Contrastive linguistics in a new key[1]

Jarle Ebeling, University of Oslo

In 1994 a group of researchers from the Nordic countries met in Lund, Sweden, for a symposium titled *Languages in Contrast: A Symposium on Text-based Cross-linguistic Studies*.[2] The purpose of the symposium was to gather researchers "active in the field of contrastive linguistics and who share an interest in linking cross-linguistic studies with corpus linguistics and the use of the computer as a tool in linguistic research" (Aijmer et al. (eds) 1996, Acknowledgments). Little did we know back then that we would meet again in Lund in 2014 to commemorate the symposium and celebrate the start of the collaborative efforts that led to the compilation of the parallel corpora known as the English-Norwegian Parallel Corpus (ENPC), the Swedish-English Parallel Corpus (ESPC) and the Finnish-English Contrastive Corpus Studies Project (FECCS).[3] Moreover, in 2014, we could also celebrate the 20th anniversary of the research network "Languages in contrast", which was generously funded by the Nordic Academy for Advanced Study (NorFa) between 1994-1996, and which meant that the parallel corpus projects could employ research assistants and a software developer (Knut Hofland) to work on the compilation of the corpora.[4]

[1] The phrase "contrastive linguistics in a new key" is borrowed from Johansson (2012: 46) who says of present-day contrastive linguistics "that the focus on immediate applications is toned down; the contrastive study is text-based rather than a comparison of systems in the abstract; and the study draws on electronic corpora and the use of computational tools".

[2] The only non-Nordic participant was Sylviane Granger from the Centre d'Études Anglaises, Université Catholique de Louvain. Sylviane Granger later became one of the compilers of the PLECI corpus, built on the same model as the ENPC/ESPC. See http://www.uclouvain.be/en-cecl-pleci.html.

[3] A Danish parallel corpus project was also proposed, but never materialised.

[4] Stig Johansson and Knut Hofland had already run experiments on the alignment of parallel texts in 1993 (Johansson & Hofland 1994), and the Swedish project "Text-based Contrastive Studies in English" had also started in 1993 (Aijmer et al. 1996).

Ebeling, Jarle. 2016. "Contrastive linguistics in a new key." *Nordic Journal of English Studies* 15(3):7-14.

Looking back, one may ask why the time seemed ripe for corpus-based contrastive studies based on parallel text corpora in 1994. For to be fair, parallel texts have existed almost since the dawn of writing, as evidenced by e.g. clay tablets with Sumerian and Akkadian aligned cuneiform texts from Assyria from the 2nd millennium BCE, let alone famous inscriptions such as the Rosetta Stone (see Véronis 2000 and Ebeling & Ebeling 2013). Furthermore, contrastive linguistics was not exactly a new discipline in 1994. Two years earlier, for instance, the 25th International Conference on Contrastive Linguistics and Cross Language Studies was held at Adam Mickiewicz University in Rydzyna, Poland, bearing evidence of the strong tradition of contrastive linguistics in the then so-called Eastern European countries (cf. e.g. Fisiak 1983 and Mair & Markus (eds) 1992). In fact, as early as 1970, a research group in former Yugoslavia translated parts of the Brown corpus and aligned the source (original) and target (translation) texts and built a parallel corpus (see e.g. Filipović (ed.) 1971). In the same decade, Jan Svartvik, one of the pioneers of corpus linguistics, initiated a project titled The Swedish-English Contrastive Studies project (Svartvik 1973) at the University of Lund.

Several important trends and developments in the late 1980s and early 1990s heralded the advent of parallel corpora and corpus-based contrastive analysis (cf. Hartmann 1997). In the late 1980s corpus linguistics was firmly established as a research paradigm, but had hitherto mostly been preoccupied with compiling monolingual corpora and building tools for exploring them. An irresistible challenge for a corpus linguist such as Stig Johansson, who had also been interested in contrastive linguistics since the 1970s (Johansson 1975), was to see if some of the techniques and tools developed for monolingual corpora could extend to bi- and multilingual corpora. Kari Sajavaara certainly thought so, but envisaged that it would take some time to develop these tools and techniques.

> It is no longer necessary for the contrastive linguist to invent examples. It is now possible to resort to corpora, where the relevant instances can be found by means of automatic searchers [reference omitted]. There is a wealth of information about principles to be applied in the compilation of unilingual text corpora [references omitted], but there is much less information about parallel corpora [references omitted]. Since a bilingual parallel corpus is different from a unilingual one, it is to be assumed that the principles for its compilation are also different. It is evident that

> within the next few years we will have more information about this problem. (Sajavaara 1996: 31)

About the same time as contrastive linguistics turned to corpora for new impetus, so did translation studies. In his 1995 book, Gideon Toury argues strongly for a descriptive, empirically based approach to translation studies.[5]

> [...], as I see it, one of the weaknesses of Translation Studies in the present phase of its evolution lies precisely in the fact that descriptivism as such is frowned upon, driving every other scholar to indulge in theorizing, very often in a highly speculative manner. (Toury 1995: 266)

The empirical basis for which Toury made a case came in the form of parallel and translation corpora of the kind developed by e.g. Mona Baker (Baker 1993).[6] A few years earlier, the translation scholar Brian Harris had written a short piece on what he called bi-texts, which he saw as a new concept in translation theory.[7] To Harris, a bi-text should be thought of not only as a complete source text (ST, original) and a target text (TT, its translation), but also as combinations of words and segments within the two texts, since

> translators do not translate whole texts at one fell swoop. They proceed a little at a time, and as they proceed each spurt, each segment forms a fragment of bi-text in their mind. Bi-text retains this structure when it is recorded on paper or in a computer: that is to say, not only is the whole text a bi-text but each segment combines ST and TT. (Harris 1988: 8)

These bi-texts could then be stored electronically and make up a hypertext base or translation memory, where the translator could search his or her own previous work. Such a hyper-bi-text system should

[5] In 1969 Hartmann writes: "I must repeat my plea for more empirical research into translation within the frameworks provided by applied and contrastive linguistics (Hartmann 1969 [2007: 24]).

[6] In July 1993, Mona Baker wrote to Stig Johansson saying that John Sinclair had showed her a copy of Stig's proposal for an English-Norwegian corpus, and that she hoped to be setting up her own corpus of translated texts soon at The University of Manchester Institute of Science and Technology (UMIST).

[7] Harris's ideas have also been reiterated in Véronis (2000) and Ebeling & Ebeling (2013).

include a search engine "programmed in such a way that when it finds an occurrence of the word [one is looking for], it retrieves and displays the whole translation unit in which it occurs" (ibid. 9).

Without specifying how, Harris foresees an even more sophisticated system whereby one can search for units similar to the one one is after, i.e. a system where what is combined is not translation units, but meaning units.

As to the display of bi-texts Harris advocates an interlinear translation display, where each line of the target text is interlaced between the corresponding line(s) of source text making up an interlinear bi-text. However, Harris (ibid. 11) adds that "[t]here remains the serious problem with interlinear bi-text, that of aligning ST and TT."

The serious problem mentioned by Harris was soon tackled by computer scientists interested in natural language processing.[8] In a series of experiments in the early 1990s Gale and Church showed how parallel texts could be automatically aligned at sentence level using sentence length in characters on text pairs already aligned at paragraph level (Gale & Church 1991, 1993). The method, though very simple, was surprisingly successful:

> The model was motivated by the observation that longer regions of text tend to have longer translations, and that shorter regions of text tend to have shorter translations. In particular, we found that the correlation between length of paragraph in characters and the length of its translation was extremely high (0.991). This high correlation suggests that length might be a strong clue for sentence alignment. (Gale & Church 1993: 89)

At the 14th ICAME conference in Zürich in 1993, Stig Johansson and Knut Hofland proposed a method for aligning English and Norwegian source (original) and target (translation) texts which can be seen as an amalgamation of several of the methods tried out in earlier experiments (see Johansson & Hofland 1994). Johansson and Hofland, though recognising the effectiveness of using sentence length in characters, wanted to explore a more linguistically-grounded method of aligning a pair of texts at sentence level and proposed to include an anchor word list as a central component. The anchor word list contains words and expressions "where the correspondence between the languages could be

[8] See Véronis (2000) for an overview of the many alignment methods explored and experiments performed at the time.

expected to be rather good" (Johansson & Hofland 1994: 30), and the object of using anchor words "was to calculate an *anchor score* which can be used in sentence alignment, expressing the number of shared anchor words" (ibid. 31).

The bilingual anchor word lists we use today contain function words, numerals, frequent and stable content words and names of days, months, countries and languages. In addition to sentence length in terms of characters and anchor words, the latest version of the alignment program, the Translation Corpus Aligner (TCA v.2) also takes special characters, e.g. %, ?, !, matching proper names and cognate words in the two languages into consideration when aligning sentences. Such bilingual word lists, of approx. 2,000 lines, have now been made for a range of language pairs.

In an investigation of the effect of the anchor word list carried out for the language pair English-Portuguese (Santos & Oksefjell 1999) it was shown that the anchor word list is essential to the success of the alignment. In fact, six out of the sixteen texts used in the investigation could not be aligned without the anchor word list, thus suggesting that a language-dependent aligner such as the Translation Corpus Aligner (Hofland & Johansson 1998) was an important step in the development of aligners for parallel texts.

As to the idea of performing contrastive analysis based on original and translated texts pitted against each other in such as fashion that similarities and differences become apparent, this owes a lot to James' (1980) belief in translation as the best *tertium comparationis*, Hartmann's (1980: 37-38) classification of types of parallel texts, and the method known as analytical comparison (e.g. Mathesius 1975), which is straightforwardly described by Čmejrková (1992: 5-6) in this way:

> As long as we knew only one language, it seemed to be the only one, writes the Czech typologist Vladimír Skalička [reference omitted]. The mother tongue seems to be the most natural, practical, and beautiful one. But awareness of foreign language solutions raises the awareness of the mother tongue and the speaker consequently asks questions he would have not asked otherwise.

We are now in a position to answer the question raised at the beginning of the second paragraph of this résumé, as to why the time was ripe for contrastive studies in a new key based on parallel corpora in 1994. It was, as is often the case, the coming together of people, ideas and technology. The people, in our context, were above all Stig Johansson, Bengt Altenberg, Karin Aijmer and Knut Hofland.[9] As to the ideas, I have attempted to point to some of their sources, knowing quite well that great injustice has been done to many linguists and other scientists, who have had, and acted on, similar ideas over the years. Finally, technology: the corpora we began compiling in earnest in 1994 could hardly have been built without the technological advances that had taken place in the 1970s and 80s, with, e.g., the advent of the personal computer and the emerging new disciplines of computational linguistics and natural language processing.

Today we know that work on parallel corpora did not stop when ENPC/ESPC/FECCS were completed: several new copora have emerged, often including other language pairs than English and the Nordic languages. Moreover, corpus-based contrastive linguistics as a new discipline has gained momemtum with for example the pre-conference workshop as a fixture of ICAME conferences.

References

Aijmer, Karin, Bengt Altenberg and Mats Johansson (eds). 1996. *Papers from a Symposium on Text-based Cross-linguistic Studies. Lund 4-5 March 1994* [Lund Studies in English 88]. Lund: Lund University Press.

Aijmer, Karin, Bengt Altenberg and Mats Johansson. 1996. Text-based Contrastive Studies in English. Presentation of a Project. In Aijmer et al. (eds), 87–112.

Baker, Mona. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. 233–250.

Čmejrková, Světla. 1992. Linguistic Comparison and Linguistic Characterology. In Christian Mair and Manfred Markus (eds), 3-8.

Ebeling, Signe O. and Jarle Ebeling. 2013. From Babylon to Bergen. On the Usefulness of Aligned Texts. *Bergen Language and Linguistics Studies* (BELLS), 3:1. <DOI: http://dx.doi.org/10.15845/bells.v3i1>

[9] Mention should also be made of the many research assistants, including the current author, who worked hard on the various corpus projects, scanning, proof-reading and aligning texts.

Filipović, Rudolf. (ed.). 1971. *Zagreb Conference on English Contrastive Projects, 7-9 December 1970. Papers and Discussion*. Zagreb: Institute of Linguistics, Faculty of Philosophy, University of Zagreb. (Also *The Yugoslav Serbo-Croatian-English Contrastive Project, B. Studies 4*. Zagreb: Institute of Linguistics, Faculty of Philosophy, University of Zagreb.)

Fisiak, Jacek. 1983. Present Trends in Contrastive Linguistics. In Kari Sajavaara (ed.), *Cross-language Analysis and Second Language Acquisition* 1 [Jyväskylä Cross-languag Sudies 9]. Jyväskylä: University of Jyväskylä. 9–38.

Gale, William A. and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (ACL), Berkeley. 177–184.

Gale, William A. and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19:3. 75–102.

Harris, Brian. 1988. Bitexts: A New Concept in Translation Theory. *Language Monthly* 54. 8–10.

Hartmann, Reinhard R.K. 1980. *Contrastive Textology*. [Studies in Descriptive Linguistics 5]. Heidelberg: Julius Groos Verlag.

Hartmann, Reinhard R.K. 1997. From Contrastive Textology to Parallel Text Corpora: Theory and Applications. In Hickey, R. and S. Puppel (eds), *Language History and Linguistic Modelling. A Festschrift for Jacek Fisiak on his 60th Birthday. Volume II* [Trends in Linguistics. Studies and Monographs 101]. Berlin: Mouton de Gruyter. 1973–1987.

Hartmann, Reinhard R.K. 2007. *Interlingual Lexicography. Selected Essays on Translation Equivalence, Contrastive Linguistics and the Bilingual Dictionary*. Tübingen: Max Niemeyer Verlag. (Essays originally published between 1969 and 2005.)

Hofland, Knut and Stig Johansson. 1998. The Translation Corpus Aligner: A Program for Automatic Alignment of Parallel Texts. In Stig Johansson and Signe Oksefjell (eds), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 87–100.

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

Johansson, Stig. 1975. *Papers in Contrastive Linguistics and Language Testing* [Lund Studies in English 50]. Lund: Liber-Läromedel/Gleerup.

Johansson, Stig. 2012. Cross-linguistic Perspectives. In Merjä Kytö (ed*.), English Corpus Linguistics: Crossing Paths*. Amsterdam: Rodopi. 45–68.

Johansson, Stig and Knut Hofland. 1994. Towards an English-Norwegian Parallel Corpus. In Udo Fries, Gunnel Tottie, and Peter Schneider (eds), *Creating and Using English Language Corpora*, Amsterdam: Rodopi. 25–37.

Mair, Christian and Manfred Markus (eds). 1992. *New Departures in Contrastive Linguistics. Proceedings of the Conference Held at the Leopold-Franzens-University of Innsbuck, Austria, 10-12 May 1991*. Volume I & II.[Innsbrucker Beiträge zur Kulturwissenschaft. Anglistische Reihe Band 4/5]. Innsbruck: AMOE.

Mathesius, Vilém. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis*. The Hague: Mouton. (Translation into English by Josef Vachek of "Obsahový rozbor současné angličtiny na základě obecně lingvistickém".)

Sajavaara, Kari. 1996. New Challenges for Contrastive Linguistics. In Aijmer et al. (eds), 17–36.

Santos, Diana and Signe Oksefjell. 1999. An Evaluation of the Translation Corpus Aligner, with Special Reference to the Language Pair English-Portuguese. In Torbjørn Nordgård (ed.), *NODALIDA'99, Proceedings from the 12th "Nordiske datalingvistikkdager*". Trondheim: NTNU. 191–205.

Svartvik, Jan (ed.). 1973. *Errata. Papers in Error Analysis*. Lund: CWK Gleerup.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond* [Benjamins Translation Library 100]. Amsterdam: John Benjamins.

Véronis, Jean. 2000. From the Rosetta Stone to the Information Society: A Survey of Parallel Text Processing. In Jean Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers. 1–24.