Translation and corpus design

Federico Zanettin

Università di Perugia

Summary
In this article I discuss the role of translated texts in different types of corpora. I first consider the role of translations in corpus-based monolingual linguistics, arguing that while translated texts are often excluded from corpora on the basis of a more or less implicit assumption that they "corrupt" the reference norm for a language, this assumption does not seem to be justified on theoretical grounds. For the same reason, translated texts should also be included in bi- and multi-lingual comparable corpora. The incorporation of subcorpora of parallel texts within comparable corpora can also offer practical advantages for contrastive studies. Finally, I provide an overview of the different types of corpora which can be used in translation studies research, and discuss the role of (sub)corpora of translations within these corpora.

1 Introduction

About 15 years ago, Baker noted that "[t]ranslated text has always had a raw deal in corpus linguistics" (1996: 175), being systematically excluded from monolingual corpora because thought of as unrepresentative of the language being studied. She quoted Lauridsen as expressing "the overall position of corpus linguist" (ibid.) when saying that "one should refrain from using translation corpora unless the purpose of the linguistic analysis is either to evaluate the translation process or to criticize the translation product on the basis of a given translation theory" (Lauridsen 1996: 67). Translations seemed to be admitted only as one component of a specific type of corpus in which each translated text is paired with a source text. This type of corpus has been called "parallel" (e.g. McEnery et al. 2008), "translation" (e.g. Granger 1996: 38, Johansson 1998: 4) or "translational" (Lauridsen 1996) corpus. Baker suggested that a different type of corpus may be used to look at the language of translation per se, without assigning to it an a priori value judgment. Like a parallel corpus, this type of corpus also includes two subcorpora, one of translations and one of non-translated texts. The latter, however, does not contain the source texts of the translations, but a set of "comparable" texts in the target language.

Translations have sometimes been mentioned also in relation to a different type of corpus which includes two or more subcorpora. This has been called "multilingual" (Baker 1995) or "comparable" (McEnery et al. 2008) corpus, and consists of two or more monolingual subcorpora in different languages, each collected according to similar design criteria. As with monolingual corpora it is often suggested that translations should not be included in multilingual comparable corpora.

In this article, I first suggest that translated texts have a role to play not only in corpus-based translation studies, where translated text is the object of study, but also in corpus linguistics more in general. I argue that translations should be included in most corpora, be they used in monolingual corpus linguistics or in corpus-based contrastive linguistics. Then, I look at the types of corpora which can be used in translation studies research, and discuss the role of (sub)corpora of translations within these corpora.

2 The role of translations in corpus-based monolingual linguistics

Monolingual corpora are either general or specialized. General language corpora are created with the aim of representing as far as possible a given (national variety of a) language, and thus to function as a reference for that language. Specialized corpora focus instead on a specific variety of language, for instance on a specific text type/genre (e.g. fiction, news, academic prose), domain/topic (e.g. biology, social sciences), production method (e.g. learners' language, translated language), or a combination of various defining features (e.g. translated academic medical language).

Regardless of whether a corpus is general or specialized, it is usually assumed that the linguistic patterns which are observed can be generalized beyond that corpus to the textual universe that corpus stands for. In other words, a corpus is taken to be a "representative sample" of a larger textual population. According to Biber (1993), corpus design should be based on language-external criteria and proceed from the definition of a sampling frame, that is a list of all possible units from which the actual ones used to populate the corpus are selected. The sampling frame should account for the whole range of genres into which the textual population to be represented may be subdivided, and a corpus should be hierarchically organized into subcorpora. The size of the subcorpora should be proportional to the relative frequency of the genres in the textual universe to be sampled, that is to say a corpus should be "balanced" (McEnery et al. 2008: 18). The overall size of a corpus and the delicacy of textual categorization will differ depending on the scope of a corpus, so that a very specialized corpus may be smaller and less stratified than a general corpus.

However, "the difficulties of determining the size of the textual universe and its sub-universes from which a corpus is to be sampled are formidabile" (Leech 2007: 139) and representativeness remains the "holy grail" of corpus linguistics, something to strive for rather than something that can reasonably be attained.[1] Ultimately, it is the sampling frame on which those who design a corpus decide to agree which defines the total linguistic population the corpus is assumed to represent and for which findings have validity, and since "the design of a corpus is a human activity … [it] will always carry the unintended influence of the designer(s)" (Ahmad 2008: 61), including their "socioeconomic origins … and their past and current working environments" (ibid.).

For instance, in the 100 million word *British National Corpus* (*BNC*), which is often taken as a benchmark for general reference corpora, "approximately 10% of the texts come from spoken, 16% from fiction, 15% from (popular) magazines, 10% from newspapers, and 15% from academic, with the balance coming from other genres" (Davies 2009: 161). The sampling frame of the BNC was devised on the basis of institutionalized text types and demographic features with the aim of creating a representative/balanced corpus of British English. However, according to Leech (2007: 136) no serious attempts were made to ensure the proportionality of the genres included, so that for instance spoken language is severely underrepresented. Even considering only written language, the BNC seems to be skewed in favour of the language used by 'educated speakers', since its written component appears to be dominated by texts published by a small group of metropolitan publishers from Southern England, while the tabloid format is under-represented in the newspaper category if this category is to be based on publication/circulation figures (Ahmad 2008: 88-90).

[1] It is perhaps because of this that the adjective "balanced" tends to be preferred to "representative" in more recent discussions of corpus design.

The sampling frame and therefore the composition of other national, general language reference corpora may differ to a larger or smaller extent: for instance, in the 400+ million word *Corpus of Contemporary American English* (*COCA*), which "was designed to be roughly comparable to the BNC in terms of text types" (Davies 2009: 161), "texts are evenly divided between spoken (20%), fiction (20%), popular magazines (20%), newspapers (20%) and academic journals (20%)" (ibid.). Still different are the sampling frames devised for general reference corpora such as the Chinese National Corpus (http://www.cncorpus.org/), the National Corpus of Polish (http://www.nkjp.pl), the Czech National Corpus (http://ucnk.ff.cuni.cz/), the Russian National Corpus (http://www.ruscorpora.ru/), the Korean National Corpus (http://www.tokuteicorpus.jp), and the Corpus de Referencia del Español Actual (CREA, http://corpus.rae.es/), to mention a few.

The composition of a corpus will affect the findings derived from the analysis, and while corpus linguistics is usually presented as a descriptive endeavor, it is also prescriptive in as far as description is a precondition for prescription. Normative reference works like grammars and dictionaries are increasingly corpus-based, which means that the standard of reference is not anymore left to the judgment of the individual scholars who compile grammars and dictionaries, but rather it emerges from the analysis of the most frequent patterns of occurrence in the collective body of texts. The standard, correct norm is therefore instituted by those who design the corpus in as much as they decide which texts are to be considered representative of a language and which texts are not.

The translation status of texts in monolingual corpora is rarely explicitly mentioned in the documentation accompanying them[2], and it is thus difficult to ascertain whether translations were purposefully included or left out. The texts in the BNC were acquired by scanning or typing in printed and spoken material, and underwent a "Britishness" test before being included in the corpus. Dunlop (1993) provides a checklist according to which all texts had to be published in Britain by someone who resided in Britain in the two years preceding text production. Still, while it seems that translations are not explicitly banned, according to one of the authors of the *BNC Handbook* (Aston 2011, personal communication) the BNC does not contain any translations.

Translation status was also not used as an explicit design criterion when compiling the COCA. However, while the texts in the BNC were individually examined before being included in the corpus, the COCA was compiled by automatically downloading texts from "authoritative" data sources, such as the online archives of US-based publishers of books, magazines, newspapers, academic journals and the like. Thus, the COCA should include at least some translated texts, since some of these sources (e.g. book publishers) publish translated works.[3] A survey of the translation status of texts in monolingual corpora in other languages has yet to be conducted, but from a summary perusal of the online documentation accompanying the corpora mentioned above it seems that translations are generally not included, perhaps as a consequence of an implicit judgement concerning their (lack of)

[2] The only exception in the corpora mentioned above is the Russian National Corpus, which includes "albeit in smaller volumes, translated works (parallel with the original texts)" (http://www.ruscorpora.ru/en/corpora-structure.html).

[3] Translated texts are not identified as such in the corpus, so it is impossible to estimate the proportion of translations in the corpus as a whole. However, some traces of translation can be found by browsing the full list of texts available at http://corpus.byu.edu/coca/ (where at least one translation, an extract of a novel by Paulo Cohelo, is explicitly acknowledged,) and by performing a corpus search for "translated from * by" (which retrieves about 40 translated texts).

representativeness. Indeed, it may be that translations are more likely to be included in a monolingual corpus if, as in the COCA, the texts are not individually screened.[4]

In contrast with what seems to be a widespread if implicit practice in monolingual corpus design, I would like to suggest that monolingual corpora should contain translated texts, or at least that "translation" should not be regarded as a criterion for exclusion. Leech (2007: 138), among others, argues that the representation of texts in a corpus should be proportional to both text production and reception. Translation is a legitimate language production activity, and translated texts are to a smaller or larger extent part of what is read by speakers of a language. In other words, translations contribute to the creation of the standard norm for a language and should therefore be part of the sampling frame for a corpus aiming to represent that language.

The practice of excluding translations from monolingual corpora seems to derive from the assumption that translations do not represent, but rather distort a language. Still, this is a prescriptive rather than a descriptive assumption, in as far as it seems to be based on a preconceived idea that what is produced under the constraint of a source text in a different language is by default deviant and not acceptable, and cannot be part of the collective body of reference for the target language.

This is not to contend that the language of translation is not different from that of non-translation. That there is a difference is in fact the hypothesis prompting many studies which compare translated and non-translated language, and which attempt to describe this difference (see below). Translated language may indeed be a specific variety of language, but all language production is subject to some kind of constraints. Even allowing for translated language being significantly 'corrupted' by instances of translationese in the pejorative sense of the term, other varieties of "original" (i.e. non-translated) language are also obviously non-standard, including experimental and highly technical writing.

The proportion of translated texts which should be included in monolingual corpora will vary depending on the language and genres considered. In most English-speaking countries, for instance, translated texts are only a minor percentage of all texts published and read, and therefore they may not exert a strong influence on the reference norm for English. Thus, in a representative English monolingual corpus translations may in fact be present in a very small percentage. Quite the opposite should happen for languages like Italian or Brazilian Portuguese, since in these countries translations represent a considerable share of all published texts. The proportion of translated texts in a specialized corpus may be, for some genres in some languages, even higher than that of non-translations. In a country like Brazil, for instance, where 90% of all published fiction is in translation, it may in fact be very difficult to design a representative/balanced corpus consisting exclusively of non-translations (Laviosa 2002: 40).

Of course, any decision as to which texts qualify for inclusion in a corpus, for example whether or not texts spontaneously produced by non-native speakers of American English should be included in a representative/balanced corpus of American English, or whether or not translated fiction should be included in a representative/balanced corpus of Italian fiction

[4] I do not examine in this paper design issues related to very large Web-based 'reference' corpora (see e.g. Kilgarriff and Grefenstette 2003, Baroni et al. 2009, Pomikalek et al. 2009), which are created on the basis of "language-internal" rather than "language-external" criteria (cf. Sinclair 2005). However, it seems likely that these corpora include translated texts since they are created automatically.

will have an impact on the results. In the final analysis, the design of a monolingual corpus reflects its compilers' idea of what is 'normal' and 'standard' usage in a language or language variety, including possibly a prejudicial view of translated texts.

3 The role of translations in corpus-based contrastive linguistics

Two or more monolingual general corpora can, in principle, be combined into a general reference comparable corpus. For example, since the sampling frames used for the Korean National Corpus, the Chinese National Corpus and the Polish National Corpus are similar to that of the BNC, these corpora are said to "form a balanced comparable corpus that makes contrastive studies for these four languages possible" (McEnery et al. 2008: 49). It should be noted, however, that it is one corpus, in this case the BNC, which acts as a model for the others, and not otherwise. Thus it may well be that, for instance, the composition of a corpus which tries to represent as far as possible what is produced and read in Chinese will differ considerably from that of a Chinese corpus whose composition closely mirrors that of the BNC, since the distribution of domains, text types and socials stratifications (the parameters used for the design of the BNC) differ in the two cultures. As argued by Leech (2007: 142), representativeness and comparability are often conflicting goals: "an attempt to achieve greater comparability may actually impede representativity and vice versa", and "as one nears to perfection in comparability, one meets with distortion in terms of representativeness, and vice versa".

Monolingual specialized corpora may of course also be combined into a bi- or multi-lingual comparable corpus. Each monolingual corpus should be designed in order to represent as far as possible, for each language, the specific genre or text type which is being compared across languages, for instance literary fiction, news writing, or any variety, text type/genre, domain etc. which happen to exist in the languages considered.

Some scholars (e.g. Granger 2003) recommend that comparable corpora should only contain 'original' as opposed to 'translated' texts. Again the implication seems to be that comparable corpora should not be 'tainted' by translation, thus effectively establishing that translations do not conform to a preconceived idea of a standard norm which classifies this method of language production as deviant, and the language produced (translated texts) as unworthy to be included in a corpus aimed at representing the norm for that language. As argued above, however, translations contribute to the creation of the norm of what is produced and read in the language of a receiving culture, for some domains and languages more than for others. Thus, since translated texts may be part of monolingual corpora aiming to represent a certain domain or language, they should also be included in multilingual comparable corpora.

As already stated, the proportion of translated texts in each monolingual subcorpus may differ considerably. For instance, a specialized comparable corpus of English and Italian fiction should contain a much higher proportion of translated texts in the Italian subcorpus, considering that translated fiction accounts for about one third or all titles published in Italy and only about 2-3% of titles published in the UK and the US.[5] It may even be the case that a specialized comparable corpus will have to include almost exclusively original texts in one language and almost exclusively translated texts in the other. As Kenny (1998: 53) explains, in fact, "it is in the very nature of translation that new genres are introduced from one literature to another, and there may be nothing 'comparable' in the host literature to a text introduced to it through translation from another textual tradition".

[5] The difference is greater when considering reception rather than production. See Zanettin (2002) for details.

A bilingual comparable corpus which includes translations may also offer a practical, operative advantage. Such a comparable corpus will in fact include a cross-section of translations for which there could be the corresponding source texts in the other language, possibly in both directions of translation (albeit, of course, not in the same quantity). These parallel subcorpora provide links between the two languages which may serve as starting point in order to compare and contrast features of the languages involved. Indeed, as argued by some scholars (Sinclair 1996, Tognini-Bonelli 1996, Teubert 1996) since links between languages are created by translation studies of contrastive units in different languages should start from the study of actual translational correspondences, and empirical evidence based on corresponding 'translational units' in a parallel corpus should inform the description of lexical equivalents in bilingual dictionaries.

4 The role of translations in corpus-based translation studies

Various attempts have been made towards a typology of the different types of corpora used in translation studies (e.g. Laviosa 1997, 2002, Johannson 2003, Zanettin 2000, 2011, Fernandes 2006, Biel 2009), a task which because of overlapping concepts and terminology has resulted in somewhat different categorizations. Generally speaking, corpus-based translation research usually involves the comparison of two subcorpora, one of which consists of translated texts.[6] This "translational" subcorpus can be compared with different types of subcorpora, depending on the purpose of the analysis. Most research in translation studies has involved either comparable monolingual corpora in which the subcorpus of translations is compared to a subcorpus of non-translations, or parallel corpora, in which the translations are paired with their source texts in (a) different language(s). In these types of corpora the composition of the subcorpus of non-translations is determined by that of the corpus of translations.

A general, representative/balanced translational (sub)corpus would have to be designed according to principles similar to those used to create general monolingual corpora. Like with general monolingual corpora, the adoption of a sampling frame for translated texts should be based on a prior definition of the overall target population to be sampled and of its internal categorization. Halverson maintains that translation theory does not offer "an adequate means of determining where the boundaries of a target population might be drawn" (Halverson 1998: 10), and suggests that a prototype conception of "translation" could be used to define the target population. A translational corpus would thus focus on professional translations, and possibly include peripheral subcorpora of other translated texts which may be looked at as less prototypical. She also argues that a translational corpus should be structured into internal categories which derive from functional and situational parameters "that are valid for the categorization of translations, not of texts in general" (ibid.: 11), and provides some suggestions as to the types of parameters that might be employed.

As far as I know, however, no attempts have been made so far to create general purpose translational corpora. All existing translational corpora are not only specialized in the sense that they contain a specific variety of texts (i.e. translations) but also in that do not aim to represent translated texts in general but only specific translated genres or text types. For instance, the Translational English Corpus (TEC, cf. Laviosa 2000), one of the first and best-known corpora of this kind, consists of one main subcorpus of translated fiction and three smaller subcorpora of other genres (inflight magazines, newspapers and biography).

[6] It is also possible, however, to compare two or more specialized translational subcorpora in the same language, just like other types of monolingual specialized corpora can be compared between them.

In addition to general criteria valid for all text corpora, the design of a translational specialized (sub)corpus will include a specification as to the source language(s) of the translation. For instance, since the TEC was created with the purpose of investigating universal features which are hypothesized to be distinctive of translated texts as opposed to non-translated texts (e.g. explicitation, simplification, disambiguation, conventionalisation, standardization, cf. Laviosa 2002, Olohan 2003, Zanettin 2011) in order to minimize the possible influence of a specific source language, it includes translations from a variety of source languages. Other studies may however involve only one source language (e.g. Baroni and Bernardini 2006).

The design of a monolingual comparable corpus of translations and non-translations poses the same problems encountered when combining other types of monolingual corpora into a bi- or multi-lingual (general or specialized) comparable corpus. Thus, the sampling frame for a representative/balanced translational corpus of fiction may be different from that of a representative/balanced corpus of non-translated fiction in the same language, since possibly not all subcategories of non-translated fiction are equally represented in translated fiction in the same language. Similarly, not all subcategories of translated fiction may have a proportional counterpart in non-translated fiction. For instance, popular fiction is hardly present in translations from Italian into English, while it represents a very large proportion of fiction translated from English into Italian (Zanettin 2002). However, (hypothesized) universal features of translation, be they characterized as absolute laws or probabilistic tendencies (Toury 2004, House 2008, Malmkjær 2008), belong to the domain of cognitive behavior. Thus, the comparable (sub)corpus of non-translations should be designed to be as similar as possible to that of translations, in order to isolate the production method as the only variable. In this case, therefore, issues of comparability will prevail over issues of balance.

If, on the other hand, the aim of a comparable monolingual corpus which includes a translational subcorpus is that of investigating the translation norms which characterize texts translated under specific social and historical circumstances, the design of the (sub)corpus which is compared to the translational one should privilege representativeness/balance rather than similarity of sampling frame, i.e. comparability.[7] As opposed to universals, translation norms belong to the domain of socially constrained behaviour, and the aim of comparison is to establish in what respect regularities of behavior of translated texts, for a certain genre, differ from the standard norm for that genre. As argued above, that standard norm may be determined to some extent not only by original but also by translated texts. In other words, translations norms will result from a comparison between a subcorpus representative of translations, for a certain genre, and a subcorpus representative of that genre as a whole. Such a specialized monolingual subcorpus will be "comparable" in that it is used as a reference, that is to compare translated language to the norm for that language (for that genre).

The design of parallel corpora may vary depending on the number of languages and translation directions involved. A bilingual, monodirectional parallel corpus will be made of a translational subcorpus in one language paired with a subcorpus containing the respective source texts in another language. A multilingual, monodirectional parallel corpus will be one in which the texts in the translational subcorpus are translated from more than one language. Such would be, for instance, the composition of an English parallel corpus having the TEC as

[7] Comparability is used here to refer to the sampling frame used. Other criteria for comparability, such as overall size of the corpus, text extent and time frame (see Zanettin 2011) should apply to all types of comparison between (sub)corpora.

the translational subcorpus. A bilingual, bidirectional parallel corpus contains four rather than two subcorpora, that is both source texts and translations in both languages. One of the first and most well-known corpus of this type is the English Norwegian Parallel Corpus (ENPC, Johansson 1998). A multilingual, multidirectional parallel corpus can contain any number of subcorpora. For instance, a subcorpus of source texts in language A and their translations both in language B and in language C. In this case, the translational relationship holds between A and B and A and C, but not between B and C, which are indeed parallel, but do not contain translations of each other. The subcorpora may become 6 if we have source texts in 2 languages (and their translations into the other languages), or 9 if have source texts in 3 languages. If we add another language we add another factor, resulting in 16 subcorpora, and so on. Multilingual multidirectional parallel corpora can in fact give raise to quite intricate configurations the more languages and directions of translations are involved (Johansson 2003).[8] Parallel corpora in which all directions of translation are covered for all the languages involved have been called "reciprocal" corpora (Teubert 1996). Multilingual parallel corpora containing many languages are of course hard do find, and are mostly confined to multilingual text productions such as international legislatures (EU, UN) and software localization.

A reciprocal corpus can, in principle, combine the benefits of both parallel and comparable corpora, since it makes possible various types of comparison. In a bilingual reciprocal corpus, for instance, each of the two translational subcorpora can be combined with the non-translational subcorpus in the same language, resulting in two monolingual comparable corpora. The two non-translational subcorpora could then be combined in a bilingual comparable corpus.[9] However, both the "comparable" monolingual corpora and the bilingual "comparable" corpus are not really comparable, since the non-translational subcorpora are designed neither to be comparable (i.e. designed according to the same sampling frame) nor to be representative. The design of the corpus as a whole is dictated by the design of the translational subcorpora, and the only design criterion for the subcorpora of non-translation is their status as source texts (Zanettin 2002).

5 Conclusions

I have suggested that translations should be included in most corpora, be they monolingual corpora used for general reference or specialized ones, as well as in bi- and multi-lingual comparable corpora, both general and specialized. This is because translations represent a sometimes substantial proportion of all linguistic production in a given culture. Even taking for granted that translated texts constitute a specific textual variety of a language, this variety should be included in all corpora aiming to represent what is actually read and produced in that language. The proportion of translated texts to be included in such corpora is language- or rather, culture-specific. A decision to exclude translations on the assumption that the language of translation "corrupts" the standard norm of reference does not seem to be justified by theoretical considerations.

On the other hand, corpora designed to investigate regularities of translation usually include a translational subcorpus. This subcorpus can be compared with different types of subcorpora in the same language, whose design will depend on the purpose of the investigation. The comparison may be with a subcorpus of non-translations compiled according to a sampling

[8] A parallel corpus may also contain more than one translation of the same source text, in each target language considered.

[9] The two translational subcorpora in the two different languages could also be compared between them, though the rationale for doing so is less clear.

frame similar to that of the translational (sub)corpus, or with a specialized reference corpus designed according to criteria of representativeness and balance. While in the first instance the analysis focuses on the features specific to translated texts as a result of the translation process, in the second instance the analysis focuses on the deviation of translated texts from the reference norm for the area of specialization considered. I have also argued that bidirectional parallel, or "reciprocal" corpora, while seemingly providing comparable data both within and across languages, are not in fact suitable for these types of comparison.

When a corpus is created, a compromise has often to be reached between ideal design criteria and practical constraints. However, while opportunistic choices may be justified, the limitations and distortions they introduce in the makeup of a corpus should not be forgotten when evaluating the results.

References

Ahmad, Khurshid (2008) Being in Text and Text in Being: Notes on Representative Text. In: Anderman, Gunilla /Rogers, Margaret (eds) *Incorporating Corpora: The Linguist and the Translator*. Clevedon, Buffalo and Toronto: Multilingual Matters. 60-94.

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

Baker, Mona (1996) Corpus-based translation studies: the challenges that lie ahead. In: Somers, Harold (ed.) *Terminology, LSP & Translation*. Philadelphia/Amsterdam: John Benjamins. 175-186.

Baroni, Marco /Bernardini, Silvia (2006) A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary & Linguistic Computing* 21 (3). 259-274.

Baroni, Marco /Bernardini, Silvia /Ferraresi, Adriano /Zanchetta, Eros (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Journal of Language Resources and Evaluation* 43 (3). 209-226.

Biber, Douglas (1993) Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4). 243-257.

Biel, Łucja (2009) Corpus-Based Studies of Legal Language for Translation Purposes: Methodological and Practical Potential. In: Heine, Carmen /Engberg, Jan (eds) *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009. Aarhus 20*. URL http://www.asb.dk/fileadmin/www.asb.dk/isek/biel.pdf < 10.08.2011>

Davies, Mark (2009) The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14 (2). 159-190.

Dunlop, Dominic (1993) TGAW22. Britishness Test for Written Corpus Texts. URL http://www.natcorp.ox.ac.uk/archive/vault/tgaw22.pdf <10.08.2011 >

Fernandes, Lincoln (2006) Corpora in Translation Studies: revisiting Baker's typology. *Fragmentos* 30. 87-95.

Granger, Sylviane (2003) The Corpus approach: a common way forward for contrastive linguistics and translation studies? In: Granger, Sylviane /Lerot, Jacques /Petch-Tyson, Stephanie (eds) *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam/New York: Rodopi. 17-29.

Halverson, Sandra (1998) Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study. *Meta* 43(4). 494-514.

House, Juliane (2008) Beyond Intervention: Universals in Translation? *trans-kom*. *Journal of Translation and Technical Communication* 1(1). 6-19.

Johansson, Stig (1998) On the Role of Corpora in Cross-linguistic Research. In: Johansson, Stig /Oksefjell, Signe (eds) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam/Atlanta: Rodopi. 3-24.

Johansson, Stig (2003) Reflections on corpora and their uses in cross-linguistic research. In: Zanettin, Federico /Bernardini, Silvia/Stewart, Dominic (eds) *Corpora in Translator Education***.** Manchester: St. Jerome. 135-148.

Kenny, Dorothy (1998) Corpora in Translation Studies. In: Baker, Mona (ed.) *Routledge encyclopedia of translation studies*. London/New York: Routledge. 50-53.

Kilgarriff, Adam /Grefenstette, Gregory (2003) Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*.29 (3). 333-347.

Lauridsen, Karen (1996) Text Corpora and Contrastive Linguistics: Which Type of Corpus for which Type of Analysis? In: Aijmer, Karin /Altenberg, Bengt /Johannson, Mark (eds) *Languages in Contrast. Papers from a Symposium on Text-based Cross Linguistic Studies*. Lund: Lund University Press. 63-72.

Laviosa, Sara (1997) How Comparable Can 'Comparable Corpora' Be? *Target* 9 (2). 289-319.

Laviosa, Sara (2000) TEC: A Resource for Studying what is "in" and "of" Translational English. *Across Languages & Cultures* 1 (2). 159-178.

Laviosa, Sara (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*, Amsterdam/Atlanta: Rodopi.

Leech, Geoffrey (2007) New Resources, or just Better Old ones? The Holy Grail of Representativness. In: Hundt, Marianne/Nesselhauf, Nadja /Biewer, Carolin (eds) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi. 133-149.

Malmkjaer, Kirsten (2008) Norms and nature in translation studies. In: Anderman, Gunilla /Rogers; Margaret (eds) *Incorporating corpora: the linguist and the translator*, Clavedon: Multilingual Matters. 49-59.

McEnery, Tony /Xiao, Richard /Tono, Yukio (2008) *Corpus-based Language Studies. An Advanced Resource Book*. London/New York: Routledge.

Olohan, Maeve (2003) *Introducing Corpora in Translation Studies.* London/New York: Routledge.

Pomikalek, Jan /Rychly, Pavel Kilgarriff, Adam (2009) Scaling to Billion-plus Word Corpora. *Research in Computing Science* 41. 3-13

Sinclair, John (1996) An International Project in Multilingual Lexicography. *International Journal of Lexicography.* 9 (3). 179-196.

Sinclair, John (2005) Corpus and Text − Basic Principles. In: Wynne, Martin (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. 1-16.

Teubert, Wolfgang (1996) Comparable or Parallel Corpora? *International Journal of Lexicography*. 9 (3). 238-264.

Tognini-Bonelli, Elena (1996) Towards Translation Equivalence from a Corpus Linguistics Perspective. *International Journal of Lexicography.* 9 (3). 197-217.

Toury, Gideon (2004) Probabilistic explanations in translation studies: Welcome as they are, would they qualify as universals? In: Mauranen, Anna /Kujamäki, Pekka (eds) *Translation Universals. Do They Exist?*. Amsterdam/Philadelphia: John Benjamins. 15-32.

Zanettin, Federico (2000) Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In: Olohan, Maeve (ed.) *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*. Manchester: St Jerome. 105-118.

Zanettin, Federico (2002) CEXI: Designing an English translational corpus. In: Kettemann, Bernard /Marko, Georg (eds) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi. 329–344.

Zanettin, Federico (2011, in print) Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies. Manchester: St Jerome.