ROUTLEDGE-WIAS INTERDISCIPLINARY STUDIES

Corpus Methodologies Explained

An empirical approach to translation studies

Edited by
Meng Ji, Michael Oakes, Li Defeng and
Lidun Hareide

WIAS 早稲田大学高等研究所
Waseda Institute for Advanced Study

ROUTLEDGE

Corpus Methodologies Explained

An empirical approach to translation studies

Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

WiAS 早稲田大学高等研究所
Waseda Institute for Advanced Study

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication Data
A catalog record for this book has been requested

ISBN: 978-0-415-71699-4 (hbk)
ISBN: 978-1-315-69412-2 (ebk)

Typeset in Galliard
by Out of House Publishing
Printed in Great Britain by Ashford Colour Press Ltd

FSC
www.fsc.org
MIX
Paper from responsible sources
FSC® C011748

Corpus Methodologies Explained

This book introduces the latest advances in Corpus-Based Translation Studies (CBTS), a thriving subfield of Translation Studies which forms an important part of both translator training and empirical translation research. Largely empirical and exploratory, a distinctive feature of CBTS is the development and exploration of quantitative linguistic data in search of useful patterns of variation and change in translation. With the introduction of textual statistics to Translation Studies, CBTS has geared towards a new research direction that is more systematic in the identification of translation patterns; and more explanatory of any linguistic variations identified in translations. The book traces the advances from the advent of language corpora in translation studies to the new textual dimensions and the shift towards a probability-variation model. Such advances in CBTS have enabled in-depth analyses of translation by establishing useful links between a translation and the social and cultural context in which the translation is produced, circulated and consumed.

Meng Ji is Associate Professor/Reader at the Department of Chinese Studies at the University of Sydney.

Lidun Hareide is Assistant Professor at Møreforsking AS, Volda, Norway.

Defeng Li is Professor of Translational Studies at the University of Macau, China.

Michael Oakes is Reader in Computational Linguistics at the University of Wolverhampton, UK.

Routledge-WIAS Interdisciplinary Studies
Edited by Hideaki Miyajima and Shinko Taniguchi,
Waseda University, Japan

1. Corporate Crime in China
History and contemporary debates
Zhenjie Zhou

2. Why Policy Representation Matters
The consequences of ideological proximity between citizens and their governments
Willy Jou, Luigi Curini and Vincenzo Memoli

3. Electoral Survey Methodology
Insight from Japan on using computer assisted personal interviews
Edited by Masaru Kohno and Yoshitaka Nishizawa

4. Corpus Methodologies Explained
An empirical approach to translation studies
Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

5. Clans and Religion in Ancient Japan
The mythology of Mt. Miwa
Masanobu Suzuki

Acknowledgements

Empirical translation studies represents a rapidly growing field of cross-lingual and cross-cultural studies. An important feature of recent development in empirical translation studies is the use of statistical research methods in the exploration of translational features at linguistic and textual levels. Recurrent patterns identified and extracted from quantitative translations bring valuable and much-needed insights into effective translation strategies and techniques to inform the teaching of practical translation, the development of translation theories and the design of new translation technologies and software to support cross-cultural communication. This book represents the joint effort of advancing empirical translation studies among four translation scholars from Australia, Norway, China and the UK.

The conceptualisation of this project was discussed and finalised among the co-authors when the first author of the book, Meng Ji, was affiliated with the Waseda Institute of Advanced Studies (WIAS), Waseda University, Tokyo, in 2012. As the first translation scholar to be awarded the prestigious WIAS Research Fellowship, she benefited greatly from the world-class research environment provided by WIAS, which was multi-disciplinary, stimulating and truly rewarding. As the first title of translation studies in the Routledge-WIAS Interdisciplinary Studies series, the publication of the book on the tenth anniversary of the foundation of WIAS reflects the tradition and aspiration of the world-leading research institute, i.e. to pursue research excellence to advance better cross-cultural communication and understanding.

List of tables

List of figures

Contents

Introduction to *Corpus Methodologies Explained*

An empirical approach to translation studies

by Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

Amidst the growing body of empirical translation studies and corpus translation studies in particular (CTS), the current volume represents the latest research in key areas of CTS such as machine translation (Chapter 1, Michael Oakes), translation genre variation and shifting (Chapter 2, Meng Ji), translation stylistics (Chapter 3, Defeng Li) and translation universals, including testing of the Gravitational Pull Hypothesis (Chapters 4–5, Lidun Hareide). The structural organization of the book is balanced between theoretical discussion and illustrative case studies. It aims to provide a focused introduction to the research paradigms which prevail in current CTS, i.e. from the development of statistical machine translation to the exploration of recurrent translational patterns called translation universals. From Chapter 1 to Chapter 5, the levels of theoretical postulation increase, as the research methods used gradually move from essentially corpus-driven (Chapter 1 and 2), via corpus-assisted (Chapter 3) to typical corpus-based translation studies (Chapter 4 and 5).

The distinction between these three main research paradigms within the current CTS, which is evolving rapidly, is largely based on the purposes and aims of the use of empirical evidence in the study of corpora. Throughout the book, the frequency-based analysis of language corpora, monolingual or multilingual, plays an instrumental role in the corpus analysis of translation. In corpus-driven translation research as exemplified by Chapter 1 (on statistical machine translation), and from a different perspective by Chapter 2 (on genre studies), corpus analysis tends to focus on the statistical modeling of linguistic and textual patterns which lead to the development of new computational language models, conceptual dimensions and analytical instruments in translation studies.

Chapter 1 offers an overview of important research paradigms in machine translation, i.e. rule-based machine translation, example-based machine translation, translation memories and statistical machine translation. The significance of this chapter is that it uses case studies in multiple languages to illustrate the rationale behind competing language and translation models. The linguistic analysis is enhanced with detailed explanations of relevant statistical procedures which allow readers to obtain an in-depth understanding of machine translation systems from Google Translate to popular computer-assisted translation (CAT) language resources like translation memories.

Chapter 2 presents a quantitative analysis of contrastive distributional patterns of part-of-speech categories in monolingual English and Chinese corpora, and corpora which contain Chinese translations of English source texts. The corpus study adopts an essentially corpus-driven approach to the analysis of the quantitative data extracted from large-scale language corpora. The statistical analysis constructs three distinctive genre classification models for English, Chinese and translational Chinese as represented by the three large-scale corpora under study. The analysis shows that English written genres have a clear focus on techniques involved in the delivery of textual information. By contrast, the genre system of original Chinese gives more emphasis to language style rather than the delivery of actual textual information. The focus on the quality and stylistic features of the language implies that the prioritization of the aesthetic value of writing exists widely in the modern Chinese genre system, which is a long-standing tradition in the Chinese language and cultural system.

The exploratory statistical analysis of translational Chinese genres reveals that the genre system of translational Chinese is more complex than that of the original languages, as three sets of criteria have emerged in the corpus analysis which underline the configuration of the translational Chinese genre system. These are (1) features related to the communicative function of translation, i.e. explicitation, simplification and interactivity; (2) source-text oriented textual and linguistic features; and (3) target-text oriented textual and linguistic features. Such corpus findings suggest that translation is a highly purposed and complex system. If we consider translational textual features like explicitation, simplification and interactivity as essentially target-audience oriented translation strategies and tactics, the corpus-driven analysis in Chapter 2 seems to suggest that the contemporary Chinese translational genre system is overwhelmingly oriented towards the target language and culture.

Chapter 3 offers an overview of translation stylistics, an important area of corpus translation research. It deploys descriptive analyses widely used in corpus-based translation studies such as the type-token ratio, standardized sentence length variation and normalized word frequency lists to explore contrastive stylistic profiles of different target versions of a source text (the case study used is from two early English translations of the Chinese literary classic *Dream of the Red Chamber* or *Hongloumeng*). The methodological framework of Chapter 3 is distinct from that of Chapter 2 in that the frequency-based analysis used in Chapter 3 is largely descriptive, whereas the quantitative methods used in Chapter 2 are more exploratory, aiming to construct new analytical instruments to make necessary preparations for further theoretical development. If we could consider the type of corpus translation research exemplified by Chapter 2 as essentially corpus-driven, the focus of the analytical strategies of Chapter 3 is to detect differences between paired translations and the source text. An important observation made in Chapter 3 regards the further analysis of the corpus findings at a social and cultural level; in other words, how to interpret the stylistic differences identified between different translations within the larger target social and cultural background – a methodological concern which points to the strengths and limitations of many similar studies on translation stylistics.

Chapters 4 and 5 reflect the theoretical branch of translation studies, which focuses on general tendencies in translations. These chapters offer two corpus-based studies of universally existent tendencies in translation, i.e. translation universals, which represent the main focus of corpus-oriented descriptive translation research. The study tests the previously untested Gravitational Pull Hypothesis (Halverson 2003, 2007, 2009, 2010). Since the Gravitational Pull Hypothesis intends to reconcile two seemingly opposing translation tendencies, full testing of this hypothesis entails testing of the mutually exclusive Over-representation of Target-Language Specific Features Hypothesis (Baker 1993, 1996) and the Unique Items Hypothesis (Tirkkonen-Condit 2001, 2004). Consequently, all three hypotheses posited on the suggested translation universal "over- or under-representation of target-language specific features" in translation studies are tested. In order to test these hypotheses, two comparable parallel corpora having the same target language but different source languages are needed. The feature to be tested must be unique to the target language in one of the language pairs, but must have a grammatical counterpart in the source language in the other language pair.

As a typical corpus-based study, Chapter 4 presents the design of the study, outlines the three hypotheses, the language pairs and the corpora used, as well as the grammatical structure that is tested in the case studies. In addition, Chapter 4 presents the first case study where the mutually exclusive Unique Items and Over-representation of Target-Language Specific Features hypotheses are tested. The Spanish gerund is used as the test object. In order to establish empirically that the Spanish gerund in fact constitutes a unique item in relation to Norwegian, a comparative study of 20 per cent of all of the Spanish gerunds in each text of the Norwegian-Spanish Parallel Corpus and their Norwegian counterparts is conducted.

Chapter 5 builds on the results from Chapter 4 in order to test the Gravitational Pull Hypothesis on the language pairs English-Spanish and Norwegian-Spanish, using the same grammatical structure (the Spanish gerund). The work presented in Chapters 4 and 5 demonstrates that the Gravitational Pull Hypothesis can be empirically tested using corpus data, and that the five core predictions of this hypothesis received support. In addition, the Unique Items Hypothesis was not upheld in translations from Norwegian, both with regards to frequent and to prototypical gerunds, and this raises important questions as to when this latter hypothesis applies and when it does not, and whether it is needed at all.

Since its inception in the 1980s, CTS has been one of the fastest growing research and teaching areas in translation studies as an independent academic discipline. The development of CTS owes much to the growing sophistication and specificity of related research methodologies. The current volume highlights three key research paradigms or sets of analytical strategies widely used in CTS: corpus-driven (statistical machine translation; exploratory corpus statistics), corpus-assisted (translation stylistics and parallel corpus comparison) and corpus-based (translation universal features or general translation tendencies) approaches. As the case studies used in each chapter demonstrate, each approach has its strengths and limitations, which reflects the very nature of empirical translation research. The delimitation of these three sets of distinct yet related research schemes contributes to the further expansion of the field, which relies to a large extent on the development of a robust, integrative and innovative methodological system for empirical translation research.

References

Baker, Mona (1993). Corpus linguistics and translation studies: Implications and applications. In Gill Francis, Mona Baker and Elena Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, pp. 233–250.

Baker, Mona (1996). Corpus-based translation studies: The challenges that lie ahead. In Harold L. Somers (ed), *Terminology, LSP and Translation: Studies in Language Engineering*. Amsterdam: John Benjamins, pp. 175–186.

Halverson, Sandra (2003). The cognitive basis of translation universals. *Target* 15(2): 197–241.

Halverson, Sandra (2007). Investigating Gravitational Pull in translation: The case of the English progressive construction. In Riita Jääskeläinen, Tiina Puurtinen and Hilkka Stotesbury (eds), *Text, Processes, and Corpora: Research Inspired by Sonja Tirkkonen-Condit*. Savonlinna: Publications of the Savonlinna School of Translation Studies 5, pp. 175–196.

Halverson, Sandra (2009). Elements of doctoral training: The logic of the research process, research design and the evaluation of design quality. *The Interpreter and Translator Trainer* 3(1): 79–106.

Halverson, Sandra (2010). Cognitive translation studies: Developments in theory and method. In Gregory M. Shreve and Erik Angelone (eds), *Translation and Cognition*. Amsterdam: John Benjamins, pp. 349–369.

Tirkkonen-Condit, Sonja (2001). Unique items – over – or underrepresented in translated language? In *The Third International EST Congress*, Copenhagen, Denmark.

Tirkkonen-Condit, Sonja (2004). Unique items – over – or underrepresented in translated language? In Anna Mauranen and Pekka Kujamäki (eds), *Translation Universals: Do They Exist?* Amsterdam/Philadelphia: John Benjamins, pp. 177–184.

1 The need for corpora in machine translation

Michael P. Oakes

Abstract

In this chapter we show that corpora, particularly parallel bilingual corpora, are essential in the development of automatic machine translation (MT) systems, whether translation memories, example-based or statistical. Specific topics examined are the Europarl corpus, similarity measures for sentence matching, the Hofland sentence aligner, automatic generalisation of translation examples through paraphrasing and the discovery of templates, statistical methods of building bilingual dictionaries, the development of MT for less-resourced languages and the evaluation of MT systems.

1. Introduction

This chapter will show that corpora, particularly parallel bilingual corpora, are almost the *sine qua non* of automatic machine translation (MT). In section 2 we will examine the four main paradigms in automatic MT, namely rule-based MT (the least dependent on corpora), translation memories (not strictly speaking "true" MT, but widely used by professional translators), example-based MT and statistical MT. In section 3 Europarl is described, a multilingual corpus built from transcripts of sessions of the European Parliament, especially for developing MT systems. In section 4 we describe how translation memory (TM) and example-based MT systems depend on finding the most similar stored examples to the sentence we wish to translate. This requires "matching", or the determination of how similar two sentences are to each other. Section 5 covers sentence-level alignment, or discovering automatically which sentence(s) of one language in a parallel corpus match which sentence(s) of the other. As a case study, we will consider Hofland's aligner, designed originally for English and Norwegian. Since gathering enough parallel corpus data can be a problem, in section 6 we discuss the automatic generalisation of translation examples – how can we make a single stored example represent a whole set of sentences? The techniques described include paraphrasing and the discovery of templates. In section 7 we look at statistical methods of building the bilingual dictionaries, with frequency information, that are widely used in automatic MT. In section 8 the topic is the development of MT for "minority" or less-resourced languages, using Cebuano and Mapudungun as case studies. Finally, in section 9, we will look at how MT systems are evaluated – to help us identify the "best" system, and to learn which improvements are possible.

2. Paradigms for machine translation

In this section we will consider the main broad methods which have been used for MT. The earliest systems were called rule-based systems, because they were heavily dependent on language-pair specific rules. At about the same time, three other paradigms were introduced. Two of these, translation memories and example-based MT, both stored large numbers of previous translations against which new translations could be compared. The difference between them was that human translators took the final decision as to which parts of the previous translations could be reused, while in example-based MT, the machine decides which fragments to reuse. Statistical MT uses purely numeric data, derived from corpora, about the probabilities of the translations of individual words (which may have more than one counterpart in the other language) and the fluency of translated text as a function of the word adjacencies in it. While traditional statistical MT systems used information about individual word correspondences, a more recent development is to consider phrase correspondences across languages.

2.1 Rule-based machine translation

The earliest MT systems, prior to the 1990s, were called rule-based systems, and were built using linguistic knowledge in what Somers (2009) calls a rationalist approach. At that time corpora were relatively rarely used in the development of MT systems, not really coming into their own until the advent of what Somers describes as the data-driven or empirical approaches which came to the fore in the 1990s. However, many people at that time were looking at how the use of "controlled languages", where the range of vocabulary and allowed grammatical structures was both limited and fixed, could improve the performance of rule-based systems. The idea was that controlled languages would contain relatively little ambiguity, and thus would be easier for MT systems to process. Various groups at this time did make use of corpora to define the range of vocabulary and grammar that an MT system should work with, and thus they had (and have) a role in developing controlled languages. The TAUM group in Montréal used the set of words and structures in a 70,000-word corpus to define a controlled language for MT, and the Eurotra MT Project used the Europarl corpus for a similar purpose (Somers, 2009). We will briefly take a look at an example of a rule-based system which is taken from Arnold et al. (1993:76–77).