New Trends in Translation Studies

Vol. 1

Phraseology in Corpus-Based Translation Studies

Meng Ji

Peter Lang

Translations of Cervantes' *Don Quijote* (1605) take pride of place among foreign literature in China. Despite the contrasts between the two cultures and the passage of four centuries the adventures and misadventures of the Castilian hero have always been popular with Chinese readers.

In this book a corpus-based stylistic study is used to explore two contemporary Mandarin Chinese translations of *Don Quijote*: those by Yang Jiang (1978) and Liu Jingsheng (1995). Utilising a micro-structural perspective this study suggests explanations for the surprising popularity of *Don Quijote* in China.

Meng Ji has a PhD from Imperial College London (2009) within the area of corpus-based translation studies focused on the study of phraseology in literary translations into Chinese. She is presently developing an interdisciplinary approach to corpus-based translation studies by integrating methodologies from disciplines including textual statistics, quantitative sociolinguistics and computational stylometry.

www.peterlang.com

Phraseology in Corpus-Based Translation Studies

New Trends in Translation Studies

Volume 1

PETER LANG
Oxford • Bern • Berlin • Bruxelles • Frankfurt am Main • New York • Wien

Phraseology in Corpus-Based Translation Studies

Meng Ji

PETER LANG
Oxford • Bern • Berlin • Bruxelles • Frankfurt am Main • New York • Wien

Bibliographic information published by Die Deutsche Nationalbibliothek.
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available on the Internet at
http://dnb.d-nb.de.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data:

Ji, Meng, 1982-
Phraseology in corpus-based translation studies / Meng Ji.
p. cm. -- (New trends in translation studies ; vol 1)
Includes bibliographical references and index.

1. Cervantes Saavedra, Miguel de, 1547-1616--Appreciation--China. 2.
Cervantes Saavedra, Miguel de, 1547-1616--Criticism and interpretation.
3. Cervantes Saavedra, Miguel de, 1547-1616. Don Quixote--Translations
into Chinese. 4. Corpora (Linguistics) 5. Translating and
interpreting. I. Title.

PQ6349.C6J52 2010
495.1'802--dc22

2010007058

ISSN 1664-249X
ISBN 978-3-0353-0014-7

Hochfeldstrasse 32, CH-3012 Bern, Switzerland
info@peterlang.com, www.peterlang.com, www.peterlang.net

Printed in Germany

Table of Contents

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

Tables and Diagrams

Abbreviations

AI	Archaic idioms
CCDQ	Parallel Castilian and Chinese Corpus of *Don Quijote* Part I
CLAS	Chinese Lexical Analyser System
CMT	Context-motivated theory
CTC	Comparable translational corpora
DV	Dependent variable
FCEX	Chinese four-character expressions
FIG	Figurative idioms
I-AI	Instantiated archaic idioms
I-FIG	Instantiated figurative idioms
IDV	Independent variable
LBS	Lexical segmenter
LCMC	Lancaster Corpus of Mandarin Chinese
MP	Morphologically patterned four-character expressions
POA	Problem-oriented annotation scheme
PTC	Parallel translation corpora
SB	Semantically bipartite four-character expressions
SCT	Segmentation of Chinese text
SP	Shortened phrases
SS	Syntactically schematic four-character expressions
SSY	Structurally symmetrical four-character expressions

Acknowledgements

The completion of this book owes a great deal to the unfailing support and encouragement of many people. I would like to give my heartfelt thanks to my PhD supervisor, Dr Juan Antonio Lalaguna, who is a great expert in the field of translation history and Spanish literature and has provided a unique perspective and invaluable guidance throughout the preparation of this work. His wisdom and encouragement are very much appreciated.

I am also grateful for the professional support from people working in the Department of Humanities at Imperial College London, especially Head of Programme Mr Mark Shuttleworth and Dr Jorge Díaz Cintas. All of them have inspired me in one way or another through their interest in my work.

I extend my sincere thanks to the Global COE Program of Gender Equality and Multicultural Conviviality at Tohoku University, Japan, especially my postdoctoral mentor Professor Ohnishi Hitoshi. The free research environment and generous financial support they provided were crucial for the smooth and timely delivery of the book.

I owe my sincere gratitude to Professor Glenn Hook at the National Institute of Japanese Studies, and the University of Sheffield, UK. His understanding and support to my research as a young academic has provided a unique opportunity for me to develop my future career.

I would like to thank my editors at Peter Lang International Academic Publishers for the efforts they put into reviewing the manuscript in a most efficient and professional manner.

Last but not least, I offer this book to my beloved parents Ji Cheng Quan 纪 成全 and Jin Ze Rong 金 泽荣 – without their unconditional support and unstinting encouragement throughout those years, all of this would not have been possible.

Introduction

My strong interest in Spanish literature and culture from the very beginning has drawn me naturally to the topic under consideration here, namely the translation of Cervantes's *Don Quijote* into Mandarin Chinese. *Don Quijote*, undoubtedly the best well-known literary work of Spain's Golden Age literature, occupies today a position of pride among translations of foreign literature into Chinese. Despite the contrasts between the two cultures and the length of time that has elapsed, the adventures and misadventures of the Castilian hero have always been well-received by Chinese readers, which is quite a miracle considering China's changing socio-political climate in the course of the last century.

In pursuing this particular topic, my general aim is to help explain *Don Quijote*'s popularity in China. To this effect I intend to approach the subject from a micro-structural perspective; that is to say, through the detailed comparison of two different versions of *Don Quijote* (Part I) separated in time by nearly twenty years: Yang Jiang's translation, published in 1978, and Liu Jingsheng's, published in 1995. I should endeavour to underline the evolving nature of the language used in translating *Don Quijote* and the stylistic distinctiveness, if any, of the translators.

The representativeness of the two Chinese versions under investigation may be said to relate to the translator's keenness to be seen to be writing within the general socio-cultural background at that moment in time, which in turn would help to explain the wide acceptance generally afforded to Yang's version and indeed, the massive commercial success enjoyed by Liu's translation in recent years.

For the purpose of outlining the stylistic profile of the authors, I concentrate on a particular feature of the language used extensively in the two target texts, i.e. Chinese four-character expressions. They constitute a very typical category of Chinese phraseology, which includes a wide variety of phrasal units, ranging from morpho-syntactically patterned words or phrases to figurative idioms, archaism and variants of idiomatic expressions.

The corpus-driven methodology of analysis, which is still in its early stages of development and promises much through the use of ever more sophisticated corpus techniques, could be used to renovate and improve traditional textual analyses. First of all, this particular approach prioritizes the generation of quantitative textual data over the unwarranted claims made by personal impression. Also, the textual data extracted from the parallel corpus will then be used as raw material in the process of ascertaining underlying linguistic patterns in the corpus texts. This has been made possible thanks to the availability of a number of standalone text mining applications, which may help yield essential statistical information from the raw data, such as keyword indexing, collocation patterns, token/type ratio, node word distribution spectrum, mutual information score, etc.

Textual data, after being properly annotated and encoded, may then be used for statistical modelling to explore the relationship between different sets of data, as well as the structure of possible contextual or co-textual factors that may contribute to the linguistic patterns thus detected in the corpus texts. The exploratory role of statistical pattern modelling in corpus-based literary studies should not be underestimated. Given the fast developing nature of the discipline as a whole, the introduction of statistical methods will greatly ease the formulation and testing of theoretical hypotheses bearing on the nature of literary translation.

As I intend to show in detail, all this statistical information has greatly enhanced the capacity for detecting and identifying certain features of a language; extracting, organizing and classifying data, and analysing and comparing large quantities of translated texts. In this regard, this book may be broadly divided into four main tasks: (1) construction of a parallel Castilian–Chinese corpus of *Don Quijote*; (2) extraction and annotation of raw corpus data; (3) detection of underlying linguistic patterns; (4) computational modelling of co-textual factors to explore the cognitive rationale behind the translator's particular use of language.

Chapter 1 begins with a discussion of various types of translational corpora developed so far and freely available. It has been noted that despite the availability of a number of statistically built parallel corpora, small-scale topic-specific corpora still dominate the mainstream of corpus-based translation studies. They are usually built to a large extent manually and entail a great deal of effort and dedication on the part of corpus builders. This has been the case in the construction of the parallel corpus of *Don Quijote*.

To be more specific on the sort of technical challenges encountered in this project, I have focused on two basically related issues: segmentation of Chinese texts and parallel alignment of source and target texts. Inevitably the solutions to the problems encountered in the performance of these two tasks have significantly affected the subsequent design of the whole project.

Chapter 2 outlines a type of problem-oriented annotation scheme, which through an exchange of two major procedures in corpus data mining, i.e. the linguistic annotation of the corpus and the automatic extraction of corpus data, could help keep in balance the developing nature of many corpus annotation tools and the labour-intensive process of manual tagging. The problem-orientated scheme developed here is a corpus-based typological study of Chinese four-character expressions, which prepares the raw data and makes them more amenable to further quantitative exploration of the parallel texts.

Based on the large amount of linguistic data automatically retrieved from the parallel corpus, Chapter 3 proceeds to a quantitative analysis of the parallel corpus, which should lead to the identification of three general phraseological patterns separating the two Chinese translations: Yang's relatively higher use of morpho-syntactically patterned four-character expressions (Phraseological Pattern 1, PP-1 hereafter); Liu's preferred use of Chinese figurative language (PP–2), and Liu's relatively higher use of Chinese archaic idioms (PP-3). The remaining part of Chapter 3 is devoted to a detailed sample-based study of PP-1. It is argued that PP-1 signals quite effectively how the translator manipulates language at the morpho-syntactic level, i.e. Yang's initial attempt at assimilating the source text content into the target linguistic system.

Chapter 4 goes on to examine PP-2 in Liu's work, pointing to a stylistic modification of language at an ideational or conceptual level. Chapter 4 also offers a sample-based analysis of PP-3 as identified in Liu's work, suggesting that the use of archaism, as an important rhetorical device, may well also have an impact on the reception of the target texts. Through the comparative study, it becomes clear that in dealing with source text figurative expressions, Liu's approach has been more creative and target-text-oriented. He massively deploys Chinese figurative expressions, which sometimes may only match with source text expressions at a pragmatic or functional rather than semantic level.

Chapter 5 moves from qualitative analysis to a rather exploratory computational modelling of textual patterns that have been identified so far. My purpose here is to bring statistical ways and means to further the study of literary translations. Quantitative analysis produces a number of very interesting findings regarding the triangular relationship between the source text and the two target texts. While the frequency of use of figurative idioms in Liu's work is generally higher than that found in Yang's work, the enhanced figurative nature of Liu's work has proved to be more responsive to the figurative features in the source text than is the case in Yang's work.

With respect to archaisms, while Liu's use of archaic idioms may be reasonably predicted by the statistical models built up on the basis of the data retrieved from Yang's work and the original, noticeable differences come to the surface from the second half of Part I of the novel. This translational phenomenon is known as style variation, which brings me to the notion of a context-motivated theory of interpretation that would help in my search to integrate quantitative methods adapted from textual statistics and cognitive linguistics.

Chapter 6 uses a context-motivated theory (CMT) with a view to exploring the nature of Liu's idiosyncratic use of Chinese archaisms in the second half of his translation of the novel. The corroboration of such an interpretative framework, which draws upon Biber's quantitative study of register variation (Biber, 1998), has been put into operation through the use of the statistical procedure known as Categorical Principal Component Analysis. Despite its wide application in quantitative sociolinguistics, this procedure has been rarely used in translation studies. The qualitative analysis performed by CMT aims to explore rather than to test the latent structure of co-textual factors that may help to reveal the cognitive nature of style-shifting in literary translations.

In the course of defining a firm line of empirical research in corpus-based translation studies as outlined above, I argue for an interdisciplinary approach to the study of style, which as will be shown in this book, will prove to be quite useful to future research. It is believed that an interdisciplinary study of translating style will benefit greatly from the cross-fertilization of research methodologies and insights obtained from different but related disciplines, such as computational stylometry, corpus phraseology, textual statistics, quantitative sociolinguistics and cognitive stylistics.

To be specific, there are two core issues in the study showing an overt predisposition towards interdisciplinary treatment. These are, firstly, the expansion of the scope of corpus-based analysis of translation from lexicogrammar to phraseology, which has been made possible thanks to the technical development in the automatic retrieval of multiword expressions; and secondly, the deployment of quantitative analysis. One of the major contributions the present work makes to the study of literary translations is the establishment of a three-level quantitative analytical framework, namely raw data generation and classification, pattern modelling and cognitive exploration of the rationale behind the observed linguistic patterns.

CHAPTER 1

Construction of a Parallel Corpus of *Don Quijote* (Part I)

For the purpose of analysing the two selected Chinese versions of *Don Quijote* (Yang, 1978; Liu, 1995), there were three methodological options open to use. I could have followed a method of interlinear textual analysis, guided by my intuition regarding the subject matter (Allen, 1979; Power, 1967); or I could have set up a comparative model to examine linguistic equivalences that may occur between source and target texts (Leuven-Zwart, 1989: 151–81); or I could have adopted a corpus-based quantitative approach which would yield the textual data required to address a range of potential research questions.

I was always attracted by the last option. My own view is that many current approaches to the study of style in translation are somehow methodologically limited and not very productive. The range and scope of current textual analysis seem to have reached a limit, especially in terms of the methods that have been used so far. Is it worth therefore asking whether corpus tools may not be the way forward? These corpus tools do provide the kind of quantitative data that would enable researchers to engage in a convincing manner with soundly based empirical analysis. Such a new perspective is bound to elicit valuable insights into language behaviour in translation, as well as to facilitate a deeper understanding of the linguistic peculiarities and idiosyncrasies that may help define the particular style of a translator. In this regard, a corpus-based study of the two Chinese translations of *Don Quijote* has considerable potential to bring methodological benefit that would help advance and renovate textual studies.

The second reason why I have decided to follow this corpus-based approach to the study of the two Chinese translations is that, should this particular approach of mine prove to be a fruitful and efficient way of analysing translated texts, it would serve as inspiration to others and also be of practical help to any future corpus-oriented research in translation studies. It is my belief that the methodological framework used in corpus linguistic research is highly replicable and largely extensible. Any contribution made in the course of refining and advancing the current use of corpora in doing textual analysis would leave substantial space for the development of future work in this particular field of translation research.

1.1. Corpus construction in Translation Studies

Rapid advances in the construction of language corpora have provided us with a powerful technical platform for the study of literature and language never seen before. In particular, Translation Studies, as a growing academic discipline in its own right, situated at the crossroads of applied linguistics and comparative literature, has been called upon to turn its attention to the use of computerized parallel corpora as a highly efficient way of exploring translation-related activities and textual phenomena (Baker, 1995: 223–43).

Baker's proposal has been a milestone in the development of Translation Studies as an individual academic discipline through its introduction of modern computational techniques and quantitative linguistic material as furnished by language corpora, which will gradually revolutionize the way we observe, analyse and conceptualize human translation as a central cross-linguistic socio-cultural endeavour, as well as the subsequent translation products. To a certain extent, it may be said that the further expansion of the linguistic branch of translation studies could be explored effectively through the use of modern language resources and techniques, and the large amount of material evidence, real language in context, as accumulated in the form of translational corpora.

From initial efforts and discussions on the construction of parallel or comparable corpora of translation (McEnery and Wilson, 1993; Teubert, 1996: 238–64), the compilation and development of electronic resources for translation studies has shown a diverse trend, ranging from small-scale topic-specific corpora to massive statistically-built parallel corpora. Also, as a crucial feature of annotated corpora, the linguistic information added to raw corpus texts has become increasingly sophisticated. This in turn has prepared the ground for further expansion of the scope of research based on the generation and description of textual data from electronic translational corpora.

This section will first be looking retrospectively at the various types of translational corpora already developed or under development, with a view to assessing the current situation as regards corpus-based translation studies, as well as to pointing to possible directions for future research in the field based on ongoing parallel/multilingual corpus engineering and natural language processing. There are the different types of translational corpora in the field:

A. Parallel (Multilingual) Translational Corpora (Both ST and TTs) (PTCs)
1. Large-scale balanced PTC;
2. (Large-scale) genre-specific PTC, e.g. newspapers, legal documentation;
3. Large-scale non-balanced and non-classified PTC;
4. Large-scale and topic-specific PTC;
4.1. Raw corpora (for shallow quantitative analysis by using tools, e.g. WordSmith Tools);
4.2. Linguistically annotated corpora (for statistical analysis and pattern modelling);
4.2.1. Syntax annotation: lemmatization, POS tagging, syntactic parsing;
4.2.2. Semantic annotation: word sense tagging;
4.2.3. Pragmatics annotation: speech acts; speech and thought presentation;
4.2.4. Problem-oriented annotation.

B. Comparable Translational Corpora (TTs only) (CTCs), e.g. TEC;[1]
1. Large-scale balanced CTC, comparable to large-scale monolingual corpora, etc.;
2. Genre-specific CTC, like TEC.

[1] Translational English Corpora (TEC) is constructed by Centre of Translation and Intercultural Studies, University of Manchester. It may be accessed at <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>

Some of them have been made available for some time, while others are still under development at this moment. As we can see, the scale and diversity achieved so far in translational corpus engineering has been quite conspicuous for a relatively short period of time; especially when we compare it to the construction of large-scale monolingual corpora, e.g. in English (e.g. the Brown Corpus of American English, 1961), which was initiated more than three decades earlier than the former.[2] The advances made in parallel corpus construction would seem to be even more prominent, if we take into account the much higher levels of difficulty implied in solving technical problems relating to parallel text matching and alignment, especially working with typologically different languages (Piao, 2002: 207–30).

General speaking, we have two major types of translational corpora: parallel translational corpora (PTC) and comparable translational corpora (CTC). There has been some confusion in the literature regarding the establishment of a consistent terminological framework for corpus type categorization (Baker, 1999: 281–98). In my view, the differences between the two may be better described and understood by looking at their underlying structural features: while a PTC contains both the source and target texts, a CTC is a compilation of translated texts only, with a view to investigating the nature and regularities of translated language (Baker, 1995: 223–43).

Within each category, PTC or CTC can be further classified according to the text types that they may cover, whether they be large-scale balanced corpora or genre-specific corpora. The purpose of building large-scale balanced corpora is to investigate general linguistic features of the language in use; whereas the compilation of genre-specific translational corpora aims primarily to address research questions regarding specific aspects of translated language or within certain text domains. In theoretical terms, both types of subcorpora may be equally explored in the construction of a translational corpus platform; however, experience shows that unlike the construction of monolingual corpora (one language only), large-scale

[2] One of the earliest well-known parallel corpora was the English–Norwegian Parallel Corpus (ENPC) (1990s).