Meta
Journal des traducteurs
Translators' Journal

META

The Corpus-based Approach: A New Paradigm in Translation Studies

Sara Laviosa

Volume 43, Number 4, décembre 1998

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

URI: https://id.erudit.org/iderudit/003424ar
DOI: https://doi.org/10.7202/003424ar

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

Laviosa, S. (1998). The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta*, *43*(4), 474–479. https://doi.org/10.7202/003424ar

THE CORPUS-BASED APPROACH: A NEW PARADIGM IN TRANSLATION STUDIES

INTRODUCTION

Only a few years ago, Baker (1993: 243) predicted that the availability of large corpora of both original and translated texts, together with the development of a corpus-driven methodology, would enable translation scholars to uncover "the nature of translated text as a mediated communicative event." Since then, a growing number of scholars in translation studies have begun to seriously consider the corpus-based approach as a viable and fruitful perspective within which translation and translating can be studied in a novel and systematic way. Contrastive linguists have also recognised the value of translation corpora as resources for the study of languages, and translator trainers have begun to design general and specialised corpora to aid the comprehension of source language texts and improve production skills.

The aim of this issue's collection of corpus-based studies is twofold. On the one hand, it attempts to outline the existing territory occupied by a new field of research in translation studies; on the other, it hopes to show that the corpus-based approach is evolving, through theoretical elaboration and empirical realisation, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation.

The studies included in this volume have been grouped into two main categories on the basis of their primary research focus. The first group consists of discussions concerning theoretical issues pertaining to the scope, object of study, and methodology of the corpus-based approach. The second is made up of empirical and pedagogical studies of translation and translating. The concluding paper by Maria Tymoczko draws on the insights provided by these studies and discusses the role that computerised corpora will play within the discipline as a whole.

THEORETICAL RESEARCH

The collection begins with Baker's discussion of the need to develop a coherent corpus-based methodology for identifying the distinctive features of the language of translation. The aim of this endeavour, she argues, is not merely to unveil the nature of the 'third code' per se, but most importantly, to understand the specific constraints, pressures, and motivations that influence the act of translating and underlie its unique language.

In the second paper, Shlesinger looks at the problems and benefits that may arise from applying a corpus-based methodology to the study of interpreting, viewed not merely as a particular type of translation, but as a distinct "mode of interlingual processing [...] shaped by its own goals, pressures and context of production." Shlesinger explores two ways in which interpreting could be fruitfully investigated with the aid of corpora. The first is direct and involves the design of new types of parallel and compa-

rable corpora. The second consists of employing existing monolingual corpora to extract material that can be used in experimental research into interpreting.

The obstacles that may slow down the development of corpus-based interpreting studies are, according to Shlesinger, the transcription of raw interpreted speech — still very costly despite recent advances in computational linguistics — and the current inability to represent and easily access the paralinguistic features (such as prosody and timing) that uniquely characterise interpreting *vis-à-vis* other forms of oral discourse. These drawbacks notwithstanding, Shlesinger proposes extending Baker's notion of comparable corpora to comprise three separate types of text in a single language: interpreted speeches from a variety of source languages, original spoken texts produced in similar settings, and written translations of source oral texts delivered in analogous circumstances. This new design would permit not only the study of interpreted texts as distinct pieces of oral discourse, but also the identification of those patterns that distinguish interpreting from written translation. The traditional parallel corpus design is also adapted by Shlesinger to include three sets of text: source language texts; their interpreted versions; their written translations. The particular advantage of this kind of corpus is that it permits the study of language- and direction-specific features of the interpreted output along with their possible interaction with extra-linguistic factors such as gender, extent of professional experience, language background, etc. providing, of course, that they are separately recorded as part of the corpus design.

Moreover, the availability of written translations would permit the identification of modality-specific factors. Finally, monolingual corpora can also be used in experimental interpreting studies, where there is a trade-off between the need to control the numerous variables affecting the process of interpreting and the need for ecological validity. According to Shlesinger, this problem can be partly overcome by creating input material extracted from large volumes of authentic texts. An example of this type of application is provided by Shlesinger's own research (in progress) into the way in which working memory operates in simultaneous interpreting. The methodology adopted in this study consists of two stages: first, the selection, from the British National Corpus, of a set of authentic utterances which vary according to pre-established variables (for example, length and frequency of words, phonological patterns, semantic features, etc.); second, the creation of texts containing these utterances, which are then given to the subjects for interpreting.

Halverson discusses the issue of representativeness in the creation of general translation corpora, with a view to addressing a more general philosophical question, which, in her view, is fundamental to the development of corpus-based translation studies: how can we match theory, data, and methodology in a coherent whole so as to ensure the comparability and integration of results obtained in this new field of research?

Through an in-depth examination of the stages involved in selecting texts which adequately represent the target population, Halverson demonstrates that prototype categories are better suited than the all-or-none classical ones to reconcile existing theoretical statements about what constitutes legitimate data in translation studies with the new methodology developed by the corpus-based approach.

While Halverson tackles a theoretical issue that concerns all empirical research carried out within the corpus-based perspective, Kenny focuses on some of the methodological and theoretical challenges presented by her particular research: the investigation of "sanitisation" in translated texts through the analysis of semantic prosody. Her study relies upon two large reference corpora, the British National Corpus and the Mannheim Korpora, and a parallel corpus consisting of literary texts in German and

their translation into English, which she is compiling under the supervision of Mona Baker at UMIST in co-operation with Dublin City University. Her main hypothesis is that target texts tend to use toned-down vocabulary compared with their sources, and this results in the creation of what Kenny terms a "sanitised version of the original." Empirical evidence of this phenomenon will be gathered by examining the semantic prosodies of as many lexical items as possible in the parallel corpus. These will subsequently be analysed against the backdrop provided by the reference corpora for both the source and the target languages.

Puurtinen outlines the aims and methodology of a study she is currently carrying out in the field of children's literature. Her research makes use of a composite corpus representing original English, original Finnish and translated Finnish from English. The initial focus of her investigation is the analysis of nonfinite constructions, taken as a measure of readability of children's books. Ultimately, her aim is to infer, through interpretation of the lexico-grammatical patterns emerging in her corpus, the ideological norms prevailing in the literary systems of English and Finnish children fiction.

Finally, Malmkjær analyses the advantages and difficulties that the study of parallel corpora presents when attempting to answer questions arising specifically from within translation studies. Among the advantages, she cites the authenticity of the texts and their availability in large quantities, both of which contribute to the identification of translational norms and the assessment of the probability that a given equivalence is representative of the relationship between the utterances produced in source and target texts of the larger parent population. Moreover, authentic texts can be invaluable to the contrastive linguist who wishes to investigate differences and similarities in language use, as opposed to language systems.

Malmkjær also points out two main problems connected with the use of parallel corpora in translation studies. The first is that the concordance lines generally used as an analytical tool do not always offer enough linguistic context to investigate features of whole texts and/or semantic phenomena such as the expression of information, ideas and concepts. There is therefore a risk that some aspects of translational behaviour may be revealed, while others are overlooked because they are inaccessible through KWIC concordances. The second difficulty is connected with the way parallel corpora are, on the whole, designed to include only one translation for each source text. This may hide an important aspect of the translational phenomenon, namely the differences existing between the various translations of the same original work. To remedy these shortcomings, Malmkjær suggests complementing norm-oriented studies, which justifiably require large amounts of text, with smaller and carefully constructed corpora which consist of one source text and as many translations of it as possible, so that in-depth investigations of entire texts can be performed. Her advice is to "select texts carefully, read them in (or even [...] key them in and perhaps align them manually, if necessary, in the case of the corpora of multiple translations)" before analysing them in greater detail. There are two advantages in combining these two different methodologies. First, the findings would be richer — since it would be possible to identify both norm-governed and idiosyncratic behaviour, and second, they would be more rigorous — given that the larger corpus could be checked against the individual cases examined in the smaller corpus. This type of methodology, Malmkjær argues, caters to the needs of both contrastive linguists and translation scholars, bringing them closer to one another in a relationship of mutual co-operation.

The section opens with Munday's analysis of shifts in Edith Grossman's translation, *Seventeen Poisoned Englishmen*, of a Spanish novel by Gabriel García Márquez. Munday uses a variety of basic tools from corpus linguistics — word frequency lists, text statistics, concordances — as aids to the inductive exploration of texts. Word frequency lists are first obtained for both source and target texts and then compared for "spotting useful areas of investigation." He uses "intercalated text," i.e. a text obtained by manually keying in the translated text between the lines of the source text. He subsequently runs concordances of this intercalated text and uses them to carry out a contextualised comparative analysis of all the instances of selected lexical items in order to examine some of the shifts "that build up cumulatively over a whole text [rather than a text fragment] as a result of the choices imposed on or taken by the translator." This type of analysis is performed not to evaluate the quality of a given translation, but to understand the decision making process underlying the product of translation and to infer from it the translational norms adopted by the translator. Munday's approach is therefore descriptive, product- and process-oriented, and data-driven. In fact, he derives his hypotheses from observing differences that appear in the parallel frequency lists and during manual construction of the intercalated text. These initial hypotheses are then investigated with the aid of additional automatic methods of analysis, for example aligned concordance lines.

Munday's preliminary study of the first 800 words of his full-text parallel corpus reveals shifts in cohesion and word order which occur over the whole text and have the effect of moving the narrative viewpoint from the first to the third person and thereby distancing the reader from the thoughts, experiences and feelings of the main character in the story. Munday's innovative work shows how the analytical tools of corpus linguistics can be used heuristically to discover patterns that cannot be discerned through manual analysis, while at the same time assess the cumulative impact that the individual choices of the translator have over the entire text.

A different type of corpus is used in Laviosa's investigation of the linguistic nature of English translated text. Her corpus consists of a sub-section of the English Comparable Corpus (ECC) (Laviosa-Braithwaite 1996). It comprises two collections of narrative prose in English: one is made up of translations from a variety of source languages, the other includes original English texts produced during a similar time span. The study reveals four patterns of lexical use in translated versus original texts: a relatively lower proportion of lexical words versus grammatical words, a relatively higher proportion of high-frequency versus low-frequency words, relatively greater repetition of the most frequent words, and less variety in the words most frequently used. The author proposes to call these regular aspects of English translated text "core" patterns of lexical use in an attempt to convey the fact that, given that they occur in both the Newspaper and the Narrative Prose Subcorpora of ECC, they may prove typical of translational English in general.

Øverås's investigation of explicitation (expressed in terms of a rise in the level of cohesion) in translational English and translational Norwegian is another elegant empirical study aimed at unveiling the specificity of the language of translation regardless of the contrastive differences existing between the two languages in contact. Her ultimate objective, however, like Puurtinen, is to go beyond mere linguistic investigation in as far as she attempts to reach conclusions about the literary translational norms prevaling in the target communities she has studied.

From the perspective of contrastive linguistics, Maia analyses the frequency and nature of the SVO sentence structure in English and Portuguese, particularly in those

cases where the subject is realised by the first person pronoun *I* and *eu*, respectively, or by a name. The corpus analysed is a small bidirectional parallel corpus comprising a Portuguese novel and its English translation, and an English novel and its Portuguese translation. The texts contain a large number of monologues and dialogues, assumed to be representative of near-speech type usage. The author considers the parallel component of the corpus appropriate for comparing how the same situation is represented in the two languages, while the original texts permit additional comparisons between the original languages on the one hand, and between the translational and non-translational variety of the same language on the other. The discrepancies observed in the frequency of personal subjects (realised by either names or pronouns) suggest, contrary to what happens in English, that the apparently subjectless V + O sentence structure is the norm, rather than the exception, in original Portuguese, and that translational Portuguese is influenced by the norms of the English language. Moreover, while the use of *I* is syntactically necessary in English, the occurrence of the Portuguese equivalent *eu* seems to be related to pragmatic factors, such as thematisation, topicalisation and emphasis. On the basis of these findings the author argues that "the flexibility of word order and the wider variation of thematisation in Portuguese in relation to English do at least allow for more subtlety in communication."

Like Maia, Ebeling regards parallel corpora suitable sources of data for investigating the differences and similarities between languages, and adopts the notion of translation equivalence as a methodology for contrastive analysis. Ebeling uses a bidirectional parallel corpus of Norwegian and English texts (the ENPC) to examine the behaviour of presentative English *there*-constructions as well as the Norwegian equivalent *det*-constructions in original and translated English, and original and translated Norwegian respectively. The corpus of original English reveals that *be* is by far the most common verb occurring in these structures, while Norwegian allows a much wider set of verbs, some in the passive voice. Ebeling's analysis of the Norwegian translation equivalents of the English *there be*-constructions reveals an optional choice of specification with *det*-constructions containing verbs other than those of existence, *have*-existentials, *det*-constructions with passives. On the other hand, the English translation of *det*-constructions with active lexical verbs often leads to despecification, as does, to an even greater extent, the translation of *det*-constructions with passive verbs. These results partly confirm the predictions put forward on the basis of evidence from the original corpora and throw new light into the assumed relationship of equivalence between two structures found in English and Norwegian.

By adopting a genuine descriptive approach to translation and, at the same time, remaining fully aware of the specificity of this act of language mediation, Maia and Ebeling are able to break new ground in their respective research fields. They also demonstrate how fruitful and exciting the co-operation between contrastive linguistics and translation studies can be, through the adoption of a common corpus-based methodology.

Finally, Zanettin's and Bowker's respective works are of particular interest to those directly involved in the applied area of translator training. Zanettin demonstrates how small bilingual corpora of either general or specialised language can be used to devise a variety of structured and self-centred classroom activities aimed at enhancing student's understanding of the source language text and their ability to produce fluent target language texts. It is not unreasonable to predict that Zanettin's idea of providing students with a "translator trainee workstation" will fairly soon become a common feature in more progressive and technologically advanced training institutions.

Bowker, still within a pedagogical perspective, reports on the results of an interesting experiment in which she compares two translations produced by a group of translator trainees. One translation was carried out with the use of conventional resources; the other with the aid of a specialised monolingual corpus, which was consulted using the analytical facilities provided by *WordSmith Tools*. The results reveal that the corpus-aided translations were of higher quality in respect of subject field understanding, correct term choice, and idiomatic expression. The author observes that, although she did not find any improvement with regard to grammar or register, the use of the corpus was not associated with poorer performance either. These experimental findings are exciting and encouraging. Other scholars will no doubt be inspired to pursue this new type of experimental work in the applied area of general and technical translator training.

Finally, Tymoczko's concluding article provides an inspiring and convincing discussion of the centrality of corpus-based studies within the entire discipline. Her heartfelt warning against the possible danger of pursuing scientific rigour as an end in itself through empty and unnecessary quantitative investigations echoes Baker's worry that the systematic research into the nature of the third code may not go beyond the study of recurrent linguistic patterns.

I think that the studies selected for this collection show that the present corpus-based translation scholars are well aware of these potential dangers and that they are keen to combat them with serious, well thought-out, and theoretically sound research.

SARA LAVIOSA
UMIST, Manchester, United Kingdom

REFERENCES

BAKER, Mona (1993): "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 and Philadelphia, John Benjamins, pp. 233-250.

LAVIOSA-BRAITHWAITE, Sara (1996): *The English Comparable Corpus (ECC): A Resource and Methodology for the Empirical Study of Translation*, PhD Thesis, Manchester, UK, UMIST.

SHLESINGER, Miriam (in progress): *Working Memory in Simultaneous Interpreting*, PhD Thesis, Ramat Gan, Israel, Bar-Ilan University.