Comparable and translation corpora in cross-linguistic research
Design, analysis and applications

Sylviane Granger
Centre for English Corpus Linguistics
Université catholique de Louvain

1. Introduction

The history of Contrastive Linguistics has been characterized by a pattern of success-decline-success. Contrastive Linguistics (CL) was originally a purely applied enterprise, aiming to produce more efficient foreign language teaching methods and tools. Based on the general assumption that difference equals difficulty, CL, which in those days was called Contrastive Analysis (CA), consisted in charting areas of similarity and difference between languages and basing the teaching syllabus on the contrastive findings. Advances in the understanding of Second Language Acquisition (SLA) mechanisms led to a questioning of the very basis of CA. Interlingual factors were found to be less prevalent than other factors, among which intralingual mechanisms such as the overgeneralization of target rules and external factors such as the influence of teaching methods or personal factors like motivation. This led to the decline of CA, but not to its death. At first, it gave rise to some drastic pedagogical decisions, which in some cases culminated in a total ban of the mother tongue in FL teaching. But research (see Odlin 1989, Selinker 1992, James 1998) re-established transfer as a major – if not **the** major – factor in SLA, which in turn led to a progressive – albeit limited – return of contrastive considerations in teaching. More importantly, the questioning of the contrastive approach to FL teaching did not impede its extension to other fields. The globalisation of society led to an increased awareness of the importance of interlingual and intercultural communication and played a major role in the revival of CL. Another factor which helped boost contrastive studies was the emergence and rapid development of corpus linguistics and natural language processing, which are increasingly focusing on cross-linguistic issues. Large bilingual corpora gave contrastive linguists and NLP specialists a much more solid empirical basis than had ever been previously available. Previous research had been largely intuition-based. Vinay & Darbelnet (1958/1995) and Malblanc (1968) are well-known exemplars of this type of approach. As the authors had an excellent knowledge of the languages they compared, these books contain a wealth of interesting contrastive statements. However, intuitions can be misleading and a few striking differences can lead to dangerous over-generalisations. For instance, the absence in English of connectors corresponding to the French ‘or’ or ‘en effet’ has led to the general conclusion that French favours explicit linking while English tends to leave links implicit (Vinay & Darbelnet 1958: 222, Newmark 1988: 59, Hervey & Higgins 1992: 49). Like many others, this contrastive claim still awaits empirical investigation. Contrastive linguists now have a way of testing and quantifying intuition-based contrastive statements in a body of empirical data that is vastly superior – both qualitatively and quantitatively – to the type of contrastive data that had hitherto been available to them.

The domain of Translation Studies (TS) underwent a similar corpus-based trend in the early 90s under the impetus of Mona Baker, who laid down the agenda for corpus-based TS (1993 and 1995) and started collecting corpora of translated texts with a view to uncovering the distinctive patterns of translation. Her investigations brought to light a number of potential ‘translation universals’ (Baker 1993) which further corpus studies are helping to confirm or

disprove (see Puurtinen 2007). Researchers in both CL and TS have thus come to rely on corpora to verify, refine or clarify theories that hitherto had had little or no empirical support and to achieve a higher degree of descriptive adequacy.

Section 2 gives an overview of the types of corpus used in cross-linguistic studies and suggests a unified terminology. Section 3 presents the different types of corpus-based comparison and section 4 highlights the respective advantages and disadvantages of bilingual comparable vs translation corpora. Section 5 gives a brief overview of some of the applications of corpus-based cross-linguistic research and the last section offers some concluding remarks.

2. Corpora in cross-linguistic research

In the corpus, scholars of contrastive linguistics and translation studies now have a common resource. Unfortunately, like in many new scientific fields, the terminology has not yet been firmly established, leading to a great deal of confusion.

Contrastive linguists distinguish between two main types of corpus for use in cross-linguistic research:

- corpora consisting of original texts in one language and their translations into one or more languages – let us call these *translation corpora*;
- corpora consisting of original texts in two or more languages, matched by criteria such as the time of composition, text category, intended audience, etc. – let us call these *comparable corpora*. (Johansson & Hasselgård 1999).

It should be noted however, that even among contrastive linguists the terminology is not entirely consistent. The term *parallel corpus* is sometimes used to refer to a comparable corpus (Aijmer et al 1996: 79, Schmied & Schäffler 1996: 41), a translation corpus (Hartmann 1980: 37) or a combined comparable/translation corpus (Johansson et al 1996).TS researchers, on the other hand, use the terms *translation corpus, parallel corpus* and *comparable corpus* to cover different types of texts. The term *comparable corpus* is used to refer to 'two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages' (Baker 1995: 234). The term *translation* (or *translational*) *corpus* is used to refer to the corpus of translated texts (see Baker 1999 and Puurtinen 2007). While in standard CL terminology, comparable corpora are usually multilingual (comparable original texts in different languages), in TS terminology they are usually monolingual (original and translated texts in the same language). Within the TS framework the term *parallel corpus* usually refers to 'corpora that contain a series of source texts aligned with their corresponding translations' (Malmkjaer 1998: 539), in other words what contrastive linguists usually refer to as translation corpora.

Over and above the terminological difference, there is a more fundamental discrepancy between the two cross-linguistic approaches. In the TS framework, translated texts are considered as texts in their own right, which are analysed in order to "understand what translation is and how it works" (Baker 1993: 243). In the CL framework they are often presented as unreliable as the cross-linguistic similarities and differences that they help establish may be 'distorted' by the translation process, i.e. may be the result of interference from the source texts.

Faced with the terminological diversity that characterises current cross-linguistic research, I feel that unified terminology is desirable and would like to suggest the general typology illustrated in Figure 1.

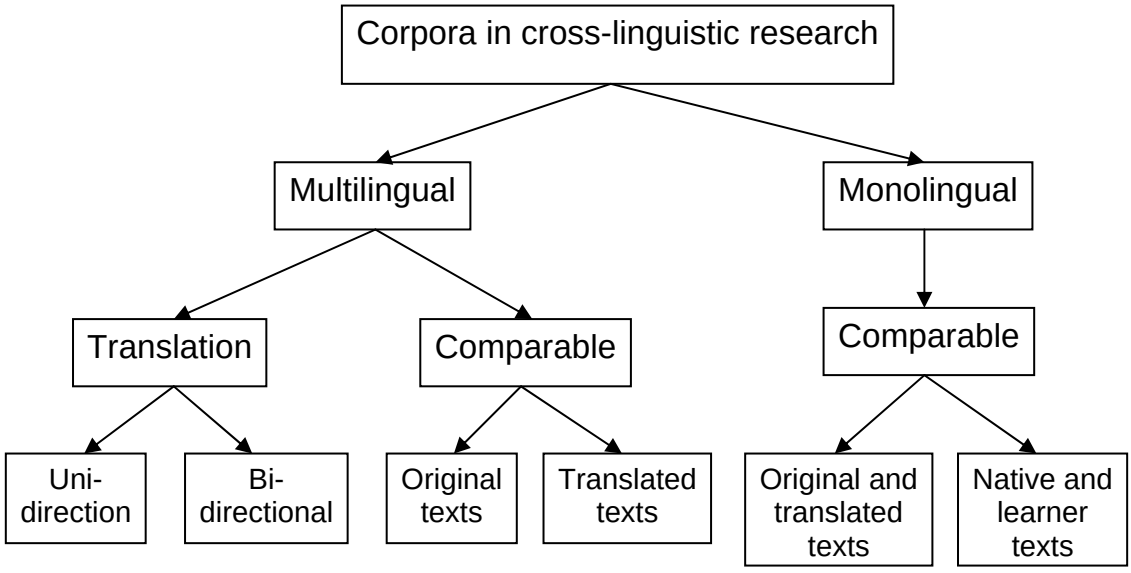

Figure 1: Corpora in cross-linguistic research

In this typology, a primary distinction is made between multilingual and monolingual corpora. *Multilingual corpora* involve more than one language. They may be of two main types: (a) *translation corpora* (which contain source texts and their translations and may be unidirectional – from language X to language Z – or bi/multidirectional) and (b) *comparable corpora* (which contain non-translated or translated texts of the same genre). The *monolingual corpora* relevant for cross-linguistic research are all comparable corpora. They may contain (a) *original and translated* texts in one and the same language or (b) *native and learner texts* in one and the same language[1]. In this typology, the term *parallel corpus* is not used in view of its ambiguity in the literature, where it has been used to refer to corpora of source texts and their translations, comparable corpora or as a generic term to refer to any type of multilingual corpus (Teubert 1996: 245).

This diagram does not include the many extralinguistic features that influence the data and therefore need to be carefully recorded, such as the translator's status (professional or student) or the direction of the translation process (into the translator's mother tongue or not).

3. Types of corpus-based comparison

With these different corpus types, a variety of comparisons can be undertaken. Table 1 presents an overview of the different types of cross-linguistic comparison and the disciplines within which they are undertaken (see also Johansson 2007a)

[1] For a description of this special type of contrastive research called *Contrastive Interlanguage Analysis*, see Granger 1996 and Gilquin 2000/2001.

	Type of comparison	Type of corpus	Discipline
1.	$OL_x \Leftrightarrow OL_y$	Multilingual comparable corpus of original texts	CL
2.	$SL_x \Leftrightarrow TL_y$	Multilingual translation corpus	CL & TS
3.	$SL_x \Leftrightarrow TL_x$	Monolingual comparable corpus of original and translated texts	TS & CL
4.	$TL_x \Leftrightarrow TL_y$	Multilingual comparable corpus of translated texts	TS

OL = original language
SL = source language
TL = translated language

Table 1: Types of corpus-based cross-linguistic comparison

The first type of comparison, between corpora of original texts in different languages (x and y), is the CL domain of expertise par excellence. However, there is a growing awareness among TS researchers of the interest of this type of research for translation studies. The second type of comparison is the most obvious meeting point between CL and TS. Researchers in both fields use the same resource but to different ends: uncovering differences and similarities between two (or more) languages for CL and capturing the distinctive features of the translation process and product for TS. The third type of comparison, which contrasts original and translated varieties of one and the same language, is the ideal method for uncovering the distinctive features of translated texts and hence seems at first sight to fall exclusively within TS. However, this type of comparison is increasingly being used by CL researchers who interpret differences between OL and TL as indirect evidence of differences between the languages involved (see Johansson & Hasselgård 1999 and Johansson 2007a). Finally, the comparison of translated varieties in different languages is quite clearly the prerogative of TS. However, it is essential that contrastive linguists pay attention to this type of study. Failing to properly understand the nature of translated texts might lead them to attribute some difference between OL and TL to interference from OL when in fact the phenomenon may simply be a manifestation of a translation universal.

4. Advantages and disadvantages of bilingual comparable and translation corpora

Table 2 summarizes the advantages and disadvantages of the two main types of multilingual corpus: the comparable corpus and the translation corpus. It appears clearly from the table that what constitutes an advantage for one type of corpus constitutes a disadvantage for the other and vice versa.

+ / -	**Translation corpora**	**Comparable corpora**
+	Text type comparability L1-L2 equivalence	Wide availability of texts Original language (reliable frequency and use)

-	Limited availability of texts Translated language (translationese & translation universals)	Text type comparability L1-L2 equivalence

Table 2: Bilingual translation vs comparable corpora

AVAILABILITY

The most easily accessible corpora for cross-linguistic research are undoubtedly comparable corpora of original languages. English is particularly well equipped with large balanced corpora such as the *British National Corpus* or the *Bank of English*. For other languages, there are electronic text collections, notably newspaper archives, that are regularly used for cross-linguistic research, but they tend to be less representative than the English mega corpora. Less widespread languages may not have any corpus resources at all or access to them may be severely limited. As regards translation corpora, however, electronic resources are scarce. It is not always possible to find translations of all texts, either because of the text type – letters and e-mail messages, for instance, are not usually translated – or because there are more translations in one direction (English to Chinese, for instance) than in another (Chinese to English). Available translation corpora tend to include older, copyright free texts (cf. project Gutenberg[2] which contains c. 30,000 free books) or alternatively, highly specialised texts such as documents from the European Union or the World Health Organization, the disadvantage of which is that it is often impossible to determine the source and target languages, a major variable for both CL and TS studies. While we are witnessing a rapid growth in the number of bilingual (and multilingual) resources, some of which can even be explored online, many high quality resources remain inaccessible to the academic community. This is the case, for instance, of the excellent English-Norwegian and English-Swedish corpora, which are only available to a limited group of researchers because of copyright restrictions.

TEXT TYPE COMPARABILITY

Translation corpora are an ideal resource for establishing equivalence between languages since they convey the same semantic content and are pragmatically and textually comparable (cf. James 1980: 178). In the case of comparable corpora, however, it is much more difficult to ensure text type comparability. Some types of text are culture-specific and simply have no exact equivalent in other languages. For example, when compiling the Lancaster Corpus of Mandarin Chinese (LCMC), McEnery & Xiao (2004) designed the corpus as an exact replica of the FLOB corpus to ensure comparability of the data. However, they encountered some difficulty, notably with the category of 'western and adventure fiction" which has no exact correspondent in Chinese and which they decided to replace by a category of 'martial arts fiction'.

L1-L2 EQUIVALENCE

Cross-linguistic comparison requires a "common platform of comparison" (Connor & Moreno 2005), a "background of sameness" (James 1980: 169) against which differences can be described. This constant, which is usually referred to as the *tertium comparationis*[3] *(TC)*, is relatively easy to establish in the case of translation corpora but constitutes a major stumbling block in the case of comparable corpora. In translation corpus studies, the TC is the relationship between a unit in the source language and its translation in the target language,

[2] Cf. http://www.gutenberg.org

[3] The term *tertium comparationis* has been used in a wide range of meanings in the contrastive literature. Connor & Moreno (2005), for instance, use the term TC for all levels of research, including the selection of corpora.

viz. translation equivalence. For example, in Aijmer's (1999) study of epistemic modality in English and Swedish, the TC is the relationship between the English modal verb *may* and the corpus-attested equivalents in Swedish (modal verbs, modal adverbs or a combination of the two). With comparable corpora, however, there is no readily available tertium comparationis. And yet, researchers need to establish one if they want to make sure that they will compare like with like. As regards grammar, James (1980: 167) reminds us that "the fact that we use the labels 'tense' or 'articles' to refer to a certain grammatical category in two different languages should not be taken to mean that we are talking about the same thing". It is therefore necessary to establish a basis for comparison. However, James (ibid: 168) hastens to point out that "comparability does not presuppose absolute identity, but merely a degree of shared similarity". In the case of articles, the TC could be "a small class of function words that occur in pronominal position and seem to indicate the specificness or genericness of the noun" (ibid: 168). This is a thorny issue whatever the languages involved but the problem is particularly acute in the case of very different language systems, such as English and Chinese (cf. McEnery & Xiao's 1999 comparison of aspect marking in English and Chinese). It is all the more important to establish a clear TC in areas such as phraseology where units such as idioms or collocations tend to be ill-defined.

RELIABILITY OF LANGUAGE

Comparable corpora have the major advantage of representing original texts in the two (or more) languages under comparison, i.e. language spontaneously produced by native speakers of those languages. They are therefore in principle free from the influence of other languages[4] and therefore arguably more reliable, especially to assess frequency and patterns of use. Translation corpora, on the other hand, display two main types of features that mark them off from original texts. On the one hand, they often contain features of what is usually referred to as 'translationese', i.e. "deviance in translated texts induced by the source language" (Johansson & Hofland 1994:26).[5] On the other hand, they also display universal features, i.e. "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993: 243). Gellerstam (1986) gives ample lexical evidence of translationese in translated Swedish. The main characteristics he lists are: a higher proportion of English loanwords, fewer colloquialisms, a higher frequency of standard 'press-the-button' translations of English words; and international words such as *lokal, massiv, drastic* used with new shades of meaning (for further examples of translationese, see Borin & Prütz 2001, Frankenberg-Garcia 2008, Wang & Qin 2008). In an interesting article, Rayson et al (2008) show how translationese can be detected fully automatically by comparing the frequencies of words and phrases in three ICT (Information and Communications Technology) corpora: a corpus of original Chinese texts, a corpus of translations of these texts into English by a proficient Chinese translator and a corpus of edited English, containing the versions of the Chinese translations corrected by a native speaker of English. The authors focus on multiword units and uncover interesting differences,

[4] This is obviously not entirely true. Newspaper texts, for example, have often been found to contain traces of the (usually) English texts on which the journalists have based their articles.

[5] The term 'translationese' is used in a range of meanings in contrastive and translation studies. It can be used in a neutral sense to refer to any source language-related feature that distinguishes translated language from original language or in a clearly negative sense to refer to features that result from "the translator's inexperience or lack of competence" (Baker 1993: 249). Gellerstam (1986: 88) uses it in the former meaning but explicitly excludes "anecdotal instances of bad translations".

such as the consistent replacement of the adjective *Chinese* by the genitive *China's* in edited English, in phrases such as *Chinese management software market.*[6]

Some linguists conclude that the balance weighs very much in favour of comparable corpora and advise linguists against using translation corpora. Teubert (1996: 247), for example, goes so far as claiming that "Translations, however good and near-perfect they may be (but rarely are), cannot but give a distorted picture of the language they represent. Linguists should never rely on translations when they are describing language (…). Rather than representing the language they are written it, they give a mirror image of their source language". Most contrastive linguists, however, are keen to point out that the two types of corpus should be used concurrently as each has its advantages and disadvantages. According to Johansson (2007a), one of the advantages of the combined use of translation and comparable corpora is that "the bidirectional translation model makes it possible to distinguish between language differences and translation effects". To this end, two different methods can be used: the first consists in identifying translation equivalents on the basis of translation corpora and then checking them against comparable corpus data; the other starts out with a contrastive description of patterns in each language on the basis of comparable corpus data and then studies translation correspondences using translation corpora. Johansson's (2007b) excellent book on multilingual corpora (2007b) contains a number of contrastive studies illustrating each method.

5. Applications of corpus-based cross-linguistic research

Any field that rests on the analysis of two or more languages can benefit from corpus-based cross-linguistic research. In Natural Language Processing research, the improvement of automated translation, notably via the creation and gradual update of translation memories (cf. Rayson et al 2008), is one of the most obvious beneficiaries. Lexicography, which now usually takes the form of electronic lexicography, also has a lot to gain. As stated by Teubert (1996: 241), "By exploiting corpora, bilingual and multilingual lexicography can reach a new quality level, a level that was just not possible without corpora". Lefer's (2009) corpus-based study of prefixation in English and French gives ample evidence of the heuristic power of the corpus approach. Her investigation shows that bilingual dictionaries often contain the exact same translations, with the exact same examples and would therefore undeniably benefit from the fresh blood brought in by authentic translation data (see also King 2007 and Granger & Paquot 2008).

In the field of foreign language teaching, the applications are numerous, both as regards pedagogical material and teaching methodology. For commercial reasons, major publishing houses have generally tended to produce generic tools, which target a wide range of EFL/ESL learners. This is the case, for example, with learners' dictionaries. As a result, L1-specific information is excluded, which strongly restricts their usefulness as a large number of learner problems are transfer-related. Now that electronic dictionaries are gradually replacing paper dictionaries, the generic vs. specific dichotomy loses some of its relevance. All dictionaries, whether primarily bilingual or monolingual, can benefit from the increased flexibility highlighted by Rundell (2007): "The old binary choices in reference publishing (monolingual or bilingual, dictionary or encyclopedia, advanced or intermediate) may no longer be relevant.

[6] We leave aside here cases where authors or translators use L1-modelled words or phrases to add a target language flavour to their writing. This is regularly the case in literary works (on the purposeful use of translationese in fiction, cf. Eberlein 2008).

Customization and personalization are likely new directions, so the current globally-marketed one-size-fits-all package will probably be unpicked". It has now become possible to include in ELT materials – grammars, dictionaries, textbooks, writing aids –both a core component that targets all users, whatever their language background, and L1-specific components that specifically address the difficulties of a particular learner population. The *Louvain EAP Dictionary (LEAD)*, which is currently under development, is entirely based on the analysis of English for Academic Purposes (EAP) words in both native and learner corpora. The dictionary allows for two types of customization: discipline (medicine, business, etc.) and learners' L1. The tool is truly bilingualized, with access to the entries via a translation of the EAP words in the learner's L1 and warnings to alert users to L1-specific error-prone items (Granger & Paquot 2009).

As regards teaching methodology, the most promising avenue is the addition of comparative concordance-based exercises to the battery of traditionally used exercises. Corpus-based analyses have led to a new inductive teaching methodology, called data-driven learning (DDL), which Johns and King (1991: iii) describe as "the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output". DDL can involve varied types of corpus data: monolingual and bilingual, native and learner – and be integrated into foreign language teaching – both general and for specific purposes (LSP) – and translator training (cf. Gilquin & Granger in press).

Aligned bilingual corpora can be used in consciousness-raising exercises: learners are presented with examples of language features such as modals, prepositions, conjuncts or pronouns in the source language and their aligned translations in the target language and asked to reflect on the interlingual similarities and differences. This inductive stage is usually followed by a series of activities, which take the form of corpus-based fill-in exercises where either the search item or the aligned translation has been blanked out. A good example of a DDL approach to grammar teaching is the *Online Chemnitz Internet Grammar*, which makes extensive use of the English-German Translation Corpus[7]. Kübler and Foucou (2007) demonstrate the important contribution of technical English corpora – both monolingual and bilingual (English-French) – in describing the use of verbs in Computer Science and preparing pedagogical material tailor-made for French-speaking ESP students. Bernardini (1997) points out the advantages of the method within the framework of translation training: "one of the reasons why translation teaching as it is generally understood (exercise and correction) is often perceived as ineffective and tentative, is that it still lacks a solid pedagogic background". In her view, traditional exercises in translation should go hand-in-hand with corpus-based learning activities which develop the skills that are immediately relevant for the education of translators, in terms of awareness, reflectiveness and resourcefulness. Corpus-based classroom activities for translator trainees may involve comparable and parallel corpora of general or specialised language. Zanettin (1998: 618-620) demonstrates how small comparable corpora of English and Italian help learners to compare the behaviour of similar discourse units in the two languages and select the translations which best adhere to the linguistic and genre conventions of the receiving culture. In order to test the effectiveness of corpus-based methods in translation training, Bowker (1998 & 2007) assigns the same translation task to two groups of translator trainees – one using a specialised monolingual corpus, the other conventional reference tools – and finds that the former outperformed the latter in several major respects, notably subject-field understanding, correct term choice and

[7] http://www.tu-chemnitz.de/phil/InternetGrammar/. See also Schmied (2009).

idiomatic expression. Admittedly, there is as yet little empirical evidence of this sort. In spite of this, Aston (1999) is optimistic about the future of corpora as translation and learning tools: "It is our experience at Forli that few trainee-translators who have used corpora would wish to be without them, notwithstanding (or because of?) the investment in time and effort required to compile corpora and to learn how to use them, and we expect that, as the number of available corpora and the quantity of suitable software increases, the use of corpora for translation and translator-training will gather further momentum, with a growth in its cost-effectiveness".

6. Conclusion

Multilingual corpora are revolutionizing the fields of both Contrastive Linguistics and Translation Studies. They provide both fields with the solid empirical foundation they need to enhance their descriptions and test their theoretical constructs as well as improve the cross-linguistic applications resulting from their respective research. The corpus has the potential to bring the two fields even closer together as both CL and TS researchers now rely on the same type of data, use the same software tools and are partly interested in the same corpus-based applications, notably reference materials – dictionaries, grammars – and teaching methods. Unfortunately, as rightly stated by Chesterman (1998: 6), "Although these are neighbouring disciplines, it nevertheless often appears that theoretical developments in one field are overlooked in the other, and that both would benefit from each other's insights". In particular, lack of familiarity with TS findings may lead CL researchers to interpret their data in terms of differences between language systems when they result from translation norms or strategies, while TS researchers may similarly misinterpret their data because of a lack of awareness of a systematic difference between the two language systems established by CL. Another more practical reason that should lead researchers in the two fields to cooperate is the shortage of corpora, which considerably hinders cross-linguistic research. We need more and better corpora for cross-linguistic research and as data collection is very time-consuming, there is a great deal to be gained from joining forces. If CL and TS pool knowledge and resources, one can safely predict a bright future for corpus-based cross-linguistic research and applications.

References

Aijmer, K. (1999). Epistemic possibility in an English-Swedish contrastive perspective. In Hasselgard, H. & Oksefjell, S. (eds.) *Out of Corpora*. Amsterdam & Atlanta: Rodopi, 301-323.

Aijmer K., B. Altenberg & M. Johansson (eds.) (1996). *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic* Studies. Lund Studies in English 88. Lund University Press: Lund.

Aijmer K., B. Altenberg & M. Johansson. (1996). Text-based contrastive studies in English. Presentation of a project. In Aijmer et al (eds.) *Languages in Contrast*: 73-85

Altenberg B. & S. Granger (eds.) (2002). *Lexis in Contrast. Corpus-based approaches*. Studies in Corpus Linguistics 7. Benjamins: Amsterdam & Philadelphia.

Aston G. (1999). Corpus use and learning to translate. *Textus* 12: 289-314. Also available online: http://www.sslmit.unibo.it/guy/textus.htm

Baker M. (1993). Corpus Linguistics and Translation Studies. Implications and Applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology*. Benjamins: Amsterdam & Philadelphia, 233-250.

Baker M. (1995) Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* 7: 2, 223-243.

Baker M. (1999). The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. *International Journal of Corpus Linguistics* 4:2, 281-298.
Bernardini S. (1997). A 'trainee' translator's perspective on corpora. Corpus use and learning to translate. Paper presented at the first international conference on Corpus Use and Learning to translate, Bertinoro, 14-15 November 1997.
Borin, L. & Prütz, K. (2001). Through a glass darkly: part of speech distribution in original and translated text. In Daelemans, W., Sima'an, K., Veenstra, J. & Zavrel, J. (eds.) *Computational linguistics in the Netherlands 2000*, Amsterdam: Rodopi,. 30–44.
Bowker L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. In S. Laviosa (ed.) *L'approche basée sur corpus/The Corpus-based Approach*: 631-651.
Bowker, L. (2007). Corpus-based applications for translator training: Exploring the possibilities. In Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Foreign Language Teaching and Research Press: Beijing, 169-183.
Chesterman A. (1998). *Contrastive Functional Analysis.* Benjamins: Amsterdam & Philadelphia.
Connor, U. & Moreno, 1. (2005). Tertium Comparationis: A vital component in contrastive research methodology. In Bruthiaux, P., Atkinson, D., Eggington, W.G., Grabe, W. & Ramanathan, V. (eds.) *Directions in Applied Linguistics: Essays in honor of Robert B. Kaplan.* Clevedon: Multilingual Matters, 153-164.
Eberlein, X. (2008). On translationese. Downloaded from http://www.insideoutchina.com
Frankenberg-Garcia, A. (2008). 'Suggesting rather special facts': a corpus-based study of distinctive lexical distributions in translated texts. *Corpora* 3/2, 195-211.
Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In Wollin, L. & Lindquist, H. (eds.) *Translation studies in Scandinavia*. Lund Studies in English 75. Malmö: CWK Gleerup, 88-95.
Gilquin G. (2000/2001) The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3(1): 95-123.
Gilquin, G. & Granger, S. (in press). How can DDL be used in language teaching? In A. O'Keeffe & M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
Granger S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual corpora and learner corpora. In Aijmer et al (eds.) *Languages in Contrast*: 37-51.
Granger S., L. Beheydt & J.P. Colson (eds.) (1999). *Contrastive Linguistics and Translation*. Special issue of *Le Langage et l'Homme* 34 :1. Peeters: Leuven.
Granger, S. & M. Paquot (2008b). From dictionary to phrasebook? In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, Barcelona, Spain, 15-19 July 2008, 1345-1355.
Granger, S. & Paquot, M. (2009). Customizing a general EAP dictionary to learner needs. Paper to be presented at the *eLexicography in the 21st century: New challenges, new applications (eLex2009)* conference, Louvain-la-Neuve, 22-24 October 2009.
Hartmann R. (1980). *Contrastive Textology*. Julius Groos Verlag: Heidelberg.
Hervey S. & J. Higgins. (1992). *Thinking Translation*. Routledge: London.
James, C. (1980). *Contrastive Analysis*. London & New York: Longman.
James C. (1998). *Errors in Language Learning and Use. Exploring error analysis*. Longman: London & New York.
Johansson, S. (2007a). Contrastive linguistics and corpora. In Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Foreign Language Teaching and Research Press: Beijing, 31-44.

Johansson, S. (2007b). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies.* Amsterdam & Philadelphia: Benjamins.
Johansson, S. & Hofland, K. (1994). Towards an English-Norwegian parallel corpus. In Fries, U., Tottie, G. & Schneider, P. (eds.) *Creating and using English language corpora*, 25-37. Amsterdam/Atlanta: Rodopi.
Johansson S., J. Ebeling & K. Hofland (1996). Coding and aligning the English-Norwegian Parallel Corpus. In Aijmer et al (eds.) *Languages in Contrast*: 87-112.
Johansson S. & H. Hasselgård (1999).Corpora and cross-linguistic research in the Nordic countries. In Granger et al (eds.) *Contrastive Linguistics and Translation*, 145-162.
Johns T. & P. King (eds.). (1991). *Classroom Concordancing*. ELR Journal (New Series) 4.
King P. (2007). Parallel concordancing and its applications. In Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Foreign Language Teaching and Research Press: Beijing, 157-167.
Kübler, N. & Foucou, P.-Y. (2007). Teaching English verbs with bilingual corpora. Examples in the field of computer science. In Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Foreign Language Teaching and Research Press: Beijing, 185-206.
Laviosa S. (1998). L'approche basée sur corpus/The Corpus-based Approach. Special issue of *META. Journal des Traducteurs* 43, 4, 473-659.
Lefer, M.-A. (2009). Exploring lexical morphology across languages. A corpus-based study of prefixation in English and French writing. Unpublished PhD dissertation. Université catholique de Louvain: Louvain-la-Neuve.
Malblanc A. (1968). *Stylistique comparée du français et de l'allemand*. Didier: Paris.
Malmkjaer K. (1998). Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? In S. Laviosa (ed.) *L'approche basée sur corpus/The Corpus-based Approach*: 534-541.
McEnery, T. & Xiao, R. (1999). Domains, text types, aspect marking and English-Chinese translation. *Languages in Contrast* 2, 2, 211-229
McEnery, T. & Xiao, R. (2004). The Lancaster Corpus of Mandarin Chinese. http://www.lancs.ac.uk/fass/projects/corpus/LCMC/
Newmark P. (1988). *A Textbook of Translation*. Prentice-Hall: Englewood Cliffs.
Odlin T. (1989). *Language Transfer. Cross-linguistic influence in language learning*. Cambridge University Press: Cambridge.
Puurtinen, T. (2007). Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals? In Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Foreign Language Teaching and Research Press: Beijing, 141-154.
Rayson, P., Xu, X,, Xiao, J., Wong, A. & Yuan, Q. (2008). Quantitative analysis of translation revision: contrastive corpus research on native English and Chinese translationese. In: *XVIII FIT World Congress*, August 4-7, 2008, Shanghai, China. Downloaded from http://eprints.comp.lancs.ac.uk/2042/1/Rayson_P_Et_Al_fit2008.pdf
Rundell, M. (2007). The dictionary of the future. In Granger, S. (ed.) *Optimizing the role of language in technology-enhanced learning*. Proceedings of the workshop organized at the Université catholique de Louvain, Louvain-la-Neuve, 4-5 October 2007.
Schmied, J. (2009). Learning English Prepositions in the Chemnitz Internet Grammar. In Granger, S. & Petch-Tyson, S. (eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges.* Language Teaching and Research Press: Beijing, 231-247.
Schmied J. & H. Schäffler (1996). Approaching translationese through parallel and translation corpora. In C. Percy, C. Meyer & I. Lancashire (eds.) *Synchronic Corpus Linguistics*. Rodopi: Amsterdam & Atlanta, 41-56.

Selinker L. (1992). *Rediscovering Interlanguage*. Longman: London & New York.
Teubert, W. (1996). Comparable or parallel corpora? *International Journal of Lexicography* 9: 238-264.
Vinay, J.P. & J. Darbelnet (1958). *Stylistique comparée du français et de l'anglais*. Montréal, Beauchemin.
Vinay, J.P. & J. Darbelnet (1995). *Comparative Stylistics of French and English*, trans. and ed. by Juan C. Sager and M.-J. Hamel. Amsterdam/Philadelphia : Benjamins.
Wang, K. & Qin, H. (2008). A parallel corpus-based study of translational Chinese. In Xiao, R., He, L. & Yue, M. (eds.) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008),* Zhejiang University, Hangzhou. Downloaded from http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Wang_and_Qin.pdf
Zanettin F. (1998). Bilingual Comparable Corpora and the Training of Translators.In S. Laviosa (ed.) *L'approche basée sur corpus/The Corpus-based Approach*: 616-630.

Additional Applications of Corpus-Based Research "Apart from the applications in linguistic research per se, the following practical applications may be mentioned. Lexicography Corpus-derived frequency lists and, more especially, concordances are establishing themselves as basic tools for the lexicographer. . . . Language Teaching . . . The use of concordances as language-learning tools is currently a major interest in computer-assisted language learning (CALL; see Johns 1986). . . . Speech Processing Machine translation is one example of the application of corpora for what computer scien