus types and methodological issues of their exploration, including the *tertium comparationis* for contrastive studies based on different types of corpora. Some glimpses are offered of recent developments and current trends in the field, including the widening scope of corpus-based contrastive analysis, concerning language pairs as well as the kinds of topics studied and the methods used. The paper ends by identifying and discussing some challenges for the field and indicating prospects and directions for its future.

Keywords: contrastive analysis, multilingual corpus, parallel corpus, comparable corpus, tertium comparationis

1. Introduction

When multilingual corpora entered the scene in the early 1990s, they represented a new development in corpus linguistics, which had been predominantly monolingual until then, as well as in contrastive studies, which got a more solid empirical basis, new methods and a general boost. Contrastive analysis used to be considered an applied discipline of linguistics, closely associated with language teaching (Johansson 2007: 1). This is apparent, for example, in Lado's famous lines "…in the comparison between native and foreign language lies the key to ease or difficulty in foreign language learning …" (1957: 1). A similar view is found in McArthur's definition of contrastive analysis as "a branch of linguistics that describes similarities and differences among two or more languages […], especially in order to improve language teaching and translation" (1992: 261). Johansson (1998) notes, however, that the type of contrastive analysis that was widespread between the 1950s and the 1970s tended to focus "on linguistic systems (and subsystems) rather than on language use" (1998: 3). After the 1970s, contrastive analysis suffered a decline in popularity because, among other things, "the high hopes raised by applied contrastive analysis [for language teaching] were dashed" (Johansson 2007: 2).[1]

[1] Text-based contrastive studies were not, however, completely absent from the scene, as evidenced e.g. by the papers collected in Fisiak (1984) and Dušková (2015); see also the overview in Mair (2018).

After “many years of marginal status” (Salkie et al. 1998: v) the interest in contrastive analysis had been revived substantially by the late 1990s: there were a number of projects and meetings concerned with the comparison of languages, and a designated journal, *Languages in Contrast,* saw the light of day. This revival has been attributed to the advent of multilingual corpora (e.g. Salkie et al. 1998, Altenberg & Granger 2002, König 2012). The renewed interest in (corpus-based) contrastive analysis was described by Johansson as “contrastive linguistics in a new key”, specifying that

- the focus on immediate applications is toned down;
- the contrastive study is text-based rather than a comparison of systems in the abstract;
- the study draws on electronic corpora and the use of computational tools. (Johansson 2012: 46)

The newly emerged type of contrastive analysis is understood as the systematic comparison of two or more languages, with the aim of describing their similarities and differences (Johansson 2007: 1). Mair (2018: 10) observes that the use of multilingual corpora takes “the descriptive comparison of languages from the level of the decontextualized system of choices to language in use”, thus echoing Johansson’s (2012) second point. Corpus-based contrastive analysis is situated somewhere between monolingual analysis and comparative/typological studies by focusing on the comparison of a small number of languages and by tending to emphasize differences between them rather than similarities (Salkie et al. 1998: vi).

An immediate advantage of the corpus-based approach was that it allowed a more systematic comparison not only of structures, but also their conditions of use (Johansson 2011: 125). Like other branches of corpus linguistics, corpus-based contrastive analysis utilizes large and principled digital collections of natural texts (corpora) and depends on both quantitative and qualitative analytical techniques (Biber et al. 1998). Due to the availability of “quantitative information on correspondences between two languages” corpus-based contrastive analysis can gain “a more objective picture of the degree of correspondence of patterns” (Barlow 2008: 105). At the same time, corpus-based cross-linguistic studies represent a step forward in terms of testability, authenticity and general empirical adequacy (Gómez González et al. 2008: xvii).

At present, corpus-based contrastive linguistics is a well-established field of research which can be distinguished from the neighbouring fields of learner corpus studies, translation studies, and typological studies (Ebeling & Ebeling 2013a: 44 ff., König 2012: 10). For example, although corpus-based contrastive analysis obviously shares important characteristics with corpus-based translation studies (Laviosa 2002), the emphasis can be said to be different. Both fields work with both parallel (translation) corpora and comparable corpora (see Section 3.1), with parallel corpora being for obvious reasons indispensable in translation studies (Bernardini 2015). But while translation studies often focus on the translation process, features of translated texts and/or the application of findings to practical translation, contrastive analysis is first and foremost interested in the description and comparison of the languages involved.

In the following sections, I will attempt to give an overview of the beginnings of corpus-based contrastive analysis, to take stock of recent developments and the present state of the field, and to indicate some potential future directions.

2. Beginnings and early developments

A seminal contribution to the rise of corpus-based contrastive analysis was the compilation and completion of the English-Norwegian Parallel Corpus (ENPC) and its sister corpus, the English-Swedish Parallel Corpus (Aijmer et al. 1996). The plans for the English-Norwegian Parallel Corpus were first presented as a pilot project at the ICAME conference in 1993 (Johansson & Hofland 1994). The paper argues as follows for using corpora in cross-linguistic research:

> Bilingual corpora provide evidence on similarities and differences between two languages. They make it possible to carry out text-based contrastive studies, while traditional contrastive studies have often focused on a comparison in the abstract of language systems, or parts of language systems, without being connected to real texts. (Johansson & Hofland 1994: 25).

The bulk of the article concerns methodological issues in the construction of the corpus, particularly methods for aligning source and target texts at sentence level. Johansson & Hofland refer to recent and contemporary work on sentence alignment in other projects, particularly that of Church & Gale (1991), before they present a pilot project in which they have further developed both the methods and the software for sentence alignment. In this experiment they use, among other things, a combination of sentence length and a so-called “anchor word list”, i.e. a simple bilingual list of words that are expected to correspond well between the original and the translated text (Johansson & Hofland 1994: 30, see also Ebeling & Ebeling 2013b: 28). The next step, not resolved yet at the time of Johansson & Hofland’s ICAME presentation, was to use the alignment to enable parallel concordancing – an application that was developed later by Jarle Ebeling (Ebeling 1999).[2] Johansson & Hofland conclude their paper by stating that “the importance of computer corpora in research on individual languages is now firmly established. If properly compiled and used, bilingual corpora will similarly enrich the comparative study of languages” (1994: 36).

Soon after the ENPC project had first been presented, it was joined by other teams who were interested in developing similar corpora of English/Swedish and English/Finnish (Aijmer et al. 1996). Within the Nordic research network “Languages in contrast” (1994-1997; see Aijmer et al. 1996) three corpus teams came together to share English original texts as well as software, expertise and experience in preparing texts for parallel corpora.[3] The interest in corpus-based contrastive studies was not limited to the Scandinavian context. Other teams elsewhere started compiling parallel corpora of other language pairs, either inspired by or independently of the ENPC project. Some examples are the INTERSECT corpus for English, German and French (Dickens & Salkie 1996), the PLECI corpus for English and French in Poitiers and Louvain (first outlined

[2] See Barlow (1995) for another, unrelated, software for alignment and parallel concordancing.

[3] The corpus projects, apart from the ENPC, were the English-Swedish Parallel Corpus (ESPC), directed by Bengt Altenberg and Karin Aijmer, and the Finnish-English Contrastive Corpus Studies Project (FECCS), directed by Kari Sajavaara.

in Granger 1996: 39; see also Gilquin 2000/2001), and later e.g. the CroCo corpus of English and German at Saarbrücken (Hansen-Schirra et al. 2012), the P-ACTRES corpus for English and Spanish (Sanjurjo-González & Izquiredo 2019), the COMPARA corpus of English and Portuguese (Frankenberg-Garcia & Santos 2003), the Dutch Parallel Corpus (of Dutch, French and English; see Macken et al. 2011) and InterCorp for a number of language pairs (Čermák & Rosen 2012).

Clearly, the 1990s and early 2000s were a favourable period for this type of work. The general technological advances made it possible to extend corpus methods to multilingual corpora, with the development of software for the alignment of translated texts and for parallel concordancing (J. Ebeling 2016), and, furthermore, corpus linguistics was becoming much more widespread due to the increased availability of personal computers and the advent of the internet, which greatly facilitated storage of and access to corpora.

3. Using corpora in contrastive studies

This section takes a look at the types of corpora that are suitable for contrastive studies. Then follows a discussion of some issues of contrastive analysis (CA) methodology in research based on multilingual corpora.

3.1 Types of multilingual corpora

For a multilingual corpus – or corpora in different languages – to be suitable for contrastive analysis, the texts must be related to each other in some way, either through translation or through text comparability. The main types of multilingual corpora available can be outlined as in Figure 1 (see also Johansson 2007: 9, Aijmer 2008: 276).[4]

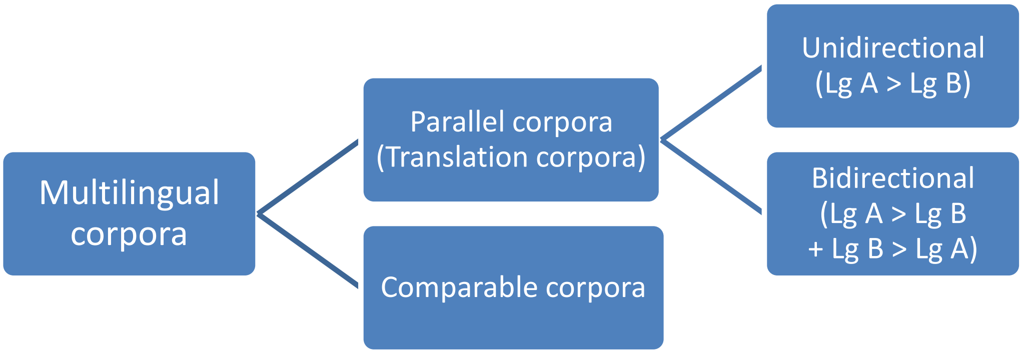

Multilingual corpora
Parallel corpora (Translation corpora)
Comparable corpora
Unidirectional (Lg A > Lg B)
Bidirectional (Lg A > Lg B + Lg B > Lg A)

Figure 1 Types of multilingual corpora for contrastive analysis.

A parallel (translation) corpus may be said to contain "the same text samples in each of two languages, in the sense that the samples are translations of one another" (Oakes & McEnery 2000: 1). In a unidirectional corpus, all the original texts are in one language and the translated texts in another, while bidirectional parallel corpora have original texts in (all) languages concerned and translations into the other language(s). For a corpus with more than two languages,

[4] The abbreviations Lg A and Lg B in Figure 1 refer to different language. In multilingual rather than bilingual corpora Lg C, D etc. may be added to the figure. See also Laviosa (2002: 34 ff.).

'bidirectional' may be replaced by 'multidirectional' (or 'multi-source'; Laviosa 2002: 37). Examples of such corpora are the English-German-Norwegian part of the Oslo Multilingual corpus (Johansson 2007: 18) and the Dutch Parallel Corpus (Macken et al. 2011), which have both originals and translations in all the languages involved.

A comparable corpus "consists of texts from different languages which are similar or comparable with regard to a number of parameters such as text type, formality, subject-matter, time span, etc." (Aijmer 2008: 276). Comparable corpora have the advantage that they are not restricted to registers that are translated, and are the only viable option for cross-linguistic studies of e.g. spoken conversation. Another advantage is that the comparison of original texts avoids the problems connected with using translations in contrastive studies (see e.g. Lauridsen 1996: 67, Mauranen 1999: 162 ff., Xiao & McEnery 2010: 7). For example, Gellerstam (1986) shows that certain lexical items are much more frequent in Swedish translations (from English) than in original Swedish texts. This phenomenon is referred to as 'translationese' (Gellerstam 1986) or 'translation effects' (Johansson 2007), and it is a reminder that translated texts, even if they are error-free and idiomatic, may be quantitatively and qualitatively different from non-translated texts in the same language.

Acknowledging the potential pitfalls of basing language comparisons on translations, Johansson notes that "translation corpora are insufficient as sources of contrastive studies […] and need to be combined with comparable corpora" (2007: 5). The bidirectional parallel corpus model represented by the ENPC can combine the two main types of multilingual corpora by having comparable original texts in both languages concerned, as illustrated in Figure 2; see also Zanettin (2011: 21).[5] Such a corpus makes it possible to use the original texts in both languages as a comparable corpus and to study relations between source and target texts in one or both directions of translation (Johansson 1998: 8). Importantly, "the translation corpus is a source of perceived similarities" while "the comparable corpus is used to control for translation effects" (Johansson 2007: 5).

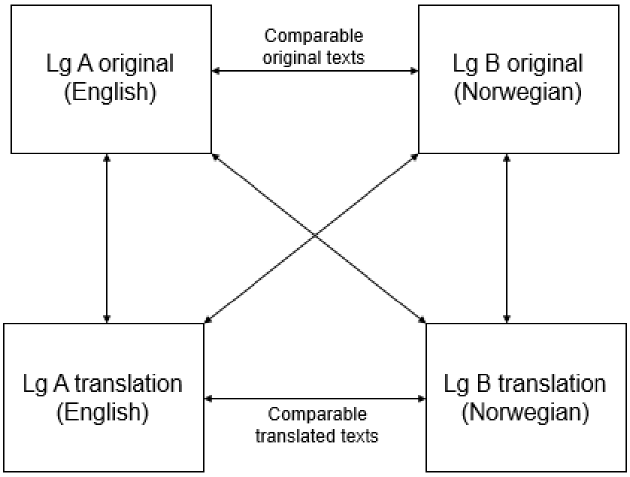

Lg A original (English)
Comparable original texts
Lg B original (Norwegian)
Lg A translation (English)
Comparable translated texts
Lg B translation (Norwegian)

Figure 2 The structure of the English-Norwegian Parallel Corpus (ENPC).

[5] The idea of a comparable bidirectional translation corpus was indeed an aim in compiling the ENPC. See further Johansson et al. 1999/2002 for details on text selection for this corpus.

3.2 Multilingual corpora and contrastive analysis methodology

The introduction of multilingual corpora for contrastive analysis (CA) required new methods for exploiting them. The development of such methods did not start from scratch but borrowed from e.g. monolingual corpus linguistics, traditional contrastive analysis and descriptive (text-based) translation studies. The choice of method for a particular study clearly depends on the type of corpus that is used. Crucially, only parallel (translation) corpora allow parallel concordancing (following alignment), while the use of comparable corpora typically requires a decision on which linguistic items to compare prior to the corpus investigation.[6]

The bidirectional corpus model is particularly well suited for contrastive studies because the translation relation can provide a *tertium comparationis* (TC) for the cross-linguistic comparison, i.e., a "background of sameness against which differences can be viewed and described" (Johansson 2007: 39). More precisely, we can assume that the pairs of source and target texts are semantically and pragmatically (near-) equivalent since the translators' task is to transfer the ideational, interpersonal and textual meaning of the original text to the translation (James 1980: 178). Hence, we can assume that the linguistic elements used for expressing these meanings are comparable across the languages. At the same time, the presence of original texts in both languages provides a way of avoiding the translation bias which can otherwise be a problem for unidirectional parallel corpora (Johansson 2007: 9 ff., Ebeling & Ebeling 2013a: 41 ff.).[7] However, an obstacle to the bidirectional model is the limited availability of translated texts for relevant language pairs and text types. For example, during the compilation of the ENPC, the original goal of collecting similar proportions of fiction and non-fiction texts had to be abandoned for lack of available translations of non-fiction, especially from Norwegian into English (Johansson et al. 1999/2002, section 1.3).

Importantly, the use of translations implies that the researcher need not determine beforehand which (pair/set of) linguistic items should be matched in the comparison: instead, cross-linguistic correspondences can be identified via *translation paradigms*, i.e. "the set of forms in the target text which are found to correspond to particular words or constructions in the source text" (Johansson 2007: 23). Such paradigms are established through parallel concordancing in aligned texts (Johansson 2012: 46f.), and are useful because they arguably "provide a blueprint of the similarities and differences between the languages compared" (Aijmer & Altenberg 2013b: 2). An example of a translation paradigm is presented in Table 1. It shows that in the fiction part of the ENPC, the English speaking verb *talk* is translated predominantly by the Norwegian verb *snakke* but also with e.g. *prate* ('chat'). *Snakke* in turn is regularly translated by either *talk* or *speak*, which is also frequent enough to merit a place in a cross-linguistic comparison of speaking verbs. Both verbs also have less recurrent members of their respective paradigms (the category 'other' comprises those that occur only once or twice). Importantly, the correspondences shown in Table 1 differ from the translations suggested in a major bilingual dictionary of English and Norwegian (Kunnskapsforlaget 2001). The use of translation paradigms is thus a systematic,

[6] More corpus-driven ways of identifying objects of comparison have been proposed, such as the n-gram method (e.g. Cortes 2008, Granger 2014, Ebeling & Ebeling 2017, Hasselgård 2017a).

[7] Filipović (1984: 109) argues that "only such a bidirectional corpus, consisting really of two corpora: a corpus of L1 and a corpus of L2, can ideally satisfy all the requirements for the primary data for CA".

usage-based way of making cross-linguistic relationships visible and identifying relevant items for the cross-linguistic comparison (see also Dyvik 2005).

Table 1 Translation paradigms for the verbs *talk* and *snakke*.

E→N		**Gloss**	**N**	**%**	**N→E**		**N**	**%**
talk	snakke	'talk'	204	76.1	**snakke**	talk (v.)	313	64.7
	prate	'chat'	14	5.2		speak	80	16.5
	fortelle	'tell'	4	1.5		say	13	2.7
	si	'say'	3	1.1		mention	9	1.9
	other		34	12.7		have a word	7	1.4
	Ø		9	3.4		discuss	6	1.2
						chat (v.)	4	0.8
						tell	4	0.8
						other	40	8.3
						Ø	8	1.7
			268	**100**			**484**	**100**

Following on from the identification of translation paradigms, it is possible to measure cross-linguistic equivalence by calculating *mutual correspondence* (MC) (Altenberg 1999).[8] This measure, which presupposes a bidirectional parallel corpus, indicates the frequency with which different (grammatical, semantic and lexical) expressions are translated into each other. It is calculated and expressed as a percentage by means of the formula

$$\frac{(A_t + B_t) * 100}{A_s + B_s}$$

where A and B symbolize different languages and t and s stand for target and source language respectively (Altenberg 1999: 254). Applied to the paradigms given in Table 1, the calculation of the mutual correspondence between *talk* and *snakke* is as follows: (204 + 313) / (268 + 484) * 100 = 68.8%. The correspondence between *speak* and *snakke*, on the other hand, is 28.2%. Particularly the match between *speak* and *snakke* is asymmetrical, since most occurrences of *speak* are translated by *snakke*, but most occurrences of *snakke* are translated by *talk*, as shown in Table 1.[9]

Comparable corpora can provide more varied data than parallel corpora since they do not depend on the existence of translated texts. However, they lack the in-built *tertium comparationis* provided by a translation relation. Instead the TC for the comparison is text comparability, typically supplemented by a perceived similarity of forms, meanings and/or functions of the linguistic items being compared.[10] An example of this is provided in Xiao & McEnery's (2010) studies of English and Chinese. Although there is no explicit discussion of TC, it appears from the studies that the cross-linguistic comparison is based on (i) corpus comparability, and (ii) grammatical categories that are regarded as equivalent

[8] The same concept is sometimes referred to as 'mutual translatability', e.g. Cosme & Gilquin 2008.

[9] See Ebeling & Ebeling (2015: 79) for the complementary measure of *reverse mutual correspondence* (rMC), which is expressed as "a percentage based on the number of times our items have each other as source".

[10] See Ebeling & Ebeling (2013a, chapter 2) for an overview of other types of TC.

(e.g. aspect marking, quantifying constructions and passives). For example, the chapter on aspect marking starts with juxtaposed grammar-based descriptions of aspect in both languages (Xiao & McEnery 2010: 11-12), thus laying the foundation for the corpus investigation. The corpora are comparable because the English corpora FLOB and Frown and the Lancaster Corpus of Mandarin Chinese were all compiled according to the sampling frame originally devised for the Brown corpus (Xiao & McEnery 2010: 8-9; McEnery & Hardie 2012: 97).

Another example is provided by Lewis (2017), who uses an English-French comparable corpus of political speeches given by politicians in government in Britain and France during the late 1990s and the early 2000s.

> The sociocultural parameters of the situations in which such texts are produced are well-defined and similar across the two languages, so that identifying comparable texts for a corpus is fairly straightforward. (2017: 145)

Lewis furthermore identifies linguistic markers of the discourse relation ‘addition’ in both languages using lists in (monolingual) grammars, dictionaries and thesauri. The TC for the study thus relies on text comparability and linguistic realizations of a semantically and functionally specified field of meaning – in this case a particular type of coherence relation.

Since comparable corpora do not contain pairs of source and target texts, they obviously do not permit the identification of translation paradigms or mutual correspondence values for linguistic items assumed to correspond to each other. For example, Xiao & McEnery’s (2010) investigation of aspect in English and Chinese would run the risk of excluding expressions that are absent from monolingual descriptions but might occur in the translation paradigm of an aspectual marker. Various techniques for gauging translational equivalence without translational evidence have been proposed, of which an interesting avenue consists in considering the collocational patterns and the semantic prosodies of the items compared (Tognini-Bonelli 1996 and 2001). However, Tognini-Bonelli’s proposal applies particularly to lexical items; not all linguistic items – and perhaps not even all kinds of lexical items – may lend themselves equally well to this type of analysis. It thus remains a challenge for contrastive studies based on comparable corpora to develop ways of measuring the nature and degree of equivalence between the items compared (see Johansson 2011: 127).

4. Ongoing developments

In the years that have passed since the introduction of multilingual corpora, the field of corpus-based contrastive analysis has grown rapidly and spread across the linguistic community. There are more and more multilingual corpora available, special events focusing on corpus-based contrastive analysis (such as workshops at ICAME conferences and the UCCTS conference series) and a large number of publications within the field. This section outlines some of the developments.

4.1 The development of corpus-based CA through the lens of *Languages in Contrast*

To give an indication of the recent developments in corpus-based contrastive studies this section surveys the articles published in *Languages in Contrast* from its first issue in 1998 until the end of 2018.[11] I have crudely divided the period into two: 1998−2009 and 2010−2018, as shown in Figures 3 and 4.[12] It should be noted that the analysis was carried out on the basis of article abstracts and thus may not be 100% accurate. A rather generous operationalization of the term "corpus-based" was applied, to include studies that may rather be "corpus-informed" (Johansson 2011: 116) as well as some studies based on pairs of translated texts that may not fulfil all the traditional criteria for a corpus in terms of e.g. size, principled collection or use of digital tools for exploration (cf. Biber et al. 1998: 4, McEnery & Hardie 2012: 1-2).

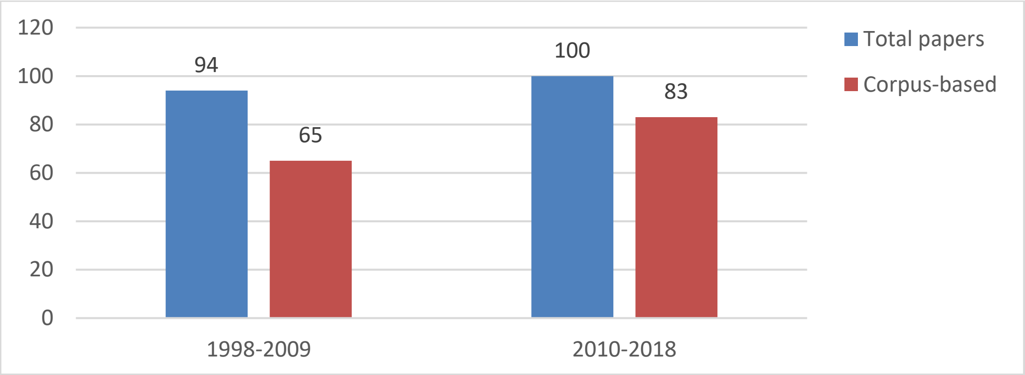
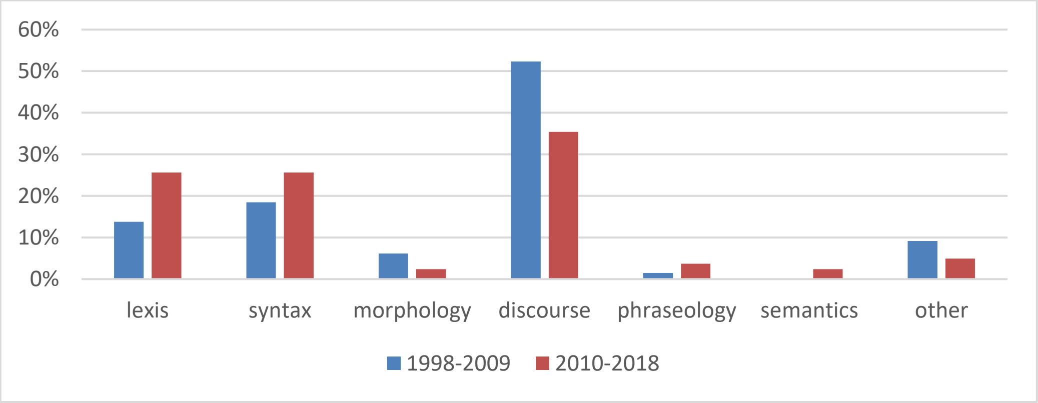


Figure 3 Corpus-based papers in *Languages in Contrast* 1998−2018 (raw numbers).

Figure 3 shows that the proportion of corpus-based papers has always been high, but it has increased in recent years. The increase should not be due to biased editorship during the last period, since all the editors of the journal have been corpus linguists. Rather it can be taken to reflect "the rapidly increasing interest in multilingual corpora" (Johansson 2012: 45) as well as the "great vitality and productivity in the field" (Aijmer & Altenberg 2013b: 3).



[11] A similar survey is found in Defranq (2015: 1).

[12] Although the two periods are of unequal length, they contain the same number of issues due to the publication history of the journal (see https://benjamins.com/catalog/lic).

Figure 4 General topics of corpus-based papers in *Languages in Contrast* 1998−2018.

It is harder to find a trend in the kinds of topics that are discussed in corpus-based papers in the journal. Again, based on article abstracts, I tried to identify the general topic of each paper. Only one topic was recorded for each paper, i.e. the one that seemed predominant. This obviously gives a simplified picture, since many papers consider linguistic features from a variety of angles. The topics are distributed as shown in Figure 4. Admittedly, the figures are skewed by the fact that there are two special issues focusing on corpus-based discourse studies (7:2 and 9:1), which explains the very high proportion for discourse in the first period. Even so, Figure 4 gives a clear indication that discourse, lexis and syntax are the three most common areas of corpus-based contrastive analysis.

A rough idea of the languages studied is given in Figure 5. The total impression is that there is a slightly increased diversification. However, the vast majority of corpus-based contrastive studies in both periods compare English and one other language ('English–X'), the most common of which are Germanic and Romance languages, but there are also comparisons with e.g. Slavic languages, Chinese and Arabic in both periods. There is a clear rise in studies not involving English ('Non-English'), which is partly due to a special issue on Romance languages (18:1) but probably also to greater availability of corpora in languages other than English. Relatively few studies include more than two languages ('Multiple languages'), and those that do are usually limited to three languages unless they are of a typological nature. Only a very small minority of these do not include English in the mix ('Multi&Non-English').

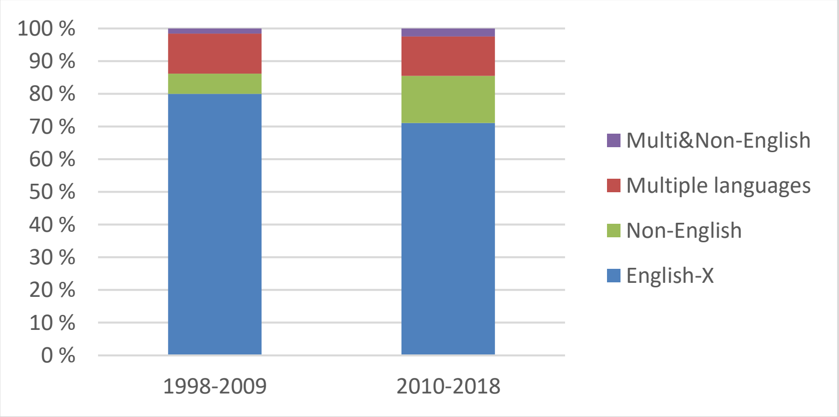


Figure 5 Language comparisons in corpus-based papers in *Languages in Contrast* 1998−2018.

The studies are based on both parallel and comparable corpora. Among parallel corpora, Europarl (Koehn 2005) is widely used, as are corpora of fiction, a text type which is frequently translated between many language pairs and thus also readily available to many researchers, who use established corpora (e.g. the ENPC, ESPC, PLECI, InterCorp) as well as more ad-hoc (usually small) collections of text pairs. For studies of other registers, comparable corpora tend to be used, often representing 'specific purposes', e.g. academic articles, political speeches, newspaper texts, and texts related to business and tourism. Many of the 'specific-purposes' corpora are very small and compiled by the respective authors.

4.2 Widening the scope: current trends

Since the 1990s the field of corpus-based contrastive linguistics has grown and diversified to include comparisons of more languages and language pairs (see e.g. Doval & Nieto 2019). The number of languages compared has increased in two ways: more language pairs are being investigated across studies, and individual studies may include more than two languages while retaining the detailed, close perspective of CA (see e.g. Viberg 2013 and 2017; van der Auwera 2012, Ström Herold & Levin 2019). At the same time, more types of lexical items are investigated (Egan & Dirdal 2017: 7-8): while early studies often focused on verbs and adverbs we now have studies of the whole gamut of word classes, including pronouns (Johansson 2007: 175 ff., Coussé & van der Auwera 2012), negators (Johansson 2007: 155 ff.) and prepositions (Egan 2013).

Another important trend is the increasing attention to register, as evidenced in e.g. Xiao & McEnery (2010), Lefer & Vogeleer (2014), Aijmer & Lewis (2017) and Kunz et al. (2018). As Aijmer & Lewis point out in the introduction to their edited volume, "the contrastive point of view highlights the dependences of the patterns on different social and cultural practices in the compared languages" (2017: 3). Monolingual studies (such as Biber 1988) have already established that lexicogrammatical features vary across registers, while cross-linguistic studies may find that registers vary in different ways or to different extents (e.g. Biber 2009, Kunz et al. 2018: 201 f.). Fløttum et al. (2006: 54) report that in their study of research articles in English, French and Norwegian, language was a less important variable than scientific discipline for the majority of the linguistic features examined.

Though not evident from the survey of topics in Figure 4, there is also increasing attention to the phraseological view of language (e.g. Sinclair 2004: 140 ff.) in contrastive studies of lexis. As noted by Egan & Dirdal (2017: 13) "contrastive phraseological studies give support to the ideas of extended lexical units and the importance of sequences in several ways". For example, Ebeling & Ebeling (2014: 209) observe that two superficially similar patterns in English and Norwegian (*for * sake* and *for * skyld*) differ as regards productivity and semantic prosody, which in turn affects their typical communicative functions.

It is also fair to say that phraseologically oriented studies have become more corpus-driven. Older CA studies usually tackled phraseology through idioms and fixed phrases; for example Cignoni et al. (1999) studied the occurrence and variability of predefined idioms in two general corpora of Italian and English. More recent studies move away from pre-defined phrases in their explorations of multi-word units in multilingual corpora. The multi-word units may be identified by retrieving collocates or clusters surrounding predefined nodes (e.g. Hasselgård 2017b), realizations of predefined collocational frameworks (e.g. Hasselgård 2016, Ebeling 2018), or lists of n-grams (e.g. Ebeling & Ebeling 2013a and 2017, Granger 2014, Čermáková & Chlumská 2017, Šebestová & Malá 2018 and Chlumská & Lukeš 2018). Particularly the last two retrieval methods present problems for multilingual application due to for example systemic morphological and/or syntactic differences between the languages compared. For collocational frameworks, the problem may be to identify similar frameworks in two languages.

For example, the framework used in Hasselgård (2016), "the N1 of the N2", could only be applied to English because Norwegian uses suffixes instead of articles to mark definiteness of nouns. By contrast, the frames explored by Ebeling (2018), *it* BE ** that* and the corresponding *det* VÆRE ** at*, worked well for both languages compared. The n-gram method has been shown to be rewarding for the language pairs English-French (Granger 2014) and English-Spanish (Cortes 2008) but to be more challenging for the pairs English-Norwegian (Ebeling & Ebeling 2017, Hasselgård 2017a) and English-Czech (e.g. Čermáková & Chlumská 2017, Chlumská & Lukeš 2018), due to systemic morphological and syntactic differences between the languages. However, a potential gain of such explorations is that they can take us closer to a cross-linguistic comparison of idiomaticity, exposing the natural and preferred patterns of expression in the languages compared.

5. Challenges for corpus-based contrastive linguistics

In one of his last papers, Johansson identified the following challenges for corpus-based contrastive linguistics:

> We need to widen the range of languages, including the variety of texts. We need multi-register corpora. We need corpora with annotation of features which cannot be easily found in raw, unannotated text. Above all, we need to learn more about how we can best exploit multilingual corpora. (Johansson 2012: 64)

Some of these challenges are currently being met while others are yet to be resolved. As indicated in the previous section, the range of languages and the variety of texts included in corpus-based contrastive studies is already being expanded. For the most part, this expansion takes place in comparable corpus analysis (see e.g. Doval & Nieto 2019), and to some extent with unidirectional parallel corpora, as discussed in Section 4.1. There are fewer newcomers among bidirectional parallel corpora,[13] most likely due to the difficulties of finding (and getting permission to use) sufficient and suitable text pairs in both translation directions required. Given the growing reliance on comparable corpora, the challenge of identifying a reliable *tertium comparationis* should not be forgotten, i.e. the establishment of a background of sameness in the absence of a translation relation. Although this challenge can be met in various ways (see above), not least in the combined use of translations and comparable data (Johansson 2007: 10), there is still no commonly agreed way to ensure that a comparison based on comparable corpora considers all the relevant forms with "similar meanings and pragmatic functions" if they belong to different formal classes (Johansson 2011: 127).

There are relatively few examples of multilingual multi-register corpora. We may mention here the two (related) corpora CroCo and GECCo, which are concerned with the language pair English and German and which contain both translated and comparable texts, both spoken and written (Hansen-Schirra et al. 2012). They can thus be used both as (bidirectional) parallel and comparable corpora. Because many registers are never or only rarely translated, multi-register

[13] An example of such a newcomer is the multidirectional Linnaeus University English-German-Swedish Corpus (LEGS); see Ström Herold & Levin (2019).

analyses largely need to rely on comparable corpora. For example, the Lancaster Corpus of Mandarin Chinese mentioned above (Xiao & McEnery 2010) can be used in conjunction with any member of the Brown family of corpora (McEnery & Hardie 2012: 97) to form a comparable corpus, as it is compiled according to the same sampling frame. Based on a similar idea, a large collaborative project is under way, namely The International Comparable Corpus (ICC), which uses the sampling frame of the International Corpus of English (ICE) for a number of languages (Kirk & Čermáková 2017).[14]

Various projects have accepted the challenge of annotating their corpora for features beyond PoS-tagging and lemmatization. Again, CroCo and GECCo can serve as examples, as they contain "several annotation and alignment layers on word, chunk, clause, and sentence level" (Hansen-Schirra & Neumann 2012: 35). Another initiative is the COST action TextLink (2014-2018), which aimed to provide multilingual resources for "discourse annotation, which considers phenomena above the sentence level such as information structure or coherence relations" (Crible & Degand 2019: 72).[15] An important observation made by Crible & Degand is that different models for annotation depend on discourse models, and the usefulness of (multilingual) discourse annotation thus depends not only on cross-linguistic comparability, but also on its relevance for specific research questions (ibid.: 95). In a similar vein, Johansson asks rhetorically:

> If corpora are annotated independently for each language, to what extent is the analysis comparable? If they are provided with some kind of language-neutral annotation […], to what extent do we miss language-specific characteristics? (2007: 306).

Semantic annotation is potentially very useful for contrastive studies as is demonstrated in Maia & Santos's (2012) cross-linguistic exploration of 'fear', where a combination of lexical clues and syntactic patterns is used for automatically identifying the domain in English and Portuguese. Other types of semantic and pragmatic annotation include evaluation/appraisal and rhetorical moves analysis (e.g. Taboada et al. 2014 and López Arroyo & Roberts 2015). Common to most of these annotation initiatives is that they are labour-intensive, requiring either a fully manual approach or extensive human intervention in (semi-) automatic processes. However, the potential gains, provided the resulting annotation can be trusted, include more or less direct access to the expressions belonging to equivalent semantic and pragmatic fields in different languages, which – even in the case of comparable corpora – could form a viable *tertium comparationis* for cross-linguistic studies.

The last point in Johansson's (2012) list of challenges concerns methods and tools for exploring multilingual corpora. To some extent the developments in the cross-linguistic comparison of phraseology (discussed above) are rising to this challenge. Other developments in corpus linguistics at large include the increased role of statistical methods. Applied to multilingual data, these have so far been more visible in translation studies (e.g. Evert & Neumann 2017) than in contrastive studies, which must be expected to follow suit (see e.g. Gries et al. [to appear] for a recent contribution).

[14] See also https://korpus.cz/icc for information about the ICC, and https://www.ice-corpora.uzh.ch/en/design.html for the design of ICE.

[15] For information on TextLink, see http://textlink.ii.metu.edu.tr/ (accessed May 2019).

Finally, it must not be forgotten that the quality of the data is crucial for the quality of the output. No matter how sophisticated the technology and the statistical methods are, one cannot obtain reliable results from a poor dataset (Wallis 2007). The usefulness of multilingual corpora presupposes that they have been "compiled with care" (Johansson 2012: 125). Hence, "[t]he mere existence of multilingual resources is not sufficient. To be maximally useful, they should be organised in a principled manner and encoded according to some accepted standard" (Johansson 2007: 305). Linguists in the business of contrastive corpus-based studies thus need to be conscious of the content of the corpora (e.g. consistency vs idiosyncrasies across texts), the distinctions between source and target texts in the case of parallel corpora, and the degree of text comparability in the case of comparable corpora so that the contrastive analysis will produce sound and reliable results.

6. Future directions

It may be assumed that the future directions of corpus-based contrastive studies lie in the present challenges and the ways in which they are being met. Many of the challenges discussed in the previous sections have not yet been resolved, and will stay with us for years to come. Thus, we can expect that practitioners in the field will continue to develop increasingly sophisticated methods for studying e.g. phraseology, pragmatics and discourse phenomena and to ground these methods in corpus linguistics as well as theoretical and methodological developments in the respective subfields. Furthermore, as an interdisciplinary enterprise, corpus-based contrastive studies will continue to be influenced by the developments in other fields of linguistics as well as in digital humanities and technology. And not least, changes in society at large will create new needs and challenges in corpus linguistics.

One such change is the increasing multimodality of texts, whether they are spoken, printed or online. Corpus linguistics is only beginning to deal with this. When Karin Aijmer places multi-modal corpora high on her wish-list for the future (Šinkūnienė 2017: 190), she is talking mainly about the study of spoken communication, where a multi-modal corpus might include gaze, gesture, tone of voice, etc. However, a written multi-modal corpus might include images, sound clips, emoji, etc., all integral components in the overall make-up of any online text. Multimodality is already being analysed contrastively on the basis of smallish corpora or (sets of) texts (e.g. Kong 2013, Kefala & Sidiropoulou 2016), but a fully-fledged corpus-linguistic perspective will require solutions to the technical challenges of representing and retrieving the different modalities in large, searchable corpora.

Contrastive corpus pragmatics is a field in its very infancy, and it can be expected that it will produce new insights into cross-linguistic and cross-cultural differences in verbal behaviour.[16] Much important work has been done on discourse markers (see e.g. Aijmer & Simon-Vandenbergen 2006), but many other pragmatic expressions are harder to identify. For example, O'Keeffe (2018) discusses the challenge of identifying speech acts in a corpus and advocates the

[16] An indication of this research interest is the establishment of the journal *Contrastive Pragmatics*, whose first issue appeared in 2020 (https://brill.com/view/journals/jocp/jocp-overview.xml).

development of reliable function-to-form approaches to corpus investigation. Such approaches are particularly interesting for contrastive analysis, since most cross-linguistic pragmatic studies will have to rely on comparable (spoken) corpora, and hence will need some reassurance that the expressions being examined are in fact comparable across the languages.

Finally, now that corpus-based contrastive studies have been with us for more than 25 years, it may be time to take a diachronic perspective. In a paper first presented at a workshop celebrating the 20th anniversary of the Nordic parallel corpus project,[17] Signe Ebeling calls for the compilation of

> parallel corpora matching the existing ones in terms of content and structure, but comprising texts of a more recent date. In a similar fashion to what has been done for the LOB and Brown corpora – with FLOB and Frown – a carefully designed ENPC 20 years on would pave the way for a new field of diachronic corpus-based contrastive studies, ensuring that such studies can be carried out in a systematic way. (S.O. Ebeling 2016: 51)

Considering the increasing amount of global media, migration and international travel, the need for insights into cross-linguistic matters is unlikely to diminish. Multilingual corpora will continue to be an invaluable source of cross-linguistic and cross-cultural information. As Johansson concludes in one of his last papers (2012: 65), "if used with care and imagination, multilingual corpora lead us beyond what we knew or did not see so clearly. This is the essence of the cross-linguistic perspective."

References

Aijmer, K. 2008. Parallel and comparable corpora. In *Corpus Linguistics. An International Handbook, Vol. 1,* A. Lüdeling and M. Kytö (eds), 275-292. Berlin/New York: Walter de Gruyter.

Aijmer, K. and Altenberg, B. (eds). 2013a. *Advances in Corpus-based Contrastive Linguistics. Studies in Honour of Stig Johansson*. Amsterdam: Benjamins.

Aijmer, K. and Altenberg, B. 2013b. Introduction. In K. Aijmer and B. Altenberg (eds.), 1-6.

Aijmer, K., Altenberg, B. and Johansson, M. (eds). 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.

Aijmer, K and Lewis, D. (eds). 2017. *Contrastive Analysis of Discourse-pragmatic Aspects of Linguistic Genres*. Springer.

Aijmer, K. and Simon-Vandenbergen, A. (eds). 2006. *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In *Out of Corpora. Studies in honour of Stig Johansson,* H. Hasselgård & S. Oksefjell (eds), 249-268. Amsterdam: Rodopi.

Altenberg, B. and Granger, S. 2002. Recent trends in cross-linguistic lexical studies. In *Lexis in Contrast: Corpus-based approaches,* B. Altenberg and S. Granger (eds), 3-48. Amsterdam: Benjamins.

[17] Papers from the workshop were published in Nordrum et al. (2016). Contributors were encouraged to present and/or suggest new approaches to parallel corpus research, to "contribute to the pitch of the future key of corpus-based contrastive linguistics" (Nordrum et al. 2016: 6).

Barlow, M. 1995. ParaConc: A concordancer for parallel texts. *Computers and Texts* 10, 14-16.

Barlow, M. 2008. Parallel texts and corpus-based contrastive analysis. In *Current Trends in Contrastive Linguistics: Functional and cognitive perspectives*, Gómez González, M. Á., Mackenzie, J.L., and González Álvarez, E.M. (eds), 101-121. Amsterdam: Benjamins.

Bernardini, S. 2015. Translation. In *The Cambridge Handbook of English Corpus Linguistics*, D. Biber and R. Reppen (eds), 515-536. Cambridge: Cambridge University Press.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 2009. *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

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

Čermák, F. and Rosen A. 2012. The Case of InterCorp, a multilingual corpus. *International Journal of Corpus Linguistics* 17: 3, 411-427.

Čermáková, A. and Chlumská, L. 2017. Expressing PLACE in children's literature: Testing the limits of the n-gram method in contrastive linguistics. In Egan & Dirdal (eds), 75-86.

Chlumská, L. and Lukeš, D. 2018. Comparing the incomparable? Rethinking n-grams for free word-order languages. In Granger et al. (eds), 40-41.

Cignoni, L., Coffey, S. and Moon, R. 1999. Idiom variation in Italian and English: Two corpus-based studies. *Languages in Contrast* 2:2, 279-300.

Cortes, V. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1), 43-57.

Cosme, C. and Gilquin, G. 2008. Free and bound prepositions in a contrastive perspective. In *Phraseology: An Interdisciplinary Perspective*, S. Granger and F. Meunier (eds), 259-274. Amsterdam/Philadelphia: Benjamins.

Coussé, E. and van der Auwera, J. 2012. Human impersonal pronouns in Swedish and Dutch: A contrastive study of *man* and *men*. *Languages in Contrast* 12:2, 121-138.

Crible, L. and Degand, L. 2019. Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory* 15(1): 71-99

Defranq, B. 2015. Contrasting contrastive approaches. *Languages in Contrast* 15:1, 1-3.

Dickens, A. and Salkie, R. 1996. Comparing bilingual dictionaries with a parallel corpus. In *Euralex 96 Proceedings*, 551-559. http://euralex.org/category/publications/euralex-1996-2/ (Last accessed May 2019).

Doval, I. and Sánchez Nieto, M. Teresa. 2019. *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*. Amsterdam/Philadelphia: Benjamins.

Dušková, L. 2015. *From Syntax to Text. The Janus Face of Functional Sentence Perspective*. Prague: Karolinum Press.

Dyvik, H. 2005. Translations as a semantic knowledge source. *Proceedings of the Second Baltic Conference on Human Language Technologies*, 27-38. Tallinn: Institute of Cybernetics at Tallinn University of Technology, Institute of the Estonian Language.

Ebeling, J. 1998. The Translation Corpus Explorer: A browser for parallel texts. In Johansson and Oksefjell (eds), 101-112.

Ebeling, J. 2016. Contrastive linguistics in a new key. *Languages in Contrast 20 Years on*. Special issue of *Nordic Journal of English Studies* 15(3): 7-14.

Ebeling, J. and Ebeling, S.O. 2013. *Patterns in Contrast.* Amsterdam and Philadelphia: Benjamins.

Ebeling, J. and Ebeling, S.O. 2015. An English-Norwegian contrastive analysis of downtoners, more or less. *Nordic Journal of English Studies* 14(1): 62-89.

Ebeling, J. and Ebeling, S.O. 2017. A cross-linguistic comparison of recurrent word-combinations in a comparable corpus of English and Norwegian fiction. In Janebová et al. (eds), 2-31.

Ebeling, J., Ebeling, S.O. and Hasselgård, H. 2013. Using recurrent word-combinations to explore cross-linguistic differences. In K. Aijmer and B. Altenberg (eds), 177-199. Amsterdam: Benjamins.

Ebeling, S.O. 2016. Does corpus size matter? Revisiting ENPC case studies with an extended version of the corpus. *Nordic Journal of English Studies* 15:3, 33-54.

Ebeling, S.O. 2018. Exploring cross-linguistic congruence: The case of two stance frames in English and Norwegian. *Bergen Language and Linguistics Studies* (BeLLS). 9(1), 69- 92. https://bells.uib.no.

Ebeling, S.O. and Ebeling, J. 2013b. From Babylon to Bergen: On the usefulness of aligned texts. *Bergen Language and Linguistics Studies* (BeLLS) 3:1, 23-42. https://bells.uib.no.

Ebeling, S.O. and Ebeling, J. 2014. *For Pete's sake!* A corpus-based contrastive study of the English/Norwegian patterns "for * sake" /*for * skyld*. *Languages in Contrast* 14:2, 191-213.

Egan, T. 2013. *Between* and *through* revisited. In *Corpus Linguistics and Variation in English: Focus on Non-Native Englishes*, M. Huber and J. Mukherjee (eds). *Studies in Variation, Contacts and Change in English* Vol 13. http://www.helsinki.fi/varieng/series/volumes/13/ (Last accessed May 2019)

Egan, T. and Dirdal, H. (eds). 2017. *Cross-linguistic Correspondences. From Lexis to Genre*. Amsterdam/Philadelphia: Benjamins.

Evert, S., and Neumann, S. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In *Empirical Translation Studies*, G. D. Sutter, M.-A. Lefer, and I. Delaere (eds), 47-80. Berlin, Boston: De Gruyter.

Filipović, R. 1984. What are the primary data for contrastive analysis? In Fisiak (ed.), 107-117.

Fisiak, J. (ed.) 1984. *Contrastive Linguistics. Prospects and Problems.* Berlin: Mouton.

Fløttum, K., T. Dahl and T. Kinn. 2006. *Academic Voices*. Amsterdam: Benjamins.

Frankenberg-Garcia, A. and Santos, D. 2003. Introducing *Compara*, the Portuguese-English Parallel Corpus. In *Corpora in Translator Education*, F. Zanettin, S. Bernardini and D. Stewart (eds), 71-88. Manchester: St Jerome.

Gellerstam, M. 1986. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia*, L. Wollin and H. Lindquist (eds), 88-95. Lund: CWK Gleerup.

Gilquin, G. 2000/2001. The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3:1, 95-123.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 37-51. Lund: Lund University Press.

Granger, S. 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14.1, 58-72.

Granger, S., Lefer, M.-A. and Aguiar de Souza Penha Marion, L. (eds). Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition). CECL Papers 1. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/cecl-papers.html (Last accessed February 2020).

Gries, S.Th., Miglio, V.G.and Jansegers, M. To appear. Quantitative methods for corpus-based contrastive linguistics. Retrieved from https://www.researchgate.net/publication/332120477_Quantitative_methods_for_corpus-based_contrastive_linguistics (Last accessed February 2020).

Hansen-Schirra, S. and Neumann, S. 2012. Corpus enrichment, representation, exploitation, and quality control. In Hansen-Schirra et al. (eds), 35-52.

Hansen-Schirra, S., Neumann, S., Steiner, E. (eds). 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German.* Berlin: de Gruyter.

Hasselgård, H. 2016. *The way of the world*: The colligational framework "the N1 of the N2" and its Norwegian correspondences. *Nordic Journal of English Studies* 15:3, 55-79.

Hasselgård, H. 2017a. Temporal expressions in English and Norwegian. In Janebová et al. (eds), 75-101.

Hasselgård, H. 2017b. Lexical patterns of PLACE in English and Norwegian. In Dirdal and Egan (eds), 97-119.

James, C. 1980. *Contrastive Analysis*. London: Longman.

Janebová, M. Lapshinova-Koltunski, E. and Martinková, M. (eds). 2017. *Contrasting English and other Languages through Corpora*. Newcastle: Cambridge Scholars Publishing.

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

Johansson, S. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies.* Amsterdam and Philadelphia: Benjamins.

Johansson, S. 2011. A multilingual outlook of corpora studies. In *Perspectives on Corpus Linguistics,* V. Viana, S. Zyngier and G. Barnbrook (eds.), 115-129. Amsterdam/ Philadelphia: Benjamins.

Johansson, S. 2012. Cross-linguistic perspectives. In *English Corpus Linguistics: Crossing Paths,* M. Kytö (ed.), 45-68. Amsterdam: Rodopi.

Johansson, Stig, Ebeling, J. and Oksefjell, S. 1999/2002. *English-Norwegian Parallel Corpus: Manual*. https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/ENPCmanual.pdf (Last accessed February 2020).

Johansson, S. and Hofland, K. 1994. Towards an English-Norwegian Parallel Corpus. In *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993,* U. Fries, G. Tottie and P. Schneider (eds), 25-37. Amsterdam: Rodopi.

Johansson, S. and Oksefjell, S. (eds). 1998. *Corpora and Cross-linguistic Research.* Amsterdam: Rodopi.

Kefala, S. and Sidiropoulou, M. 2016. Shaping the glo/cal in Greek–English tourism advertising. A critical cosmopolitan perspective. *Languages in Contrast* 16:2, 191-212.

Kirk, J. and Čermáková, A. 2017. From ICE to ICC: The new International Comparable Corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP), Birmingham, 24 July 2017,* P. Bański, M. Kupietz, H. Lüngen, P. Rayson, H. Biber, E. Breiteneder, S. Clematide, J. Mariani, M. Stevenson, T. Sick (eds), 7-12. Mannheim: Institut für Deutsche Sprache. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6249/file/Kirk_Cermakova_From_ICE_to_ICC_2017.pdf (Last accessed February 2020)..

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005. http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf (Last accessed February 2020).

Kong, K.C. 2013. A corpus-based study in comparing the multimodality of Chinese- and English- language newspapers. *Visual Communication*, *12*(2), 173-196. https://doi.org/10.1177/1470357212471594 (Last accessed February 2020).

König, E. 2012. Contrastive linguistics and language comparison. *Languages in Contrast* 12:1, 3-26.

Kunnskapsforlaget. 2001. *Engelsk Stor Ordbok: engelsk-norsk / norsk-engelsk.* Oslo: Kunnskapsforlaget.

Kunz, K., Lapshinova-Koltunski, E., Martínez Martínez, J.M., Menzel, K. and Steiner, E. 2018. Shallow features as indicators of English–German contrasts in lexical cohesion. *Languages in Contrast* 18: 2, 175-206.

Lado, R. 1957. *Linguistics across Cultures*. Ann Arbor: The University of Michigan Press.

Lauridsen, K. 1996. Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In K. Aijmer et al. (eds), 63-71.

Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam/New York: Rodopi.

Lefer, M.-A. and Vogeleer, S. (eds). 2014. *Genre- and Register-related Discourse Features in Contrast*. Special issue of *Languages in Contrast*, 14:1.

López Arroyo, B. and Roberts, R.P. 2015. Unusual sentence structure in wine tasting notes. A contrastive corpus-based study. *Languages in Contrast* 15:2, 162-180.

Macken, L., De Clercq, O. and Paulussen, H. 2011. Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta*, *56* (2), 374–390. https://doi.org/10.7202/1006182ar (Last accessed February 2020).

Maia, B. and Santos, D. 2012. "Who's afraid of ... what?" – in English and Portuguese. In *Aspects of Corpus Linguistics: Compilation, Annotation, Analysis*, S.O. Ebeling, J. Ebeling and H. Hasselgård (eds). *Studies in Variation, Contacts and Change in English*, Vol. 12. http://www.helsinki.fi/varieng/series/volumes/12/

Mair, C. 2018. Contrastive analysis in linguistics. In *Oxford Bibliographies in Linguistics*, M. Aronoff (ed.), Oxford/New York: Oxford University Press. DOI: 10.1093/OBO/9780199772810-0214.

Mauranen, A. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2:2, 161-187.

McEnery, T. and Hardie, A. 2012. *Corpus Linguistics*. Cambridge: Cambridge University Press.

Nordrum, L., Ebeling, S.O. and Hasselgård H. (eds). 2016. *Languages in Contrast 20 Years on*. Special issue of *Nordic Journal of English Studies* 15:3.

Oakes, M. & T. McEnery. 2000. Bilingual text alignment – an overview. In *Multilingual Corpora in Teaching and Research,* S.P. Botley, A.M. McEnery and A. Wilson (eds), 1-37. Amsterdam: Rodopi.

O'Keeffe, A. 2017. Corpus-based function-to-form approaches. In *Methods in Pragmatics*, A.H. Jucker, K.P. Schneider and W. Bublitz (eds), 587-618. Berlin/ Boston: De Gruyter Mouton.

Salkie, R., Aijmer, K. and Barlow, M. 1998. Editorial. *Languages in Contrast* 1:1, v-xii.

Sanjurjo-González, H. and Izquierdo, M. 2019. P-ACTRES 2.0: A parallel corpus for cross-linguistic research. In Doval & Sanchez Nieto (eds), 215-232.

Šebestová, D. and Malá, M. 2018. Testing the contrastive application of the n-gram method to typologically different languages: The case of English and Czech children's literature. In Granger et al. (eds), 157-158.

Sinclair, J. 2004. *Trust the Text. Language, Corpus and Discourse*. London and New York: Routledge.

Šinkūnienė, J. 2017. Corpora and corpus linguistics revisited: an interview with Karin Aijmer", *Kalbotyra*, 70, 184-191. doi: 10.15388/Klbt.2017.11202.

Ström Herold, J. and Levin, M. 2019. The Obama presidency, the Macintosh keyboard and the Norway fiasco: English proper noun modifiers and their German and Swedish correspondences. *English Language and Linguistics* 23, 827-854.

Taboada, M., Carretero, M. and Hinnell, J. 2014. Loving and hating the movies in English, German and Spanish. *Languages in Contrast* 14:1, 127-161.

Tognini-Bonelli, E. 1996. Towards translation equivalence from a corpus linguistics perspective. *International Journal of Lexicography*, 9:3, 197-217.

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

van der Auwera, J. 2012. From contrastive linguistics to linguistic typology. *Languages in Contrast* 12:1, 69-86.

Viberg, Å. 2013. Posture verbs: A multilingual contrastive study. *Languages in Contrast* 13:2, 139-169.

Viberg, Å. 2017. Saying, talking and telling: Basic verbal communication verbs in Swedish and English. In Egan & Dirdal (eds), 37-75.

Wallis, S. 2007. Annotation, retrieval and experimentation. Or: you only get out what you put in. In *Annotating Variation and Change*, A. Meurman-Solin and A. Nurmi (eds). *Studies in Variation, Contacts and Change in English*, Vol 1. http://www.helsinki.fi/varieng/series/volumes/01/

Xiao, R. and T. McEnery. 2010. *Corpus-based Contrastive Studies of English and Chinese*. New York/London: Routledge.

Zanettin, F. 2011. Translation and corpus design. *SYNAPS – A Journal of Professional Communication* 26, 14-23. https://openaccess.nhh.no/nhh-xmlui/handle/11250/2394011.