

## Corpus-based Approaches to Translation Studies\*

### APPROCHES FONDÉES SUR LES CORPUS DANS L'ÉTUDE DE TRADUCTION

SHEN Guo-rong<sup>1</sup>

**Abstract:** Recent years have witnessed a significant growth of corpus-based translation studies that appeared in the beginning of the 1990s. Corpus linguistics has provided a new weapon for translation studies, broadened the research scope and introduced a brand-new thought pattern for translation scholars. This paper introduces the design and application of Translational English Corpus. Besides, it makes an objective assessment to corpus-based translation studies and analyses the potential of Translational English Corpus.

**Key words:** Corpus; Corpus linguistics; Translation studies; Advantages; Limitations

**Resumé:** Ces dernières années ont connu une croissance importante des études de traduction à base de corpus qui est apparue au début des années 1990. La linguistique de corpus a fourni une nouvelle arme pour les études de traduction, élargi le champ de recherches et introduit un mode de pensée tout nouveau pour les spécialistes de la traduction. Cet article présente la conception et l'application de corpus translationnel en anglais. En outre, il fait une évaluation objective sur des études de traduction à base de corpus et analyse le potentiel de corpus translationnel en anglais.

**Mots-clés:** corpus; linguistique de corpus; études translationnelles; avantages; limitations

## 1. INTRODUCTION

With the increasing international exchange and accelerated globalization, corpus linguistics has now become main stream. Corpus-based research is increasingly influential in many areas of language studies.

---

\*This paper is completed in two projects researching and also as a result of them, namely, Approaches to Interpreters Training in Interpretative Theory Context (one of the 2010 Social Researches Granted by the Education Office of Henan Province ) and On Interpreters Self-training By Corpora (the Elenventh Five Years Reasearches Granted by the Education Oiffice of Hnena Province,2010).

<sup>1</sup> Director of English Department, School of Freign Languages, Henan University of Technology, China.

\* Received 22 November 2010; accepted 15 December 2010.

As part of it, translation study is not an exception. Recent years have especially witnessed a significant growth of corpus-based translation studies that appeared in the beginning of the 1990s.

## **2. TYPES OF CORPUS USED IN TRANSLATION STUDIES**

It is widely accepted that Mona Baker is the first scholar applying corpus to explain translation phenomenon in the middle of 1990s. The following 1990s saw the adoption of the corpus-based approach to translation studies, with other translation theorists like Gideon Toury, Kirsten Malmkjaer, Miriam Shlesinger as the leading figures. According to Baker, corpora applicable to translation research include mainly parallel corpora, comparable corpora and multilingual corpora.

Parallel corpora, consisting of original texts and their translated versions, have traditionally been the most popular data for research in translation studies. They can tell us a great deal about those patterns of language use specific to certain target texts, and should thus be very informative regarding particular translation practices and procedures used by the translator. Multilingual corpora are sets of two or more monolingual corpora in different languages and, as such, can be explained in finding out more about the typical means employed by two or more languages to express similar meanings. In a sense this limits the theoretical value of multilingual corpora when what is being investigated is the phenomenon of translation itself. Comparable corpora can help to overcome the difficulty involved in researching the nature of translation. This kind of corpora refers to two separate collections of texts in the same language, one consisting of original texts, the other of translations in that language comparable of linguistic features which are typical of translated texts. This is central to research into translation universals. (Kenny 68)

Among the three kinds of corpus, the former two are especially widely in use. They play a positive role in the studies of often highly abstract features of source and target texts, in testing out hypothesis in an explicit, empirical way, in the extraction of bilingual terminology and so on.

## **3. GENERAL PROCEDURES OF USING CORPORA IN TRANSLATION STUDIES**

No matter in what kind of translational field corpus will be used, the general procedures or the basic steps are the same.

First, the researcher needs to set up a goal, that is, to put forward a hypothesis he/ she wants to verify. For example, he/she may want to work out the frequency of a certain article in different types of texts, the more preferred words in a specific writing, or even the usage of punctuations.

Once the aim is clear, the second step is to take action purposely, that is to establish a suitable corpus, which is the most tedious and important task. Of course, the corpus can be big or small. The design criteria of corpus crucially depend on the envisaged use of the corpus and center on the idea that corpora should be somehow representative of a particular type of language production and reception, design criteria typically involve decisions such as whether spoken and written languages are to be included, what text types should be accounted for, and whether text samples or full texts are to be included.

Since everything is ready, we just need to carry out our investigation patiently. Just as Liu Jingguo (73) hold that we need to follow a certain rule and contrast the corpora or the data carefully. Thus it is necessary to be sensitive to minor difference between one text and the other. But this is so easy that everybody can make it. However, we should not stop ourselves to this superficial level but to find out things deeply buried. We can conclude our findings, analyze the results and their causes, and finally apply what we have learned to our own translation activity. For example, corpora can tell us how frequently certain language items or structures are used. This kind of information is useful when we try to select what to focus on, and decide what to focus on in the translation. We all know that translation can be divided into tourist style, technical style, economic style, political style and so on. Through corpus, we can conclude the most frequently used word in a certain type of article, and this will make our translation more professional and native.

## **4. APPLICATION OF THE CORPUS IN TRANSLATION**

The integration between corpora and translation studies has generated a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description and the practice of translation, that is also to say corpus-based translation studies come into two broad areas: theoretical and practical. In theoretical terms, corpora are used mainly to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the linguistic features and their frequencies in translated texts and original texts. While in the practical approach, corpora provide a workbench for translators and a basis for developing applications like machine translation and computer-assisted translation systems. According to a conference held in Shanghai Communication University on corpus and translation, corpus-based research has given rise to a new way of conceptualizing, studying and teaching translation. To be more concrete, its practical use can mainly be seen in the following several fields.

### **4.1 Investigating the Style of a Translator**

The first application of corpus in translation is to investigate the style of a translator under the leadership of Mona Baker. Baker, professor and dean of the centre for translation studies, held that “in the translation activity, the version must be marked with the translator’s subjectivity, just as when holding an object, fingerprint would be left on it” (227). Therefore, she put forward the hypothesis that translator had their own styles. In order to verify her hypothesis, Baker persuaded the University of Manchester Institute of Science and Technology (UMIST, Manchester) to establish a large Translational English Corpus. After the foundation of the TEC, Baker began her investigation.

#### **4.1.1 Content of the Investigation**

As her goal is to investigate the style of a translator, the first task is to choose the typical translator. Baker chose British translators Peter Bush and Peter Clark as her research targets and contrasted their works including three novels and two biographies translated by Bush and three novels translated by Clark. The originals of Bush’s three novels come from three authors using three different kinds of languages while those of Clark come from two Arabian writers utterly different in their styles. She described the style of the two translators from these aspects, such as type/token ratio, contact/full form, lexical density, sentence length, key word in context etc. I will choose type/token ratio as an example to illustrate how the corpus-based approach is used to investigate a translator’s style.

You may get confused about what type/token ratio is. As the name indicates, it is a rate. Once texts have been selected for inclusion in a corpus, a decision has to be made as to how they should be represented in electronic form. Basic mark-up may involve indicating the main divisions in a text, for instance, the addition of descriptive headers to individual texts. The level of mark-up that a corpus is subjected to will have implications for the kind of electronic processing the corpus can undergo. Zhang Meifang (60) once admitted that raw corpora, i.e., untagged, unparsed corpora, can be treated as sequences of characters in running text that are delimited by spaces. Such running words or tokens can be counted and the number of times a single form or type appears can be calculated. The frequency of occurrence of single types can thus be worked out for the whole corpus. The number of different types compared to the total number of tokens in the corpus yields the type/token ratio for that corpus. Type/token ratios tell us something about the variety of the vocabulary used in a corpus.

As the definition of type/token ratio is clear, we can go on our research about different styles of Bush and Clark through comparison between their type/token ratios. According to the data collected in the TEC, the type/token ratio of Clark is 45/38 which is lower than 49/86, type/token ratio of Bush. This indicated that Clark uses a small vocabulary and has a narrow scope in his choice of words.

#### **4.1.2 Analysis Made by Baker**

In the second section of this paper, I have mentioned we need a deep analysis to further our understanding to translation. Baker set a model for us. She didn’t stop her footsteps to the phenomenon, but

dug out the essence of the difference, i.e. the cause behind them. After personal visits to both Bush and Clark, Baker discovered that the version is closely related with the translator's cultural orientation, his living circumstance and type of the source text. I will take cultural orientation as an example.

First, it makes impact on the translator's language style. Translators with different cultural orientation show different tendencies for the selections and comprehensions of words. As the choice of certain words and phrases differs, the personal style is coming into shape. Besides, the different choice of words may furtherly make a difference in syntactic structure and sentence pattern. Secondly, cultural orientation can influence a translator's selections of source text. Translation is in a sense a process of making decisions. If the source writing has several different versions, the decisions on what to translate will evolve into on which source text to translate. And this initial decision of translator will duly manifest the translator's style and undoubtedly influence the final appearance of the version. Thirdly, cultural orientation affects a translator's comprehensions of source text. The culture and customs of the society, the personal experiences, the knowledge of history all belong to the category of cultural orientation and will lead to the differences in the translator's understanding and perception of the author and his work.

## **4.2 Application of Corpus in General Translation**

Many translation course are devoted to specialized subject fields, such as legal translation, medical translation, or economic translation. Translators can extract a corpus of translations pertaining to a particular subject field and examine them to determine if a problem is specific to one particular source text or if it is difficult to decide as it is also manifesting itself in other texts dealing with a related subject. For example, when translating a text in the subject field of law, students might have difficulty in constructing the proper syntax because one particular source text is difficult to choose word. If it is not properly worded, the translation will influence the target language production.

Lavion believed that "with the help of corpora, the translator can easily find out the most suitable words and expressions" (78). Besides, they can search some fixed phrases or collocations popular in the native country. What's more, corpora provides a better complementary resource that allows translators to see terms in a variety of contexts simultaneously, which in turn enable them to be more context-based and make the items in the translation more native. We often hear such a dialogue between students and teachers.

T: I'm afraid you have to use another word. This word is not suitable.

S: Sorry, sir. But I looked it up in the dictionary, it couldn't be incorrect.

T: I know. There is no grammar mistake in it, but people nowadays seldom use this word and it is not native. It is hard to tell the reason, but that is their habit. If you want to learn a foreign language well, you have to respect their habit and just imitate.

From the dialogue, we can learn that to be native is really important for foreign language learners, especially for translators because they bear the responsibility to convey everything significant either embodied or implied in the original and convey them accurately in the version. Besides, we can also see the translators must be responsible for their choice of items and the correspondent explanation, otherwise it might mislead the readers.

In conclusion, corpora can provide larger and more up-to-date resources and can supply translators with better materials that are needed in explanation.

## **5. ASSESSMENT OF THE METHODOLOGY IN TRANSLATION STUDIES**

Different people may have various opinions towards the same thing. That is also true when it comes to corpus-based translation studies. Since its very appearance, corpus has been connected with DTS (descriptive translation studies). Along with the development of corpus-based research, some researchers have expressed their concerns. Tymoczko warns against the possible danger of pursuing scientific rigor as

an end in itself through empty and unnecessary quantitative investigations. Malmkjaer also worries that the universals gained from corpus might restrict creative usages in translations. In a broad sense, corpus-based methodology is one branch of descriptive translation studies. Yet, DTS might also degenerate from descriptive research into prescriptive one. Kenny also admitted that corpus researches are not completely objective and the conclusions drawn by researchers are not totally indisputable since they involve the interpretation of researchers (69). Therefore, the researcher must be aware of its advantages and its limitations. In this way, the corpus can be utilized to its best. Then what are these merits and demerits?

### **5.1 Advantages in Translation Studies**

Corpora are increasingly being created and studied to answer questions arising within the pure and the applied branches of translation studies. One of the distinctive strengths of corpus-based translation studies lies in its flexible and sound methodology. Two additional features of corpus-based methodologies are versatility and comparability. They can be applied in different areas of study and for different purposes, while using similar techniques. As such, they have the potential for cutting across the separate fields of interest within translation studies and for encouraging intra- and inter-disciplinary communication. Already a corpus-based approach is emerging in the related disciplines of contrastive linguistics and bilingual lexicography, which specifically use translational corpora not to evaluate the product of translation or to improve translator's performance, but to investigate linguistic phenomena which are of interest to translation scholars and practicing translators. Conversely, the insights provided by corpus-based translation studies can enhance research in these disciplines, by providing evidence, for example, about language pair- and translation-specific phenomena. Moreover, the results of corpus-based translation studies can be used as sources of hypotheses for the study of intra-lingual mediated events, such as excerpting of the production of simplified readers, which can be of interest to the monolingual corpus linguist.

Second, I believe that, as corpus-based translation studies establishes itself as a coherent approach within the discipline, theorists, teachers, trainees and professional translators will realize that the relationship between description and practice is strong and fruitful. This is because, on the one hand, their descriptive analysis of corpora focuses on a large amounts of empirical data and regards the intuitive insights of individual practitioners as important sources of hypotheses, and on the other, the practitioners can draw on the findings provided by descriptive scholars to reflect on their own performance in the light of what has been identified as typical translational behavior. The fact that a corpus-based methodology forces scholars to refine and clarify their initially abstract and vague concepts should also make it easier for practitioners to participate in scholarly debates which, at the moment, seem remote and intimidating.

It is obvious that the former two advantages are described from the height of theory and seem to be abstract. To be concrete, in the field of translation studies, the advantages of the corpus-based approach consist in that corpora provides a broad sample of the originals and the versions extending over a wide selection of variables such as the author and the translator's gender, age, class etc. this allows generalizations to be made about language, author or translator as the corpus is as wide and as representative as possible. Besides, the usage of corpus-based approaches is convenient and efficient. A scholar who has access to a corpus can call up all the examples of a word or phrase from millions of words of texts in a few seconds. It can provide up-to-date information about language which is beneficial to lexical studies. Besides, the choice of word can be more precise since a large number of examples are examined. What's more, the ability to call up word combinations rather than individual words, and the existence of mutual information tools which establish relationship between co-occurring words mean that we can treat phrases and collocations more systematically than was previously possible. In general, in the field of translation studies, the advantages of the corpus-based approach consist in that it provides scholars with empirical data which enable them to make objective statements, rather than those that are subjective, or based upon the individual's own internalized cognitive perception of translation. Therefore, it is really an effective way to our comprehension of words and phrases and translation study.

## 5.2 Limitations of the Corpus-based Approach

Corpora and the software for processing them undoubtedly provide translation scholars with powerful tools for studying the very nature of translation. Some translation theorists have, however, sounded a note of caution. In accordance with the classification of corpora in the first part, we will describe the limitations of those three different types respectively.

Malmkjaer warns that the bulk of statistical evidence provided by corpora may lead scholars to “treat as marginal, if not exactly ignore, problematic cases.” He also argues that the selection of translated texts for inclusion in a parallel corpus can affect what the observer notices to an undesirable degree, and that “a parallel corpus still only provides, for each instance, the result of one individual’s introspection, albeit contextually and contextually informed, thus making a case for corpora that contain several translations of a single source text” Finally, Malmkjaer stresses 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” (Malmkjaer 152).

Comparable corpora too have their problems: 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 introduces to it through translation from another textual tradition. This difficulty is similar to one faced by scholars working with lesser-used languages: the only exemplars of many (non-literary) text types in Irish Gaelic, for instance, are translations, mainly from English; there are no native texts with which to compare translation. The effects of the economy of translation have also been felt by Maria Tymoczko (79), whose choice of texts for selection into their English-Norwegian bidirectional, parallel corpus has been limited by the fact that a large number and a wide range of text have been translated into Norwegian, but far less in the other direction.

Perhaps the greatest challenge that faces corpus-based research into translation stems from the fact that corpus linguistics has always been data driven: it has proceeded from the bottom up, using concrete facts to make generalizations about particular languages (Baker 232). Much current translation scholarship, however, proceeds top down: theorists are interested in finding evidence to support abstract hypotheses. Translation studies thus makes very particular demands on corpora, and ongoing research in translation studies may lead to new ways of looking at corpora, just as corpora are already leading to new ways of looking at translation.

## REFERENCES

- Baker, Mona. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*. In *Text and Technology: In Honour of John Sinclair*, ed. Mona Baker, Gill Francis & Elena Tognini-Bonelli. Amsterdam & Philadelphia: John Benjamins.
- Baker, Mona. (2007). *Corpus-based Translation Studies in the Academy*. *Journal of Foreign Languages*, 5.
- HU Zhuang-lin. (2005). *Linguistics: A Course Book*, 3rd edition. Beijing: Beijing University Publishing House. (in Chinese)
- Kenny, M. (2004). *Introducing Corpora in Translation Studies*. London/New York: Routledge.
- Lavion S. (2002). *Corpus-based Translation Study Theory, Findings, Application*. Amsterdam: Rodopi.
- Laviosa, Sara. (1998). The Corpus-based Approach to Translation Studies. *Special Issue of Meta*, 43, 4-7.
- LIAO Qi-yi. (2000). *Corpus-based Translation Studies*. Beijing: Foreign Languages Teaching and Research Press, (5), 59-64. (in Chinese)
- Malmkjaer, Kirsten. (1998). *Linguistics and the Language of Translation*. Edinburgh: Edinburgh University Press.
- Tymoczko, Maria. (1998). *Translation in a Postcolonial Context*. Shanghai: Shanghai Foreign Languages Education Publishing House.

ZHANG Mei-fang. (2001). *English-Chinese Translation Textbooks in China (1949-1998)*. Shanghai: Shanghai Foreign Languages Education Publishing House.

ZHANG Mei-fang. (2001). From Experience, Text to Deconstruction of the Concept of Translation - A Glimpse of Translation Textbooks in Western Countries. *The Journal of PLA Foreign Language Institute*. Luoyang, China.