Meta
Journal des traducteurs
Translators' Journal

META

Love thy Neighbour: Will Parallel Corpora Endear Linguists toTranslators?

Kirsten Malmkjaer

Volume 43, Number 4, décembre 1998

L'approche basée sur le corpus
The Corpus-based Approach

URI: https://id.erudit.org/iderudit/003545ar
DOI: https://doi.org/10.7202/003545ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Malmkjaer, K. (1998). Love thy Neighbour: Will Parallel Corpora Endear Linguists toTranslators? *Meta*, *43*(4), 534–541. https://doi.org/10.7202/003545ar

Article abstract

This paper analyses the advantages and difficulties that the study of parallel corpora presents when attempting to answer questions arising specifically from within translation studies.

LOVE THY NEIGHBOUR: WILL PARALLEL CORPORA[1] ENDEAR LINGUISTS TO TRANSLATORS?

KIRSTEN MALMKJÆR
The University of Cambridge, Cambridge, United Kingdom

Résumé

Cet article analyse les avantages et les inconvénients de l'approche basée sur les corpus parallèles, particulièrement lorsqu'il s'agit de répondre à des questions soulevées par les études traductologiques.

Abstract

This paper analyses the advantages and difficulties that the study of parallel corpora presents when attempting to answer questions arising specifically from within translation studies.

1. TRANSLATION STUDIES AND LINGUISTICS

Snell-Hornby (1988: 14-15) famously dismisses linguistically oriented traditions within translation studies as having led to a dead end, and as being "dated and of mere historical interest." Her main quarrel with the linguistically oriented tradition is its interest in the concept of translation equivalence, a concept she discusses at length (1988: 15-22), and which, as she points out, is integral to one of the arguments presented in Jakobson's famous article "On linguistic aspects of translation." Jakobson (1959: 233-234) writes:

> Equivalence in difference is the cardinal problem of language and the pivotal concern of linguistics [...] No linguistic specimen may be interpreted by the science of language without a translation of its signs into other signs of the same system or into signs of another system. Any comparison of two languages implies an examination of their mutual translatability; widespread practice of interlingual communication, particularly translating activities, must be kept under constant scrutiny by linguistic science.

Here, Jakobson's concern is with the use of translation equivalents in comparative linguistics, i.e. the use of translation in linguistic research. From the point of view of what translation studies might expect to receive in return from comparative linguistics, though, two problems immediately arise. One concerns the method of eliciting translation equivalents and the second concerns the kinds of equivalents that are being elicited.

The most common elicitation method in comparative linguistics at the time was (and probably still is) introspection on the part of bilinguals. This gives rise to the question: whose authority can we trust, given that (Snell-Hornby 1988: 20),

> As anyone with experience in translation knows all too well, the opinions of the most competent translators [and, we might add, of bilinguals] can diverge considerably.

The second problem, the problem pertaining to the type of equivalent being elicited, is that introspection on isolated words and sentences is not sufficiently context sen-

sitive to yield anything more than citation forms and all-things-being-equal equivalences. In real translation, of course, all things are rarely equal.

The question to be addressed in this paper is whether the study of parallel corpora can effect a narrowing of the gap (as perceived by scholars like Mary Snell-Hornby, among others) between comparative linguistics and translation studies. Clearly, raising this question implies a belief that the two disciplines have something to learn from each other, a belief which I happily confess to holding. There is, in my view, no doubt that linguists and translators ought to be the best of friends; their areas of interest, however one wants to look at them, and however they may differ, have language and linguistic activity at the centre. Yet there is, as we have seen, disaffection bordering on hostility among some translation scholars with regard to linguistics; there is also a degree of indifference to translation among some linguists. This unhappy state of affairs has come about at least in part because of a perception among translation scholars that linguists do not understand the nature of translation, and because of a perception among linguists that translation scholars tend to prefer anecdote to theory, subjectivity to empiricism, and widely scattered data-snippets to generalisations. Obviously, the previous statement is itself a generalisation of the grossest proportions: there exist traditions within translation studies which look very favourably upon linguistics (see, for example, Nida 1959, 1964; Catford 1965; Neubert 1984; Wilss 1977; Reiss 1971; Koller 1972, 1979), and, as we have seen, at least one tradition exists within linguistics which looks favourably on translation, namely contrastive linguistics. In the view of some translation scholars, the trouble is that this is a linguistic subdiscipline whose proponents frequently appear to have no idea what translation proper is, and who have little interest in becoming better acquainted with the process upon which they are fond of drawing. That process, as translators see it, is a far cry from the introspective method of eliciting isolated equivalents from bilinguals (who are not usually translators) which comparative linguists tend to employ. One might expect that research based on large, machine-readable corpora of parallel texts would seem nearer the mark, since they focus on pairs of Source and Target Texts which may be expected to have been produced in and for the market place by professional translators. Let us examine this assumption carefully, beginning by listing some of the many benefits which undoubtedly accrue from such research.

2. WHAT PARALLEL CORPUS STUDIES CAN DO

The claims made for parallel corpus studies include at least the following (see Baker 1993a, 1993b, 1996; Burgess and Kohn 1995; Church and Gale 1991; Marinai, Peters and Picchi 1991, 1992).

i) Parallel corpus studies can reveal characteristics of translated texts, such as tendencies towards explicitness and avoidance of repetition.

ii) Comparison between the translation part of the corpus and a corpus of texts of the same genre, written in the target language for the translation corpus, reveals a tendency towards what we might call the Eliza Doolittle phenomenon: the translated texts, more than the texts in the control corpus, tend to contain those TL phrases, structures, and so on, which, from a comparative point of view, seem particularly characteristic of the TL.

iii) When the translation-part of the corpus is used in conjunction with a corpus containing the Source Texts (together constituting a parallel text corpus), the method can promote sense-disambiguation, and can help to identify translation norms and to create machine translation programmes and bilingual dictionaries.

iv) The method can be exploited for language learning/teaching purposes and for translator training.

v) The study of parallel text corpora can promote "studies of the linguistic phenomena involved in the process of transferring information, ideas, concepts from one language to another" (Marinai, Peters and Picchi 1991: 63-64, 1992: 222).

There is little cause to doubt the first four of these claims. The fifth claim raises a host of questions concerning the stability of concepts under variation in language, even if we replace the notoriously problematic transfer metaphor (see Reddy 1979) with the notion of realisation. These questions are fundamental in more traditional comparative linguistics too; however, I shall not discuss them here, although mention will be made of the types of corpus research likely to promote the goals expressed in claim (v) in sections 3 and 4 below.

Now, it would appear that large, machine readable corpora of pairs of real life Source and Target Texts might also be able to provide reassurance about several specific worries to which the introspective method of comparative linguistics gives rise.

The unease at using one or a few informants' introspection about isolated equivalence relationships between isolated terms, constructions or sentences might fade in the face of the great masses of real, commissioned translations of real, socially functional texts.

The worry about variability might be calmed because the mass of data enables variability to be expressed in probabilistic terms.

The worry about context-sensitivity might disappear because the mass of data is embedded in a mass of co-text and has been produced in a mass of context. The corpus can show the statistical likelihood that a particular equivalence relationship will obtain, given specifiable co-textual and sometimes even contextual factors. It is also possible to take account of norm differences across genres.

Looking at real translations should also be a way of adding an important dimension to comparative/contrastive linguistics. Instead of focusing on the relationships between the languages as static systems, it should be possible to consider the relationships between the languages in use, since in pairs of Source and Target Texts the languages are indisputably being used to some purpose.

Will, then, the use of such corpora endear comparative linguists to translation scholars? I for one would greatly welcome a warming of relations. However, I hope, at the same time, that we are not about to experience a period of head-over-heels infatuation, and while I believe the claims for the method listed at the beginning of this section, I would like to spend some time considering the potential difficulties involved in drawing on parallel corpora in research intended to be relevant to translation studies.

3. SOME POTENTIAL DISADVANTAGES OF PARALLEL CORPORA

Consider the following example:

Hans Christian Andersen writes, in the story commonly known in English as *The steadfast tin soldier*,

Der var engang fem og tyve Tinsoldater, de vare
There were once five and twenty tin soldiers, they were

alle Brødre, thi de vare født af en gammel Tinske.
all brothers, for they were born of an old tin spoon.

The provision of an English gloss for this text fragment presents no difficulties, and it would be reasonable to expect that only minimal changes would be needed to adjust the gloss to proper English use-conventions. The sentence may seem a bit peculiar insofar as biological relationships (being brothers, being born) are set up between inanimates (tin soldiers, tin spoons), but there is no reason to suppose that this need interfere with translation. However, as is so often the case, reality, in the form of a selection of published translations of this text, holds a few surprises (I highlight the terms which function as translational equivalents of *brødre* and of *født*):

1.	BROTHERS all of them, because they were	MADE
2. all	BROTHERS, because they had all been	BORN
3. all	BROTHERS, for all had been	MADE
5. all	BROTHERS, because they had all	COME
7. all	BROTHERS because they had been	MADE
8. all	BROTHERS, for they all	CAME
9.	BROTHERS, for they had all	SPRUNG
10. all	BROTHERS because they were	MADE

Here, *født* seems to pair up with a number of terms which introspection might not have provided as translation equivalents: "made," "come" and "sprung." Nor is any bilingual dictionary likely to list any of these as equivalents for *født*.

As this tiny corpus shows, for each individual instance in a particular context, translators' preferences can differ just as much as they might in cases of context-free introspection. In fact, in this case they probably differ more; I suspect that most people would straightforwardly produce "born" if they were asked to produce an English equivalent for *født*. Yet only one of the translators produces this equivalent. The rest choose either "made," "come" or "sprung."

Let us suppose that one of these translations had been selected for inclusion in a corpus. I assume that most instances of *født* in the rest of such a corpus would be translated as "born." If so, then translation (2), if it happened to be the one selected for inclusion in the corpus, would draw very little attention to itself in this place in the text.

In the case of any one of the other translations, however, the analyst might notice the oddity. Say the oddity was noticed and found sufficiently exciting to merit investigation. A search of the immediate context for the pair will show the semantic anomaly which pertains to both *brødre* ("brothers") and *født* ("born") in the context of inanimates. Since only *født*, and not *Brødre* receives an unexpected equivalent (from the all-things-being-equal point of view), a tentative conclusion might well be that in English the term "born" lends itself less happily to figurative uses than the term "brothers." This conclusion might then be tested by examining a monolingual English corpus which, I assume, would confirm it, since it is exactly the conclusion which the corpus based COBUILD dictionary seems to have arrived at.

COBUILD has 6 headings for brother. One heading gives the literal, biological meaning, and 5 give figurative meanings. In the case of *born* there are 12 headings. 10 are literal-biological, and 2 figurative ("When an idea, organization etc. is born, it comes into existence;" "If something is born of a particular activity or emotion, it exists as a result of it").

It is interesting that the majority of the translators in my sample seem to have reacted with great sensitivity to the norm of use for this term in English which the COBUILD dictionary reflects. If one of this majority was included in the corpus, the analyst might suggest that the translated norms of use for the English term "born" overrode any preconceived equivalence relationship between *født* and "born." If, on the

other hand, translation (2) was to be included, the preconceived relationship would obtain, and there would be no cause to investigate the pairing in this place at all.

These scenarios are, of course, imagined. However, there seems to be some justification for my flights of fancy in Marinai, Peters and Picchi's description of their own procedure (1991: 65):

> The Parallel Text Retrieval System operates in two stages. In the first [...] as many links as possible are established between translation equivalents in the two texts [...]; in the second, the query system uses the links in order to construct, in real time, parallel contexts for any form or occurrence of forms contained in the texts.

The links that are initially established rely on preconceived notions of translation equivalences between pairs of terms or expressions; the equivalence-pairs *født* — "made/sprung/come" would be most unlikely to be among these preconceptions. However, it might be discovered in a search focused for example on the *brødre* — "brothers" pair, since the context provided for this pair would probably be large enough to reveal the *født* — "made/come/sprung" pairing too. As Marinai, Peters and Picchi say (1991: 70):

> A particular advantage of using the system is that, as the parallel contexts are constructed for any word present in the Corpus (and not only for those forms for which a translation equivalent link has been created), real-world translations for words or expressions which are not to be found in any dictionary can be accessed and recorded.

It follows that a great deal hinges on the extent of the context displayed, and Marinai, Peters and Picchi's examples revolve around one or two sentences. Of course, it is possible to search as large a context as one wishes, but in practice one or two sentences is the maximum context computers are asked to search, and it is very common to present the results of the search in concordance lines containing between five and ten words either side of the nodes. The reasons for this restriction of the search and presentation fields are practical: a search using wider contexts would slow the process down so considerably that the advantage of quantity would be lost. Chomsky (1957: 15-17) has remarked in this connection that the method "is incapable of modeling certain syntactic constraints such as agreement over long distances" (Church and Gale 1991: 40). It seems to me likely that similar difficulties might arise in cases where the target is not so much a syntactic phenomenon as a semantic one, including the realisation of information, ideas, and concepts.

In the case of the *født* — "made/come/sprung" pair, of course, it looks as if the context will be perfectly adequate. The semantic anomaly in terms of which the overriding of the straightforward equivalence relationship is to be explained will be obvious, because *Tinske* ("tin spoon") is only the fourth word after *født* ("born"). The problem is that the context would *not* be large enough if an analyst wanted to test whether this anomaly not only explains but *justifies* the choice.

And justification might, in fact, be called for in the case of my little example, because a very strong argument might be made to the effect that the minority translation (2) is preferable to the majority's alternatives, given the nature of the text, and given that the original, I believe, contravenes a norm for Danish which is equivalent to the norm for English. We are dealing with a Source Text in which the language that is being used to talk about the tin soldier constantly teases the reader with the possibility that the tin soldier has certain human emotions, cognitive abilities and characteristics — such as having been born, for example. It could be argued, therefore, that there is reason in this case to override a Target Language norm. It is not my intention to engage in that debate; I simply want to make the point that no such debate would be aided by

evidence from a parallel corpus constructed and searched according to currently predominant conventions. But translators often have to justify the choices they make when they are translating literary texts, which are certainly not the only text type in which the selection of specific linguistic items can be highly significant.

We may suspect, then, that the method by which parallel corpora are typically constructed and searched may very easily render them incapable of providing some of the types of information which would be helpful to translators; and probably the kind of information which would be required to show the varied realisations of concepts in different languages. More specifically, parallel corpora seem to have the potential to fail translators in two ways (at least):

> i) Like metaphors (Lakoff and Johnson 1980), parallel corpora highlight some phenomena while hiding others, and the distribution of phenomena between the hidden and highlighted categories may not be the one most helpful to translators.
> ii) Although parallel corpora provide evidence of how languages relate to each other in use, we still only get one individual's introspection on each individual instance contained in the corpus. But, as Snell-Hornby remarks (see section 1 above) and as my small corpus confirms, translators' opinions may differ on individual instances in individual contexts. We may suspect that where they differ most is where investigation might prove particularly fruitful. As parallel corpora are constructed at the moment, these cases would not come to light.

To sum up, then, the problems pertaining to the use of parallel corpora in the furtherance of translation studies include: first, the metaphor-like problem of highlighting and hiding. Secondly, the problem that a parallel corpus still only provides, for each instance, the result of one individual's introspection, albeit contextually and cotextually informed. Thirdly, the problem that in order to be able to provide any kinds of *explanation* of the data provided by the corpus, rather than mere statistics, analysts really need substantially more context than computers tend to search and display. Biber and Finegan (1991: 209) remark on the propensity of corpus linguists to treat their data as a collection of individual sentences, ignoring the potential offered by the corpora for studying features of text. It looks as if this tendency is in danger of being transferred wholesale to parallel text studies.

4. A WAY AHEAD?

Stig Johansson (1991: 305-6) suggests that in spite of the undoubted advantages of large corpora, there is still something to be said "for smaller, carefully constructed sample corpora which can be analysed exhaustively in a variety of ways." I believe that one very fruitful way forward for corpus linguistics would be to begin to supplement the vast, quantitatively oriented parallel corpus studies with smaller corpora consisting of source texts and as many translations of each source text as possible. The problem, of course, would be that there are not many genres which include texts that have had several translations made of them, so that anyone wishing to use this methodology would probably be forced either to rely on literary texts or to commission the translations. I personally think that literary stylistics is not so far removed from non-literary stylistics that the use of literary texts would be a particular disadvantage, though it might be difficult to find the texts in machine-readable form. Another difficulty is that literary texts which have been translated many times tend to be non-contemporary. However, it seems clear that text analysis and interpretation *must* begin to play a part in studies of parallel corpora, if we are to get anywhere near to being able to say anything about how

ideas/concepts/information are/is realised in different languages. But it may also be important for more modest projects. As Leech (1991: 15) remarks, "successful analysis depends on a division of labour between the corpus and the human mind." Some literature on corpus linguistics suggests a worrying tendency to try to divide the labour in such a way that the mind's share is minimal or expended during the corpus construction phase. Sometimes it almost seems as though availability in machine readable form were the sole criterion for the texts' inclusion in the corpus.

For the type of research I have in mind, it would be necessary to select texts carefully, read them in (or even, perish the thought, key them in and perhaps align them manually in the case of corpora of multiple translations), and analyse them very carefully indeed. If such corpora were added to the vast norm generators, corpus linguistics would be able to cater both for the linguist's love of generalities and for the translator's love of the instance. The two types of corpus could also act as useful checks on each other, and there might be some hope that their mutuality might come to be reflected in the relationships between linguists and translation scholars.

Note

1. By "parallel corpus" I mean a corpus containing ST-TT pairs.

Source for Andersen's original text

H.C. Andersens Eventyr: Kritisk udgivet efter de originale eventyrhæfter med varianter ved Erik Dal og kommentar ved Erling Nielsen. I:1835-42. 1963. II:1843-55. 1964. III: 1856-65. 1965. IV:1861-66. 1966. Copenhagen: Hans Reitzels Forlag.

Sources for the translations

1) *Hans Christian Andersen: Fairy Tales*. Translated by Marie-Louise Peulevé. Odense: Skandinavisk Bogforlag. No Date (but pre-1968).
2) *Hans Andersen's Fairy Tales: A Selection*. Translated from the Danish by L.W. Kingsland. London: Oxford University Press. 1959.
3) *Hans Christian Andersen: Fairy Tales*. Translator not named. London etc.: Hamlyn. 1959.
4) *Tales from Hans Andersen*, Translated by Stephen Corrin, 1978, London: Guild Publishing. 1989.
5) *Hans Andersen: His Classic Fairy Tales*. Translation Copyright Erik Haugaard 1976. London: Lynx. 1988.
6) *Hans Christian Andersen: Eighty Fairy Tales*. Translated by R.P. Keigwin. Odense: Skandinavisk Bogforlag. 1976. Introduction by Elias Bredsdorff. New York: Pantheon Books. 1982.
7) *Hans Christian Andersen: Fairy Tales*. Translated by Reginald Spink. London: Everyman's Library. 1960.
8) *Hans Christian Andersen: Stories and Fairy Tales*. Selected, translated and illustrated by Erik Blegvad. London: Heinemann. 1993.

REFERENCES

BAKER, Mona (1993a): "Corpus Linguistics and Translation Studies: Implications and Applications", Mona Baker, Gill Francis and Elena Tognini-Bonelli (Eds), *Text and Technology: In Honour of John Sinclair*, Amsterdam/Philadelphia, John Benjamins, pp. 233-250.

BAKER, Mona (1993b): *Multilingual Databases*, Report on a feasibility study on multilingual lexicography funded during 1990-91 by the Council of Europe under contract no. 57/89.

BAKER, Mona (1996): "Corpora in Translation Studies: An Overview and some Suggestions for Future Research", *Target*, 7 (2), pp. 223-243.

BIBER, Douglas and Edward FINEGAN (1991): "On the Exploitation of Computerized Corpora in Variation Studies", Karin Aijmer and Bengt Altenberg (Eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London and New York, Longman, pp. 204-220.

BURGESS, Gordon J.A. and Janos KOHN (1995): "The Use of Parallel Concordancing for Literary and Linguistic Text Analysis", Paper presented at EUROCALL 1995, Valencia.

CATFORD, J. C. (1965): *A Linguistic Theory of Translation*, London, Oxford University Press.

CHOMSKY, Noam (1957): *Syntactic Structures*, The Hague, Mouton & Co.

CHURCH, Kenneth W. and William A. GALE (1991): "Concordances for Parallel Text", Paper presented at the Seventh Annual Conference of the UW Centre for the New OED and Text Research, *Using Corpora*, September 29-October 1, 1991, St. Catherine's College, Oxford, England.

JAKOBSON, Roman (1959): "On Linguistic Aspects of Translation", Reuben A. Brower (Ed.), *On Translation*, Cambridge, Mass., Harvard University Press, pp. 232-239.

JOHANSSON, Stig (1991): "Times Change, and so do Corpora", Karin Aijmer and Bengt Altenberg (Eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London and New York, Longman, pp. 305-314.

KOLLER, Werner. (1972): *Grundprobleme der Übersetzungstheorie. Unter besonderer Berücksichtigung schwedisch-deutscher Übersetzungsfälle*, Bern, Francke.

KOLLER, Werner (1979): *Einführung in die Übersetzungswissenschaft*, Heidelberg, Quelle & Meyer.

LAKOFF, George and Mark JOHNSON (1980): *Metaphors we Live by*, Chicago and London, The University of Chicago Press.

LEECH, Geoffrey (1991): "The state of the art in corpus linguistics", Karin Aijmer and Bengt Altenberg (Eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London and New York, Longman, pp. 8-29.

MACKENZIE, Rosemary (1994): "A Quality-based Approach to Teaching Specialised Translation", Paper presented to the 1st International Congress on Translation and Interpreting: Present Trends, Universidad de Las Palmas de Gran Canaria, 24-26 Feb. 1994.

MALMKJÆR, Kirsten (1994): "Translating Customer Expectations into Teaching", Catriona Picken (Ed.), *ITI Conference 7 Proceedings: Quality – Assurance, Management and Control,* London, ITI, pp. 143-155.

MARINAI, Elisabetta, PETERS, Carol and Eugenio PICCHI (1991): "Bilingual Reference Corpora: A System for Parallel Text Retrieval", *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, Oxford, St. Catherine's College, pp. 63-70.

MARINAI, Elisabetta, PETERS, Carol and Eugenio PICCHI (1992): "Bilingual Reference Corpora: Creation, Querying, Applications", Ferene Kiefer, Gabor Kiss and Julia Pajzs (Eds), *Papers in Computational Lexicography: Complex 92*, Budapest, Linguistics Institute, Hungarian Academy of Sciences, pp. 221-228.

NEUBERT, Albrecht (1984): "Text-bound Translation Teaching", Wolfram Wills and Gisela Thome (Eds), *Translation Theory and its Implementation in the Teaching of Translating and Interpreting*, Tübingen, Narr, pp. 61-70.

NIDA, Eugene A. (1959): "Principles of Translation as Exemplified by Bible Translating", Reuben A. Brower (Ed.), *On Translation*, Cambridge, Mass., Harvard University Press, pp. 11-31.

NIDA, Eugene A. (1964): *Toward a Science of Translating. With Special Reference to Principles and Procedures Involved in Bible Translating*, Leiden, Brill.

REDDY, Michael J. (1979): "The Conduit Metaphor — A Case of Frame Conflict in our Language about Language", Andrew Ortony (Ed.), *Metaphor and Thought*, Cambridge, Cambridge University Press, pp. 284-324.

REISS, Katharina (1971): *Möglichkeiten und Grenzen der Übersetzungskritik. Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*, München, Hueber.

SNELL-HORNBY, Mary (1988): *Translation Studies: An Integrated Approach*, Amsterdam/Philadelphia, John Benjamins.

VIENNE, Jean (1994): "Towards a Pedagogy of 'Translation in Situation'", *Perspectives: Studies in Translatology*, 1/1994, pp. 51-59.

WILLS, Wolfram (1977): *Übersetzungswissenschaft. Probleme und Methoden*, Stuttgart, Klett.