New Frontiers in Translation Studies

Vincent X. Wang
Lily Lim
Defeng Li *Editors*

New Perspectives on Corpus Translation Studies

Springer

New Frontiers in Translation Studies

Series Editor

Defeng Li
Center for Studies of Translation, Interpreting and Cognition,
University of Macau, Macao SAR, China

Translation Studies as a discipline has witnessed the fastest growth in the last 40 years. With translation becoming increasingly more important in today's glocalized world, some have even observed a general translational turn in humanities in recent years. The New Frontiers in Translation Studies aims to capture the newest developments in translation studies, with a focus on:

- Translation Studies research methodology, an area of growing interest amongst translation students and teachers;
- Data-based empirical translation studies, a strong point of growth for the discipline because of the scientific nature of the quantitative and/or qualitative methods adopted in the investigations; and
- Asian translation thoughts and theories, to complement the current Eurocentric translation studies.

Submission and Peer Review:

The editor welcomes book proposals from experienced scholars as well as young aspiring researchers. Please send a short description of 500 words to the editor Prof. Defeng Li at Springernfits@gmail.com and Springer Senior Publishing Editor Rebecca Zhu: Rebecca.zhu@springernature.com. All proposals will undergo peer review to permit an initial evaluation. If accepted, the final manuscript will be peer reviewed internally by the series editor as well as externally (single blind) by Springer ahead of acceptance and publication.

More information about this series at https://link.springer.com/bookseries/11894

Vincent X. Wang · Lily Lim · Defeng Li
Editors

New Perspectives on Corpus Translation Studies

Springer

Editors
Vincent X. Wang
Department of English
University of Macau
Macao, Macao

Lily Lim
School of Languages and Translation
Macao Polytechnic Institute
Macao, Macao

Defeng Li
Department of English
University of Macau
Macao, Macao

ISSN 2197-8689 ISSN 2197-8697 (electronic)
New Frontiers in Translation Studies
ISBN 978-981-16-4917-2 ISBN 978-981-16-4918-9 (eBook)
https://doi.org/10.1007/978-981-16-4918-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Acknowledgements

The editors express sincere thanks to Prof. Defeng Li, Series Editor of New Frontiers in Translation Studies, for his full support and timely advice throughout the process of completing this edited volume. Our heartfelt appreciation extends to the whole editorial team of Springer, including Vidyaa Shri Krishna Kumar, Umamagesh Perumal, Rebecca Zhu, Carolyn Zhang, Nobuko Hirota and Fiona Wu, for their efficient communication and professional handling of the publishing matters, from the initial stage to final production. We are always grateful to all the anonymous reviewers who have provided their constructive comments and good suggestions that contributed so much to the quality of the volume.

Without any doubt, we are deeply indebted to all our contributors, who presented their original research and shed new insights in this volume.

Our heartfelt thanks go to each and every individual who is involved in the compilation of this volume.

Introduction

This collected volume consists of twelve chapters, contributed by notable researchers from Europe, Russia, China, Hong Kong and Macao, working on corpora of a variety of languages—e.g. German, Dutch, Italian, Spanish, Russian, Finnish, and, of course, English and Chinese. The chapters are classified into the three clusters of themes: (a) translation pedagogy, (b) translation norms and styles and (c) cognition and translation equivalents.

Centring on the theme of translation pedagogy, Sara Laviosa and Gaetano Falco propose to redraw James Holmes's (1988) map of translation studies to make Translation Pedagogy—which is proposed both for Language Teaching and Translator Training—the primary sub-branch of Applied Translation Studies. Laviosa and Falco argue for the shift from "translation *for* language teaching towards the emerging view of translation *in* language teaching" (italics in the original), in the European context of multilingualism, where plurilingual individuals' ability to perform cross-linguistic mediation is required by the Common European Framework of Reference for Languages (CEFR). Insomuch as plurilingal learners/users need a level of translation competence, translator training should enter the language classroom. This means a greatly increased need for translator training, and, at this juncture, corpus-assisted training can play a vital part, as illustrated and argued by Laviosa and Falco.

Moving from Laviosa and Falco's visions for the future of translator training programmes, Lily Lim contributes to translation pedagogy by evaluating the potential of corpus tools and resources, from the perspective of (trainee) translators, to resolve the translation problems surrounding the suffix –ism. The suffix entails a broad range of meaning, as noted by Lim, and poses challenges for (novice) translators. Lim demonstrates that a large-scale English-Chinese parallel corpus contains a decent repertoire of –ism words and their corresponding lexical items in Chinese. Through the lens of the Chinese lexical items, inquisitive translators should be able to tease out the meanings conveyed by –ism words, using rather basic corpus tools and skills. In the same vein, Vincent Wang uses the same parallel corpus to study the prefix de-, whose senses are identified and classified by the Chinese lexical items corresponding to de- verbs. The studies of affixes by Lim and

Wang extended the scope of research on sense disambiguation with translated texts (see Johansson 2007: 28) from words to morphemes, and further demonstrate that (big) data-driven research of translation between typologically distant languages at the morphological level is in fact cross-disciplinary. It involves and also informs contrastive language studies, translation studies and translators' ICT literacy. Returning to the theme of translation pedagogy, the chapters by Laviosa and Falco, Lim, and Wang all point to the pressing need for developing empirical studies that look into translator's (effective) interaction with corpus technology.

On the theme of translation norms and styles, the five chapters in this volume made distinct contributions. Libo Huang presents an overview of the literature on translator's style in terms of the types, (new) trends and diachronic development. Huang further draws on his expertise in the field to pinpoint the directions for future development. The other four chapters in this section involve the construction of specific parallel corpora for revealing translators' style and translation norms. The Dutch Parallel Corpus (DPC) 2.0 is introduced by Ryan Reynaert, Lieve Macken, Arda Tezcan and Gert De Sutter, which is updated to 2.0 purposely to include exceptionally rich metadata. Not only the direction of translation is determined—i.e. from Dutch to English, from Dutch to French or the other way round—also specified is the translator's gender, age group, academic degree, experience as a translator, and his or her L1 and L2 in the translation assignment. The metadata further covers the translator's status as a freelancer or in-house staff, domain/s of expertise, style guides used, revision, and whether it is a collaborative translation task. In addition, specific text-related information—e.g. text provider, channel, intended audience, register—is gathered. The arduously and meticulously updated corpus with extensive metadata enables systematic investigation of translation norms and features in relation to the parameters of the translator and the texts involved. Moving the setting from Europe to Hong Kong, Oi Yee Kwong interrogates her self-constructed Chinese–English two-way parallel corpus to contrast the lexical choices made by interpreters and translators. The corpus is composed of speeches given at the Hong Kong Legislative Council (LegCo). More specifically, the interpreting data refers to the transcribed speeches which were verbally delivered by the officials and rendered by the simultaneous interpreters at the LegCo, while the translation data is the original speeches and the written translation published in the proceedings. Subtle semantic differences were detected between the interpreted and the translated texts, and Kwong shares her insights by interpreting the results in the light of the cognitive constraints imposed on the interpreters. The extensive chapter by Maria Kunilovskaya and Gloria Corpas Pastor investigates the correlation between translationese and register in English-to-Russian translation. Built on the investigators' previous studies and insights on translationese, Kunilovskaya and Corpas Pastor constructed a macro-corpus that contains four distinct registers—general media, popular science, fiction and commentary—which are analysed in terms of morphological, syntactic and text-level properties. Using a range of sophisticated tools—e.g. supervised and unsupervised machine learning, data visualisation, text classification models—the study shows a clear distinction between the translated and the non-translated texts, and identifies register-specific features of translationese. The chapter by Kan Wu and Dechao Li zooms is to investigate a single genre—i.e. martial arts fiction—as well as to a specific linguistic feature—normalisation—while four different translators of the works by the same martial arts novelist (Louis Cha) are compared. The results show different degrees of lexical normalisation invoked by the translators, and this leads to the investigators' thought-provoking discussion on the findings in relation to readers' reception of the translated work and the translators' motivations.

The final theme concerns cognition and translation equivalents. Chu-Ren Huang and Xiaowen Wang revisit the principles of translation proposed by Yan Fu from the multi-brain and cross-cultural perspectives. Huang and Wang provide readers a fresh look at the order of Yan's principles in terms of importance. The findings on different cultural connotations of "first" and "three" in English and Chinese are revealingly supported by evidence from comparable corpora. The cognitive perspective is also taken by Zi-yu Lin in his examination of the Chinese character 柴 *chái*. Lin illustrates with abundant examples—from big data as well as from small data—that metonymic chains (转喻链 *zhuǎnyù liàn*) are indeed at work in the practice of Chinese–English translation. Coming to the perennial topic of translation equivalents, Mikhail Mikhailov points out that translation scholars need to be mindful of the direction of translation in their study of parallel corpora. Mikhailov proposes a pair of terms—translation equivalence (Teq) and translation stimulation (Tst)—to mark the difference. The value of the terms is supported by interesting Finnish<->Russian translation examples taken from the bidirectional parallel corpus. Alignment at the sentence level is examined by Silvia Hansen-Schirra, Jean Nitzke and Silke Gutermuth, in a parallel corpus that contains Standard German and German Easy Language (Geasy Corpus). Although intralingual parallel corpus presents an under-researched area, Hansen-Schirra, Nitzke and Gutermuth argue with corpus evidence that Geasy Corpus enables the studies of the translation strategies between the two varieties of German and leads to the description of the characteristics of Easy Language translation.

In summary, the volume features recent attempts to construct corpora for specific purposes—e.g. multifactorial Dutch (parallel), Geasy Easy Language Corpus (intralingual), HK LegCo interpreting and Translation corpus—and showcases sophisticated and innovative corpus analysis methods. It proposes new approaches to address classical themes—i.e. translation pedagogy, translation norms and equivalence, principles of translation—and brings interdisciplinary perspectives—e.g. (contrastive) linguistics, cognition and metaphor studies—to cast new light to translation problems. It is our aim that the volume can serve as a timely reference for the researchers as well as postgraduate students who are interested in the applications of corpus technology to solving translation and interpreting problems.

May 2021

Vincent X. Wang
Lily Lim
Defeng Li

References

Holmes, J. S. (1988). *Translated! Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Contents

Translation Pedagogy

Translation Norms and Styles

Cognition and Translation Equivalents

Editors and Contributors

About the Editors

Vincent X. Wang associate professor of the University of Macau and a NAATI-certified translator, received his MA and Ph.D. in Applied Linguistics from the University of Queensland (2006). His research interests are in interlanguage pragmatics, corpus-based contrastive language studies, and discourse and pragmatics in translation. He published journal articles in *Sage Open*, *Target*, *Journal of Language, Literature and Culture*and TESOL-related periodicals, book chapters with Springer, Routledge and Brill, conference papers with PACLIC and CLSW, and a monograph *Making Requests by Chinese EFL Learners*(John Benjamins). His recent research draws on big data and corpus linguistics methodologies to investigate language properties, discourse, and the use of conceptual metaphors in social events such as COVID-19.

Lily Lim holds a Ph.D. in applied linguistics (University of Queensland), a master's degree in software engineering (University of Macao), Certificate of Training Techniques (Escolas da Armada, Portugal) and Certificate of Chinese–Portuguese Interpreting Training (Comissão Europeia). She has been both a practising interpreting and trainer for conference interpreters for two decades. She is currently Associate Professor and Coordinator of the Chinese–English Translation Program at the School of Languages and Translation, Macao Polytechnic Institute. Her recent research covers computer-assisted interpreter and translator training, and corpus-based language studies. She has published papers in *ReCALL*, *Babel*and *The Interpreter and Translator Trainer*; chapters with Springer, Rodopi, Routledge and Cambridge Scholars Publishing, and an edited book with Springer and a monograph with Bookman.

Defeng Li is Professor of translation studies and Director of the Centre for Studies of Translation, Interpreting and Cognition (CSTIC) at the University of Macao. Prior to his current appointment, he served as Chair of the Centre for Translation

Studies and Reader in Translation Studies at SOAS, University of London; Director of the MA in Translation and Associate Professor at the Chinese University of Hong Kong; Dean and Chair Professor at Shandong University; and (Visiting) Chair Professor at Shanghai Jiao Tong University. He is currently President of World Interpreter and Translator Training Association (WITTA). He has researched and published extensively in the fields of cognitive translation studies, corpus-assisted translation studies, curriculum development in translator training, research methods in translation studies, professional translation (e.g. business, journalistic, legal translation), as well as second language education.

Contributors

Gloria Corpas Pastor Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK;
University of Malaga, Malaga, Spain

Gert De Sutter Empirical and Quantitative Translation and Interpreting Studies (EQTIS), Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Gaetano Falco Dipartimento LeLiA, Università Degli Studi Di Bari Aldo Moro, Bari, Italy

Silke Gutermuth Johannes Gutenberg University, Mainz, Germany

Silvia Hansen-Schirra Johannes Gutenberg University, Mainz, Germany

Chu-Ren Huang Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

Libo Huang Xi'An International Studies University, Xi'An, Shaanxi, China

Maria Kunilovskaya Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK

Oi Yee Kwong Formerly The Chinese University of Hong Kong, Hong Kong, China

Sara Laviosa Dipartimento LeLiA, Università Degli Studi Di Bari Aldo Moro, Bari, Italy

Dechao Li Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

Lily Lim School of Languages and Translation, Macao Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macau

Zi-yu Lin Macao Polytechnic Institute, Macao, China

Lieve Macken Language and Translation Technology Team (LT3), Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Mikhail Mikhailov Languages Unit, Tampere University, Tampere, Finland

Jean Nitzke Johannes Gutenberg University, Mainz, Germany;
University of Adger, Kristiansand, Norway

Ryan Reynaert Empirical and Quantitative Translation and Interpreting Studies (EQTIS), Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Arda Tezcan Language and Translation Technology Team (LT3), Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Vincent Xian Wang Department of English, University of Macau, Avenida de Universidade, Taipa, MACAU, Macau SAR

Xiaowen Wang School of English Education, Guangdong University of Foreign Studies, Guangzhou, China;
Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong SAR, China

Kan Wu School of Foreign Languages, Zhejiang University of Finance and Economics Dongfang College, Haining, Zhejiang, China

Translation Pedagogy

Using Corpora in Translation Pedagogy

Sara Laviosa and Gaetano Falco

Abstract In view of recent developments in applied linguistics and translation studies, this paper argues that translation pedagogy is now a broad and burgeoning area of transdisciplinary research and practice whose goal is to address questions concerning teaching methods, testing techniques and curriculum planning in language teaching as well as translator training. Starting from this inclusive stance, the paper firstly proposes to redraw James S. Holmes's outline of applied translation studies. Secondly, it provides a critical analytical overview of corpus use in pedagogical translation at the advanced levels of linguistic competence in language B, as described in the Companion Volume to the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2020). Thirdly, it overviews exemplary corpus use in translator training. These two sub-domains of applied corpus-based translation studies are viewed through the lens of two major competence models that have been elaborated in Europe in recent years. So, corpus use in language teaching is illustrated in the light of the new descriptors of the CEFR (Council of Europe 2020). Corpus use in translator training is illustrated in the light of the new European Master's in Translation (EMT) competence framework for 2018–2024 (Toudic and Krause 2017). After an introduction that outlines the background to the study, our paper critically reviews a sample of novel corpus-based teaching methods, and reveals commonalities and differences as to the place and role of corpora in 21st century translation pedagogy. The paper concludes by offering some recommendations for future research and practice.

(Sara Laviosa is the author of Sects. 1 and 2 of this paper. Gaetano Falco is the author of Sects. 3 and 4).

S. Laviosa (✉) · G. Falco
Dipartimento LeLiA, Università Degli Studi Di Bari Aldo Moro, Via Garruba 6,
70122 Bari, Italy
e-mail: sara.laviosa@uniba.it

G. Falco
e-mail: gaetano.falco@uniba.it

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_1

Keywords Translator training · Translation pedagogy · Corpora · Pedagogical translation

1 Introduction: Expanding Holmes's Vision for Translation Studies

In James S. Holmes's outline of his vision for translation studies, the major area of research in the applied branch of the discipline is translator training, that is the teaching of translating "in schools and courses to train professional translators" (1988: 77). This area of scholarly enquiry addresses questions concerning teaching methods, testing techniques and curriculum planning. The coordinated noun phrase "translator training and education" is also used in the literature to denote the same area of research (cf. Kelly and Martin 2020; Washbourne 2020). In this chapter, we use the original compound term, "translator training". Holmes's configuration of applied translation studies presupposes that translator training and pedagogical translation "need to be carefully distinguished" (1988: 77). Also, it envisages that extensive and rigorous research aimed at assessing the effectiveness of translating as a technique in foreign-language teaching and a test of foreign-language acquisition be undertaken in a separate area, namely "translation policy". The findings of these envisaged studies would enable translation scholars to give informed advice on "what part translating should play in the teaching and learning of foreign languages" (1988: 78).

Since the turn of the century, ethnographic, experimental and survey-based studies carried out by translation scholars and educational linguists worldwide have produced ample empirical evidence for the effectiveness of L2 translating as a teaching method and a means of assessing language proficiency at all levels of instruction (Laviosa 2020; Laviosa and González-Davies 2020). This growing body of transdisciplinary and interdisciplinary research is inspired by the tenets upheld by the so-called "multilingual turn", an important paradigm shift that foregrounds "multilingualism, rather than monolingualism, as the new norm of applied linguistic and sociolinguistic analysis" (May 2014: 1). The multilingual turn is endorsed and promoted by the Council of Europe. Crucially, the CEFR Companion Volume with New Descriptors highlights the importance of mediating between individuals with no common language as one of the abilities that form part of plurilingual and pluricultural competence. These two aspects of plurilingualism are intimately interrelated and constitute the goal of modern languages education in the twenty-first century. Plurilingual individuals are able to call flexibly upon a single, interrelated, uneven and developing plurilinguistic repertoire that they combine with their general competences and various strategies to accomplish a host of communicative tasks involving intercultural interaction (Council of Europe 2020).

Mediation tasks, in particular, require that the user/learner is able to act as a social agent who creates bridges and helps to construct or convey meaning, sometimes within the same language, and sometimes from one language to another (cross-linguistic mediation). The context can be social, pedagogic, cultural, linguistic or professional. Mediation involves the integration of receptive, productive and frequent interactive abilities. There are different types of mediation tasks, each requiring specific integrated abilities that are carefully described in the CEFR. These are (a) mediating a text (within the same language and between languages), (b) mediating concepts and (c) mediating communication. Mediating a text between language A (the learner's best language) and language B (the learner's new language) includes the following oral and written tasks:

- relaying specific information given in a particular section of an unabridged text;
- explaining data presented in graphs, diagrams or charts;
- processing a text, e.g. summarizing it;
- translating a text.

At the higher levels of linguistic proficiency (C1 and C2) the abilities required to translate a written text in writing, which is the focus of the present discussion, are as follows:

C1 Can translate into (language B) abstract texts on social, academic and professional subjects in his/her field written in (language A), successfully conveying evaluative aspects and arguments, including many of the implications associated with them, though some expressions may be over-influenced by the original.

C2 Can translate into (language B) technical material outside his/her field of specialization written in (language A), provided subject matter accuracy is checked by a specialist in the field concerned—(Council of Europe 2020: 103).

According to the competence model presented in the CEFR, translating a written text at C1 and C2 levels involves processing the source message and articulating it in the target language. The key functional abilities required to transfer meaning from one language to another are (a) comprehensibility of the translation, (b) adherence to the relevant norms in the target language and (c) capturing nuances in the original. Therefore, the CEFR fully legitimizes translation in language learning and teaching as a cross-linguistic mediation activity that plurilingual individuals can carry out in a personal, social, academic or professional context. Furthermore, the CEFR reappraises translating not just as an exercise in contrastive grammar, a means of achieving communicative competence or a test of students' knowledge of the target language but, most importantly, as a valuable skill in its own right. A competent plurilingual individual develops this skill in degree programmes where one or more languages are taught up to C1 or C2 level. One can readily detect a significant shift from the traditional view of translation *for* language teaching toward the emerging view of translation *in* language teaching. The latter concept draws on the principles of four major educational philosophies: technological, social reformist, humanistic and academic. As Guy Cook contends (2010: 109–112), from a technological perspective, in today's increasingly multilingual and multicultural societies, translation is a much-needed skill for personal, educational, social and professional reasons. From a social reformist perspective, translation can promote liberal, humanist and democratic values, because it facilitates language and cultural encounters with an understanding of difference. From a humanistic educational perspective, students look upon translation as a form of bilingual instruction. From an academic perspective, translation fosters the study of linguistics.

Moreover, with regard to the widely held dichotomy between pedagogical and professional translation, the CEFR affirms:

"Translating a written text in writing" is by its very nature a more formal process than providing an impromptu oral translation. However, this CEFR descriptor scale is not intended to relate to the activities of professional translators or to their training. Indeed, translation competences are not addressed in the scale. Furthermore, professional translators, like professional interpreters, develop their competence through their career. [...] On the other hand, plurilingual users/learners [...] sometimes find themselves in a situation in which they are asked to provide a written translation of a text in their professional or personal context. Here they are being asked to reproduce the substantive message of the source text, rather than necessarily interpret the style and tone of the original into an appropriate style and tone in the translation, as a professional translator would be expected to do—(Council of Europe 2020: 102).

The distinction drawn by the CEFR between pedagogical and professional translating is subtle and lies, in our view, at the heart of the difference made in translation theory between translation conceived as transfer of meaning (consonant with the instrumental model) and translation viewed as an interpretive act (consonant with the hermeneutic model) (cf. Laviosa 2019; Venuti 2017). We argue that, in order to gain a proper understanding of the relationship between these two forms of mediated communication, i.e. educational translation on the one hand and professional translation on the other, we need to compare and contrast the competence model presented in the CEFR with the translation competences required of professional translators. To this end, it is useful to consider the model of professional translation competence presented in a document titled *European Master's in Translation Competence Framework 2017* (Toudic and Krause 2017). The EMT is a network of Master's level study programmes that was developed in 2009 by higher education institutions in partnership with the European Commission's Directorate General for Translation (DGT). The *EMT Competence Framework 2017* has been drawn out in response to three main developments that have occurred in the provision of translation services in the last decade. These developments are (a) the impact of technology, (b) the continuing expansion of English as a lingua franca, and (c) the role of artificial intelligence and social media in communication. The new framework builds on the "Wheel of Competence", which was designed in 2009 by the members of the EMT network (Gambier et al. 2009), and views translating as a process to meet individual, societal or institutional needs.

The aim of the *EMT Competence Framework 2017* is to consolidate and enhance the employability of graduates with Master's degrees in translation throughout Europe. It considers translation a multi-faceted profession and recommends that translator training at Master's degree level should equip students not only with a deep understanding of all the processes taking place when conveying meaning from one language to another but also with the ability to perform and provide translation service in line with the highest professional and ethical standards. The framework defines five complementary areas of competence, all equally important:

- LANGUAGE AND CULTURE (TRANSCULTURAL AND SOCIOLINGUISTIC AWARENESS AND COMMUNICATIVE SKILLS)
- TRANSLATION (STRATEGIC, METHODOLOGICAL AND THEMATIC COMPETENCE)
- TECHNOLOGY (TOOLS AND APPLICATIONS)
- PERSONAL AND INTERPERSONAL
- SERVICE PROVISION

We will now expound on each competence area, in turn, highlighting the skills that a graduate with a B.A. Hons. or a Master's degree in modern languages will be able to build on in translator training at the postgraduate level. The competence area named LANGUAGE AND CULTURE includes all the general and language-specific linguistic, socio-linguistic, cultural and transcultural knowledge and skills that constitute the basis of advanced translation competence. The framework recommends that language A (the main target language) should be mastered at CEFR level C2 or with native or bilingual proficiency. The other working languages should be mastered at CEFR level C1 and above. A graduate with a B.A. Hons. or a Master's degree in modern languages will possess the prerequisites for being admitted to an EMT programme since he/she will have an excellent command of language A and will have achieved level C1 or level C2 in at least one other working language (i.e. language B, the main source language).

TRANSLATION competence should be understood in the broadest sense, encompassing not only the actual meaning transfer between two languages but also all the strategic, methodological and thematic skills that come into play before, during and after the transfer phase per se, from document analysis to final quality control procedures in domain-specific, media-specific and situation-specific types of translation. The latter include public service translation, interpreting, localization and audio-visual translation. Translation competence includes also the ability to use machine translation, the automatic conversion of text from one natural language to another (cf. Kenny 2020). A graduate in modern languages would have gained an adequate general understanding of the meaning transfer phase between languages as one of the many processes taking place in professional translating. Therefore, this knowledge and the associated key functional abilities that he/she will have acquired in one or more target languages at C1 of C2 level (comprehensibility, accuracy and fluency of the written target text) will be a valuable asset in translator training.

The other competence areas are specific to translation teaching in Master's degree programmes aimed at students who wish to pursue a professional career in translation. TECHNOLOGY includes all the knowledge and skills used to implement present and future technologies during the different phases of the translation process (cf. Olohan 2020). It also includes the basic knowledge of machine

Table 1 Technological knowledge and skills

• Use the most relevant IT applications, including the full range of office software, and adapt rapidly to new tools and IT resources
• Make effective use of search engines, corpus-based tools, text analysis tools and CAT tools
• Pre-process, process and manage files and other media/sources as part of the translation, e.g. video and multimedia files, handle web technologies
• Master the basics of MT and its impact on the translation process
• Assess the relevance of MT systems in a translation workflow and implement the appropriate MT system where relevant
• Apply other tools in support of language and translation technology, such as workflow management software

(Toudic and Krause 2017: 9)

translation and the ability to utilize it when needed. As we can see in Table 1, the ability to use computerized corpora as translation aids is an integral part of the area of competence devoted to technological tools and applications. In the Wheel of Competence designed in 2009 by the members of the EMT network (Gambier et al. 2009), this particular skill was a component of the information mining competence, which included knowing how to use tools and search engines effectively (e.g. terminology software, electronic corpora and electronic dictionaries).

The PERSONAL AND INTERPERSONAL area of competence includes all the so-called "soft skills" that hone graduate adaptability and employability, namely planning and managing time, stress and workload; complying with deadlines, instructions and specifications; use of social media; self-evaluation and collaborative learning. Finally, SERVICE PROVISION covers all the skills relating to the provision of language services in a professional context, from client awareness and negotiation to project management and quality assurance.

If we compare the CEFR and the EMT models, we can identify three core areas of competence, knowledge and skills that would have been acquired in modern languages degree programmes and would be valuable assets when undergoing translator training at the postgraduate level. These areas are:

- a general understanding of one of the processes involved in professional translating, namely the meaning transfer phase between the source and the target language;
- plurilingual and pluricultural competence and integrated receptive and productive communication skills as prerequisites for developing linguistic, cultural and translation skills at Master's degree level;
- the ability to translate written texts on social, academic and professional subjects as an asset for honing the capacity to translate a broader range of texts in domain-specific, media-specific and situation-specific translation assignments.

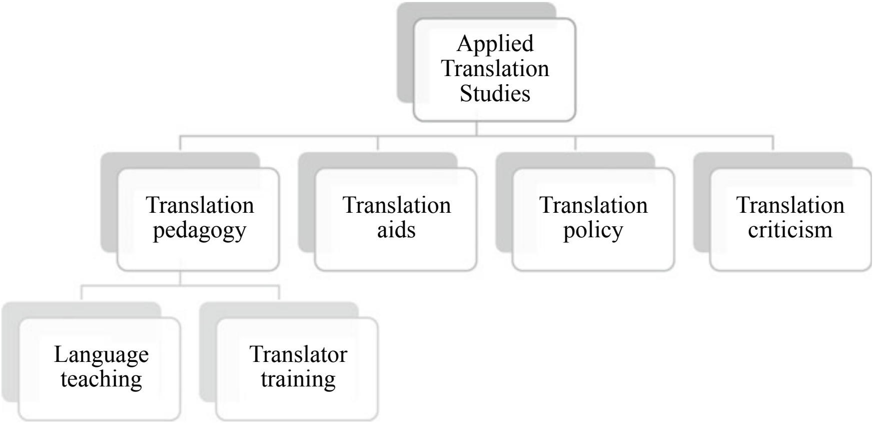


Fig. 1 Redrawing the outline of applied translation studies

On the basis of the comparative analysis presented here, translation pedagogy in higher education can be conceived as a continuum that starts with translation in language teaching and then progresses towards translator training. Going back to Holmes's delineation of applied translation studies, we propose that teaching translation as cross-linguistic mediation in undergraduate or postgraduate degree programmes in modern languages and translator training in Master's degree programmes would be considered offshoots of the pedagogic sub-branch of applied translation studies. We also propose that this research domain be re-named with the superordinate term "translation pedagogy" (see Fig. 1).

This new configuration of applied translation studies brings about far-reaching changes in the whole field of scholarship and beyond. Firstly, it fully recognizes the status of translation in language teaching as a research domain in its own right, rather than considering it merely a topic to be covered in translation policy research programmes. Secondly, the relocation of educational translation to its rightful place within the pedagogic sub-branch of the applications of translation studies widens the range of topics that can be explored in dedicated research projects aimed at investigating not only teaching approaches and methods but also testing techniques and curriculum design. Thirdly, in order to elaborate such research programmes, it is crucially important to engage in a constructive dialogue with relevant neighbouring fields, most notably second language acquisition studies (SLA), language-teaching methodology, languages for specific purposes (LSP), educational linguistics and philosophy of education.

It is within this new delineation of the pedagogic sub-branch of applied translation studies that we are going to examine the place and role of monolingual and bilingual corpora in translation pedagogy in Sects. 2 and 3 of our paper. But before we do this, it is worth reflecting on the impact that the relocation of the pedagogic translation may have on the research questions that translation policy will address in future. We envisage that the translation scholar working in this research area will continue to give advice on what needs to be translated in a given socio-cultural situation as well as on what the social and economic status of the translator is and should be. Examples of salient topics that may be investigated by the policy scholar are the translation production in crowdsourcing environments such as Wikipedia (see McDonough Dolmaia 2020); the role of translation in promoting and preserving languages of lesser diffusion such as Welsh, Corsican and Scots (see Baer 2020); the relationship between translators and their work environment such as international publishing (see Kershaw 2020); the increasing visibility of translators and their redefinition as creators and re-creators of their texts (see Summers 2020).

With regard to the other two sub-branches of applied translation studies, we envisage that, like translation policy and translation pedagogy, they will maintain their individuality and visibility by virtue of which they will continue to engage in an intradisciplinary dialectical relationship with the other domains of applied, theoretical and descriptive translation research as well as forge interdisciplinary relationships with adjacent fields of study. For example, translation aids will interface with disciplines as varied as lexicography, terminology, LSP studies, computational linguistics and artificial intelligence. Translation criticism, which extends beyond translation quality assessment, will interface with comparative literary studies, reception studies, cultural studies, stylistics, publishing studies and book history. From an intradisciplinary perspective, translation aids and criticism will influence teaching methods and testing in both language teaching and translator training, and translation policy will influence curriculum design, particularly in postgraduate translator training programmes.

2 Corpora in Language Teaching

In this section, we survey a small but representative sample of corpus-based pedagogic procedures that language and translation teachers, who are also practising translators, explain and illustrate in two textbooks aimed at university students majoring in English (Stewart 2018) and Spanish (Carreres et al. 2018). The book we are going to consider first is *Italian to English Translation with Sketch Engine: A Guide to the Translation of Tourist Texts* published in 2018 and authored by Dominic Stewart (University of Trento). The intended target readership is composed of students of English (C1 level) with Italian as language A. The activities consist of authentic translation tasks assigned by the author during the teaching of the module Lingua Inglese I that forms part of the curriculum design of the first-year postgraduate degree in *Mediazione linguistica, turismo e culture*. The module comprises 15 lessons of 90 min each (Dominic Stewart, personal communication via email, 5–7 May 2020).

After an introductory chapter that outlines the translation principles underpinning the teaching method and describes the recommended language resources, the book is organized into 15 teaching units, each containing:

- a short abridged text of about 250 words to be translated for an envisaged international, non-specialist target readership consisting of travellers requiring clear and accurate information on tourist sites in Italy;
- a proposed translation sentence by sentence, which is based on successful renderings submitted by students;
- a discussion on unsuitable equivalents or appropriate alternatives arising from renderings submitted by the students.

The translations were carried out with the aid of large, general target-language corpora together with online language resources. The target-language corpora are the British National Corpus (BNC), containing 100 million words of British English offering a broad range of text types, 90% of written texts and 10% spoken, and the web-derived corpus ukWaC, containing 2 billion words retrieved from websites in the .uk Internet domain, and searched through the corpus software Sketch Engine. The additional online language resources are monolingual English dictionaries, learner's monolingual English dictionaries, monolingual Italian dictionaries and bilingual Italian-English dictionaries.

By way of example, we now illustrate how students benefitted from searching the BNC and ukWaC to solve translation problems arising at different levels of cross-linguistic analysis. After examining all the 15 lessons illustrated in the textbook, we grouped the main lexical and grammatical mismatches that students encountered when translating Italian tourist texts into English into four main categories:

(I) Noun phrases containing toponyms.
(II) Subject-specific terminology.
(III) Polywords.
(IV) Language-specific collocations:

i different node words with semantically equivalent collocates
ii different collocates with semantically equivalent node words.

With regard to noun groups with place-names, students searched the equivalent superordinate words in the BNC and ukWaC (e.g. island, lake, lagoon, mount, pass, plateau, stream and valley), and were able to identify the correct grammatical structure and word order of the following noun phrases, thus producing cohesive, coherent, comprehensible, fluent and accurate target language texts:

- l'altopiano di Brentonico → *the Brentonico Plateau*
- l'altopiano di Malga Fanta → *the Malga Fanta Plateau*
- l'isola di Barbana → *the island of Barbana / the Isle of Barbana / Barbana Island*
- il lago Pra de Stua → *Lake Pra de Stua*
- la laguna di Grado → *the lagoon of Grado / Grado's Lagoon*
- il monte Baldo → *Mount Baldo*
- il passo di Fittanze della Sega → *the pass of Fittanze della Sega*
- il passo di Xomo → *the Xomo Pass*

- il torrente Brasa → *the Brasa stream*
- il torrente Caglieron → *the Caglieron stream*
- la valle dell'Adige/la vallata dell'Adige → *the Adige Valley / Adige Valley*

The frequent use of terms belonging to specialized fields of knowledge such as history, military history, geography, history of art, religion, architecture, gastronomy, transport, and arts and crafts is a feature of the language of tourism. As Maria Vittoria Calvi observes in connection with the discursive practices that characterize the description and promotion of tourist sites,

> Sul piano lessicale, si evidenzia l'uso frequente di unità terminologiche *mutuate* da altri settori correlati (storia dell'arte, geografia, gastronomia, ecc.) e solitamente non risemantizzate (Calvi 2012: 21, original emphasis).

By searching the BNC students were able to identify accurate and fluent equivalents of the following historical terms and expressions:

- il primo conflitto mondiale → *the First World War / World War I*
- l'ultima Guerra → *the Second World War / World War II*
- il dopoguerra → *the end of World War II*

Similarly, with the aid of the BNC and monolingual learner's dictionaries, students discovered several suitable equivalents for the geographical terms *salita* and *gobba*:

- siamo a 2/3 della salita detta della Polsa → *You are now two-thirds of the way up the ascent/climb/rise known as the Polsa/called Polsa*
- poco sotto la gobba del Cornetto, m 1543 → *below the hillock/hummock/hump/bump/mound/knoll Gobba del Cornetto, 1543 m*

Furthermore, a simple query search of the BNC and ukWaC revealed these renderings of the Italian name of the religious order founded by St. Francis of Assisi in 1209:

- *frati francescani minori → Franciscan friars / Franciscan Friars / Franciscan monks*

  Polywords are short lexical phrases that allow no variability and are continuous (cf. Nattinger and De Carrico 1992). By searching ukWaC, students identified the following equivalents in the order of preference based on the frequency of occurrence:
- secondo la tradizione, (non-restrictive appositive set off by a comma) → *tradition has it that* (261) / *according to tradition,* (non-restrictive appositive set off by a comma) (202) / *by tradition,* (non-restrictive appositive set off by a comma) (196) / *tradition holds that* (24)
- conosciuta in tutto il mondo → *recognised worldwide* (167) / *known worldwide* (161) / *famous worldwide* (54) / *recognized worldwide* (49) / *worldwide known* (14) / *worldwide famous* (8) / *worldwide recognised* (5) / *worldwide recognized* (2)
- a ricordo di → *in memory of* (6,121) / *as a memorial of* (78)

As translators and language and translation teachers know very well, collocation does not always travel across languages and cultures, hence one cannot “assume that semantic equivalents across languages have analogous collocational networks” (Stewart 2018: 11). Large general corpora in the target language can aid learners to investigate thoroughly this aspect of language use. We offer two examples of language-specific collocations examined in the textbook. The first regards the adjective *panoramico* and the English equivalent *panoramic*. While *panoramico* collocates with the node word *scorcio,* often in the plural form, as in *scorci panoramici*, the literal translation *panoramic glimpses* occurs only twice in ukWaC and does not occur in the BNC. Instead, the collocation *panoramic view(s)* is recorded in monolingual dictionaries and is very frequent in both corpora. Therefore, students realized that a comprehensible, accurate and fluent translation of the original collocation *scorci panoramici* is *panoramic views*.

The second example concerns the different collocates of the semantically equivalent node words *parete di roccia* and *rock face*. In the source text, *parete di roccia* occurs with the attributive adjective *impraticabile*, which, when referring to places, means “che non si può percorrere” (that cannot be crossed or run through/ across) (Vocabolario della Lingua Italiana di Nicola Zingarelli). However, the equivalent attributive adjective *impracticable* means “it is impossible to do in an effective way” (Cambridge Advanced Learner’s Dictionary) and collocates with abstract nouns such as *ideas*, *proposals* or *suggestions*. Indeed, by searching the BNC and ukWaC super-sensed corpus using the Concordance function of Sketch Engine at the time of writing this paper, the following frequencies are found:

	Impractical ideas	Impractical proposals	Impractical suggestions
BNC	0	2	0
ukWaC super-sensed	1	1	1

A word sketch of the adjective *arduous* (suggested by the teacher) revealed a set of node words belonging to the same semantic field of natural scenery that *rock face* belongs to, namely *path*, *climb*, *descent* and *terrain*. At the end of this careful search, where the teacher guided as a facilitator of the learning process, the students reached a consensus and rendered the original collocation *una impraticabile parete di roccia* with *an arduous rock face.* The following are some of the findings of the usage frequency of *arduous rock face* at the time of writing this paper:

	Arduous rock face
BNC	0
ukWaC super-sensed	0
English Web 2015	0
Google	1 (with a metaphorical meaning)

Therefore, the English rendition of *an arduous rock face* can be considered a good example of innovative collocation in English through analogy. Yet, this collocation is still very infrequent. It is also not quite accurate because *arduous* means "difficult, needing a lot of effort and energy" (Cambridge Advanced Learner's Dictionary). An accurate translation equivalent of the attributive adjective *impraticabile* is *impassable*, as recorded in English-Italian bilingual dictionaries. A word sketch of *impassable* in the BNC shows that it collocates with a variety of nouns belonging to the semantic field of natural scenery, e.g. *morass*, *landslide*, *rapids*, *dams*, *cliffs*, *country lanes*. In the English Web 2015 corpus, *impassable* collocates with nouns as varied as *terrain*, *ravine(s)*, *swamp(s)*, *thickets*, *crevasse (s)*, *gorge(s)*, *mountains*, *waterfall(s)*, *torrent(s)*, *canyon(s)*, *peaks*, *forest(s)* and *woods*. But *rock face* is not included in the long list of nouns modified by *impassable*. A Google search conducted at the time of writing of this paper reveals 51 occurrences of *an impassable rock face*, all referring to mountain climbing. The example of cross-linguistic mismatches at the level of collocation we have examined here highlights the importance of using multiple online learning resources when translating texts written in a subject-specific field such as nature-based tourism in alpine areas.

Summing up, when translating tourist texts with the aid of corpora, students worked individually and collaboratively in the language laboratory and engaged in group discussions guided by the teacher. They were able to solve a variety of problems arising from lexical and grammatical discrepancies between the source and the target language. In doing so, they became aware of the stylistic norms of the target language in the specific field of tourism, and, in most cases, they were capable of producing intelligible, accurate and fluent translations. They also acquired transferable technological skills that could be valuable assets if they wished to undertake translator training with a view to pursuing a professional career in translation. However, students were never encouraged to use Free Online Machine Translation (FOMT) engines such as Google Translate, despite empirical evidence showing the increasing use of these computer-assisted translation tools for various language learning tasks such as reading, writing and grammar assignments (see Enríquez Raído et al. 2020).

The second book we are going to overview is *Mundos en palabras: Learning Advanced Spanish through Translation* published in 2018 and authored by Ángeles Carreres and María Noriega-Sánchez (University of Cambridge, UK) and Carme Calduch (Queen Mary University of London). The intended target readership consists of advanced undergraduate students of Spanish (C1 level) with English as language A. The aim of the book is to develop cross-linguistic and cross-cultural awareness as well as foster the ability to translate a wide range of authentic texts from English to Spanish. The pedagogic approach adopted is task-based language learning and the activities are designed around two key tenets, i.e. translation is conceived as a form of mediated communication and learning as a collaboration among peers and between students and the teacher. This stance is in line with the approach adopted by the CEFR, where mediation "focuses on the role of language in processes like creating the space and conditions for communicating and/or learning, collaborating to construct new meaning, encouraging others to construct or understand new meaning, and passing on new information in an appropriate form" (Council of Europe 2020: 90).

The book is divided into 12 chapters. The first two chapters expound on the concept of translation underpinning the pedagogic approach adopted in the coursebook and introduce a number of key concepts, such as translation equivalence, translation strategy and translation competence, among others. Chapter 3 deals with the use of lexicographical and terminological resources and tools that students need when undertaking translation tasks either in class or by distance learning. The remainder of the coursebook presents authentic translation activities that focus on text types as varied as recipes, fiction, poetry, humour, theatre, advertising and audiovisual texts. The last chapter is devoted to the translation of language varieties such as Spanglish. The companion website contains (a) complementary exercises that require the support of online language learning resources, (b) additional activities, (c) downloadable learning materials and (d) suggested answers to most exercises. The latter is meant to be pointers for reflection and self-evaluation.

Corpora are introduced in a dedicated section of Chap. 3 entitled, "Los corpus lingüistícos". The authors first describe the main features and uses (language learning, acquisition of subject-specific terminology and professional translating) of the general corpora of the Real Academia Española. Then, they refer students to the activities contained in the companion website. The pedagogic objective is threefold, i.e. acquire practical knowledge of two corpora, in particular, Corpus del Español Actual (CREA) and Corpus de Diacrónico del Español (CORDE), develop the ability to use them autonomously when needed and reflect on their usefulness for language learning and translating. By way of example, one of the exercises in the companion website focuses on collocation and asks students to search the polysemic verb *echar* in a subcorpus of CREA that represents periodicals published in Colombia. After retrieving the first set of KWIC concordance lines, students look for the collocational patterns associated with three different meanings of the transitive verb *echar*, i.e. *deshacer algo* (defeat); *reprochar* (reproach); *culpar* (blame), and then copy in their worksheet the actual verbal context in which *echar* conveys each of the above meanings. After completing all the corpus-based activities provided in the companion website, students carry out the following reflection task:

Actividad 13.

> Tras haberte familiarizado con las búsquedas en corpus con las actividades de la Plataforma Digital, anota tres casos en los que crees que los corpus te pueden ayudar en tus traducciones y en tu aprendizaje del español (Carreres et al. 2018: 83).

As a concluding remark, we can say that Stewart's and Carreres et al.'s coursebooks make a valid contribution to fulfilling the long-term prediction made by Guy Cook at the end of his landmark work on educational translation:

> If the benefits of TILT [Translation in Language Teaching] were to be recognized in theory as well as practice by those in positions of power and influence as well as by rank-and-file teachers, it would have positive repercussions, and would initiate activity and innovation in

> many areas beyond classroom practice itself. New materials would need to be written, new tests designed, and new elements introduced into teacher education (Cook 2010: 156).

With regard to the use of corpora, Stewart's textbook, in particular, focuses on two of the three areas of convergence between teaching and language corpora earmarked by Geoffrey Leech (1997, quoted in McEnery et al. 2006: 97). These areas are "teaching to exploit" and "exploiting to teach". The former means providing students with technical expertise so that they can utilize corpora for their own learning purposes. The latter means using a corpus-based, data-driven learning approach to teaching language and linguistics courses. However, the third area, "teaching about", is beyond the scope of both books, since it refers to the teaching of corpus linguistics as an academic subject in curricula for linguistics and language-related degree programmes at undergraduate and postgraduate levels.

3 Corpora in Translator Training

In this section, we offer an overview of a representative sample of recent studies concerning the integration of corpora in translator training as a result of the spread of technological tools and expertise in the translation profession (Wong Shuk Man 2015). Indeed, as has been advocated by many scholars, there is a need for designing syllabi which, besides providing students with language and communication skills and tools, include modules intended to develop technological competence (EMT Annual Report 2019; Gouadec 2007; Samson 2005; Sikora 2014; Pym 2012; Torrés-Simón and Pym 2019).

Among them, Daniel Gouadec (2007) contends that a well-trained translator should possess appropriate technological knowledge and skills in addition to knowledge of terminology management systems for translation purposes, good documentation and research skills, as well as familiarity with technical and scientific writing. Technological knowledge and skills entail "familiarity with database management systems and electronic data management (XML/XSL/SML), proficiency at using translation memory systems, knowledge of proof-reading, revision and post-editing techniques, knowledge of technologies and software used in the processes of document production and management" (Gouadec 2007: 331–332).

It is worth pointing out that we use the terms "competence", "skills" and "knowledge" in accordance with the definitions that are provided in *The European Qualifications Framework for Lifelong Learning (EQF)*, and are also upheld in the *EMT Competence Framework 2017*. The definitions are as follows:

> competence "means the proven ability to use knowledge, skills and personal, social and/ or methodological abilities, in work or study situations and in professional and personal development. In the context of the European Qualifications Framework, competence is described in terms of responsibility and autonomy".
>
> skills "means the ability to apply knowledge and use know-how to complete tasks and solve problems. In the context of the European Qualifications Framework, skills are described as

> cognitive (involving the use of logical, intuitive and creative thinking) or practical (involving manual dexterity and the use of methods, materials, tools and instruments)".
>
> knowledge "means the outcome of the assimilation of information through learning. Knowledge is the body of facts, principles, theories and practices that is related to a field of work or study. In the context of the European Qualifications Framework, knowledge is described as theoretical and/or factual" (European Commission Education and Culture 2008).

Therefore, competence in translation technology is not just a matter of automatic work, it involves critical thinking, creativity and methodology. Disregarding this key aspect of translation technology implied widening the gap between theory and practice, which has always been one of the conundrums of translation studies. In this regard, Lynne Bowker warns against the "siloization" of technological tools as in many translator training programmes, where

> the tools are only seen and used in 'core' courses—i.e. courses with a specific focus on technology—rather than being integrated across a range of applied courses in the translator training program. The resulting gap between theory and practice does not provide students with an accurate picture of how they are likely to work—and in fact may be expected or required to work—in many professional contexts. To truly learn how tools fit into the translation process, technology-related tasks must be contextualized rather than severed from realistic experience (Bowker 2015: 97).

In fact, the word *technology* derives from two ancient Greek words, *τέχνη* and *λόγος* transliterated as *téchne* and *lógos*, respectively. *Téchne* means art, skill, craft, and especially the principles or methods employed in making something or attaining an objective. *Lógos* means speech, word and the utterance by which inward thought is expressed. So, literally, *technology* means words or discourse about the principles and methods used in making something or achieving a goal. Only lately, the word *technology* has come to mean something narrower than the original sense. In line with its etymology, we view technology holistically as a system, a process, forms of knowledge and new discoveries, as well as a set of tools that involve continuous advancement.[1] In the context of translator training, we regard "technological competence" as the knowledge of various tools, in particular electronic tools, e.g. word processors, computer-aided translation (CAT) tools, the Internet, terminological databases, corpora, as well as the skills needed to use these tools correctly and appropriately, together with the systematic use of these tools in project management and project workflow. Ideally, a successful translator training course should include all these elements. This chapter focuses on the role of corpora, including web-derived corpora and pre-constructed and do-it-yourself (DIY) corpora, available both offline and online, as well as other "associated processing tools such as concordancers [that] may find a place in a documentation course on a translator education program" because learning how to design and compile DIY corpora would enhance trainees' critical thinking, evaluation and decision-making skills (Bowker 2015: 91).

[1] https://web.engr.oregonstate.edu/~funkk/Technology/technology.html.

Moving on to examining recent pedagogical research, translator training generally adopts a socio-constructivist approach, which posits that learning how to translate results from the student's decision-making process. Research has shown that a learner-centred approach combined with the use of monolingual source-language and target-language corpora, as well as bi/multilingual comparable and parallel corpora can have beneficial effects. Furthermore, over the last years, translator training has been boosted by the incorporation of modules on the use of technological tools. CAT tools, machine translation, translation memories, collaborative translation platforms have been introduced in many Master's degree programmes on translation studies across the world.

Consequently, there has been a considerable growth in the publication of handbooks, papers, guidelines for academic curriculum design and reports on translator training experiences in the classroom, not to speak of conferences, seminars and roundtables. At the same time, researchers, scholars, professionals, international institutions and stakeholders from the professional world of translation and interpreting have promoted projects (OPTIMALE, EMT, MUST, MELLANGE, PACTE) and meetings in order to pool together their own experiences. What emerges from this heterogeneous "universe" is the shared view that some changes are needed in terms of translator training approaches; in particular, the need for a shift towards learner-generated training, where students are not simply consumers but the protagonists of their own learning experience.

With regard to corpus-based pedagogic approaches, Cécile Frérot (2016) offers a comprehensive review of research into the usefulness of corpora and corpus tools. In particular, following Alison Beeby et al. (2009), Frérot distinguishes between two teaching styles that can be subsumed under a general socio-constructivist orientation. These styles are "corpus use for learning to translate" and "learning corpus use to translate". In the former, teachers design corpus-based translation-related tasks so that students focus on a particular translation issue and analyse a given set of preselected data. In the latter, students play a central role as they are involved in designing and compiling DIY corpora, as well as identifying strategies and tools to search the corpora by themselves. In so doing, students learn how to use corpora efficiently and strategically to solve real-life translation problems. Frérot also highlights the added value of corpora for enhancing the quality of students' specialized translation, especially with regard to terminology, collocational patterns, genre and discourse.

Research into the use of corpora, including different types of software for compiling and analysing pre-constructed and/or DIY corpora, has grown steadily in recent years in a wide array of academic fields as well as in translator training, especially in Master's degree programmes in specialized translation. This growth has been boosted by the increased availability of free online corpora, concordancers, search engines and other platforms, which enable trainees to compile their own corpora or search ready-made ones. Furthermore, parallel, multilingual websites, such as the European Commission's system of multilingual display, allow users, including translator trainees, to manage a large collection of parallel electronic texts on various subject-specific domains covered by the European

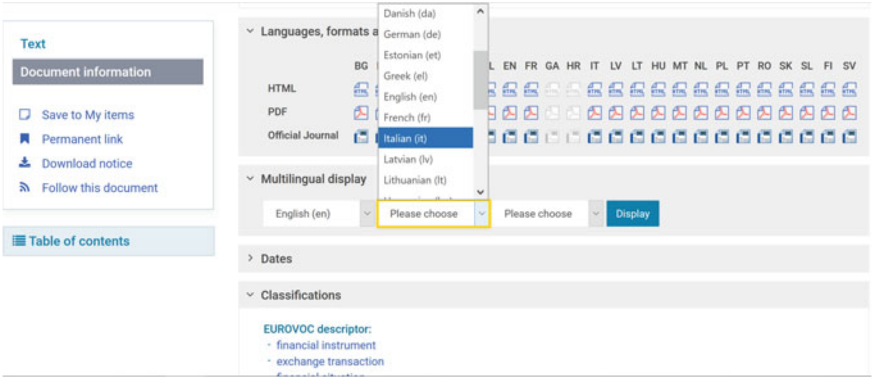
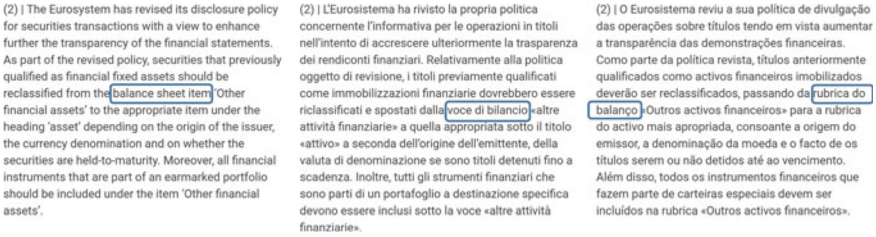


Fig. 2 European Commission Multilingual Display



(2) | The Eurosystem has revised its disclosure policy for securities transactions with a view to enhance further the transparency of the financial statements. As part of the revised policy, securities that previously qualified as financial fixed assets should be reclassified from the balance sheet item 'Other financial assets' to the appropriate item under the heading 'asset' depending on the origin of the issuer, the currency denomination and on whether the securities are held-to-maturity. Moreover, all financial instruments that are part of an earmarked portfolio should be included under the item 'Other financial assets'.

(2) | L'Eurosistema ha rivisto la propria politica concernente l'informativa per le operazioni in titoli nell'intento di accrescere ulteriormente la trasparenza dei rendiconti finanziari. Relativamente alla politica oggetto di revisione, i titoli previamente qualificati come immobilizzazioni finanziarie dovrebbero essere riclassificati e spostati dalla voce di bilancio «altre attività finanziarie» a quella appropriata sotto il titolo «attivo» a seconda dell'origine dell'emittente, della valuta di denominazione se sono titoli detenuti fino a scadenza. Inoltre, tutti gli strumenti finanziari che sono parti di un portafoglio a destinazione specifica devono essere inclusi sotto la voce «altre attività finanziarie».

(2) | O Eurosistema reviu a sua política de divulgação das operações sobre títulos tendo em vista aumentar a transparência das demonstrações financeiras. Como parte da política revista, títulos anteriormente qualificados como activos financeiros imobilizados deverão ser reclassificados, passando da rubrica do balanço «Outros activos financeiros» para a rubrica do activo mais apropriada, consoante a origem do emissor, a denominação da moeda e o facto de os títulos serem ou não detidos até ao vencimento. Além disso, todos os instrumentos financeiros que fazem parte de carteiras especiais devem ser incluídos na rubrica «Outros activos financeiros».

Fig. 3 *Balance sheet item* and its translations into Italian (*voce di bilancio*) and Portuguese (*rubrica do balanço*)

Commission Directorate-General for Translation, e.g. economics, commerce, customs laws, health, employment, IT, tourism, immigration and many others. The platform displays queries in the source language and one or two target languages out of the 24 EU official languages (see Figs. 2 and 3).

The European Commission Directorate-General for Translation has developed various tools, such as terminological databases (IATE, EuroVoc) and machine translation engines (eTranslation), to support their in-house translators as well as trainees who attend translation internships at the Directorate-Generals (DGs) of the European Commission. EU documents are an important pedagogical aid whenever students need to be trained to design and compile DIY parallel corpora for language and translation learning. Moreover, over the last two decades, there has been increasing interest in designing web-derived corpora which are domain-specific and very large in size. Currently, we have corpora consisting of billions of words, e.g. iWeb (Intelligent Web-based Corpora), or NOW (News On the Web), which are representative of different varieties of the English language.[2] Another source of

[2] Both corpora have been designed, compiled and are continuously updated by Mike Davies.

“corpus colossal”,[3] in terms of size, language variety and subject differentiation is Sketch Engine, which collects a number of corpora in different natural languages, including corpora of parallel texts, such as the European Commission DGT, Eur-lex, Europarl, OPUS, the Bible and the Quran corpora. Notably, some of them are also specialized. Therefore, they are an important source of translation data and a valuable resource for creating bilingual glossaries of specialized terms.

Sketch Engine, in particular, incorporates the BootCat toolkit,

> a suite of perl programs implementing an iterative procedure to bootstrap specialized corpora and terms from the web, requiring only a small list of “seeds” (terms that are expected to be typical of the domain of interest) as input. The basic idea is very simple: Build a corpus by automatically searching Google for a small set of seed terms; extract new (single-word) terms from this corpus; use the latter to build a new corpus via a new set of automated Google queries; extract new terms/seeds from this corpus and so forth. The final corpus and unigram term list are then used to build a list of multi-word terms. These are sequences of words that must satisfy a set of constraints on their structure, frequency and distribution (Baroni and Bernardini 2004).

Actually, the operation described here can be regarded as an interesting case of the principles of Frame Semantics at work, namely, the continuous expansion of the semantic frames starting from a very limited number of seed terms and the concepts around them by repeatedly retrieving the related data on Google to make the original conceptual frames grow as the body of the encyclopaedic knowledge expands. By “encyclopaedic knowledge” we mean the whole of linguistic and non-linguistic knowledge to which a word or a group of words potentially provide access. Unlike “dictionary knowledge”, which is merely concerned with word meaning, i.e. with words “as neatly packaged bundles of meaning”, “encyclopaedic knowledge” involves knowledge from a pragmatic perspective, i.e. knowledge of word use in the specific context of a conceptual domain (Evans and Green 2006: 208).

The iterative procedure illustrated by Baroni and Bernardini (2004) soon found other applications in the field of translation practice, education and training. Sara Castagnoli (2006), for example, trained her students attending the School for Interpreters and Translators of the University of Bologna, Forlì, Italy, to use the BootCat toolkit in order to generate corpora automatically and autonomously. One of the two modules taught in her course introduced students to corpus annotation, POS tagging and collocation extraction in order that they could “consider terminological work both as an autonomous discipline and as a component of the translation process” (Castagnoli 2006: 162). Castagnoli’s is just one of many pedagogic practices that demonstrate how corpora can find important applications in the context of translator training.

Similarly, working with a group of second-year students attending the second-year MA programme in specialized translation at the Cologne University of Applied Sciences (Institute of Translation and Multilingual Communication), Ralph

[3] The expression was coined by *The Economist* (January 20, 2005).

Krüger (2012) designed an introductory course on the key aspects of real-life translation projects. One major task included the compilation of DIY corpora using the Internet. Students learnt to use tools such as WebCorp Live, and apply strategies for querying the Internet itself as a macro-corpus. Significantly,

> The students' feedback on the use of corpora was largely positive. They particularly appreciated the availability of a high-quality translation corpus which provided immediate solutions to various translation problems. The parallel-text corpus was, for the most part, not used as an independent resource. The students mainly used it as a "back-up" corpus to check whether the terminology and structural patterns found in the target texts of the translation corpus were also present in original target-language texts (Krüger 2012: 522).

Moreover, Ana Frankenberg-Garcia (2015) reports on a training experience with a multilingual group of 13 students attending an MA in Translation at the University of Surrey during the academic year 2013/14. One module in her course focused on the hands-on use of corpora for translation practice. The syllabus was not meant to provide students with theoretical insights into corpus linguistics and translation studies; it rather focused on the practical use of corpora for translation purposes. Using WebBootCat, included in Sketch Engine, students crawled the web in order to compile DIY specialized corpora that they could use for translations assigned both in the classroom and for homework (Frankenberg-Garcia 2015: 357–358).

In her paper, Clara Inés López Rodríguez (2016) presents the results of a three-year research project carried out within the framework of CombiMed: combinatory lexis in Medicine: cognition, text and context (FFI2014-51,899-R), funded by the Spanish Ministry of Economy and Competitiveness, and the teaching innovation action Tradusaluda: audiovisual resources for the promotion of health in Europe: accessible subtitling and translation (PID 14–39), funded by the University of Granada. In particular, the article describes how quality corpora were employed in a course of scientific and technical translation from English to Spanish, with a special focus on terminological variation as evidence of language creativity. Creativity is seen as an important aspect in the cognitive processes involved in translation since it promotes the coinage of neologisms as well as the attribution of new meanings to existing words, the metaphorization of general nouns and the re-organization of syntax. These phenomena were investigated in the context of technical and scientific translation. More specifically, the students taking part in the project were trained to compile DIY monolingual corpora in English and Spanish as well as parallel corpora with the help of various online platforms, e.g. Sketch Engine, WebCorp,[4] Aranea Project No Sketch Engine,[5] Exemplar Words in context,[6] BNC,[7] Corpus of Global Web-Based English (GloWbE),[8] CREA (Corpus de

[4] http://www.webcorp.org.uk.

[5] http://ucts.uniba.sk.

[6] http://www.springerexemplar.com/.

[7] http://www.natcorp.ox.ac.uk/

[8] https://www.english-corpora.org/glowbe/.

Referencia del Español Actual), amongst others. Again, this experience underscores the central role of learners in the translation training process.

In a similar vein, Anne Lise Laursen and Ismael Arinas Pellón (2012) present the results of a concurrent course in specialized translation between Spanish and Danish. In particular, using two sets of parallel texts, i.e. the EU 4th and 7th Directive (i.e. EU Financial Reporting Legislation), and the International Accounting Standards (IAS), available both in the Spanish and Danish versions, students were trained to retrieve terminological equivalents in the accounting field in Spanish and Danish, using the AntConc concordance. As a result, trainees acquired technological skills as well as linguistic and thematic competence, i.e. the ability to identify stylistic, genre-related, terminological features in the two languages. Finally, Hind Alotaibi's (2017) study provides evidence of the attention that Arabic countries are giving to corpus-driven translation training. Alotaibi reports on the findings of a project carried out with a group of students at the College of Languages and Translation, King Sand University. The students were involved in the compilation of a 10-million word Arabic-English parallel corpus, consisting of texts from different domains, including specialized ones, such as medicine, law and science.

4 Concluding Remarks

Within the outline of applied translation studies that we have redrawn from a plurilingual perspective on language education, we can reasonably predict that translating with the aid of corpora will play a key role in stimulating the creation of novel multilingual learning resources and materials as well as the design of new teaching procedures and testing techniques in language learning and translator training, given the growing impact of technology on present-day electronically mediated communication, the study of languages and the language industry at large (see Crystal 2018: 452–476). It is fair to say that the use of corpora in translation-oriented language education is still in its infancy. There is a long way to go before corpora are systematically integrated into language teaching at the university level. In order to achieve this goal, we need to engage in empirical research aimed at assessing the benefits of corpus-aided translation for language learning. And to that end, we need to promote closer cooperation between educational linguists and translation studies scholars (see Laviosa and González-Davies 2020).

In sharp contrast, the use of corpora in translator training is growing rapidly. Recent innovative experimental research undertaken in South China Normal University, which shows the distinct advantages of translating with the aid of the parallel corpus of the Hong Kong Parallel *cum* Comparable Corpus (Liu 2020), is highly promising and offers an excellent model for future longitudinal studies that may be carried out with other language combinations and in other educational settings. The widespread scholarly interest in corpus-assisted translation teaching reflects the significant changes that are taking place in education, in general, and in translator training, in particular. The incorporation of new Information Technologies in education has reshaped the contents from the old media to the new digital ones, a phenomenon known as "remediation", a term coined by Bolter and Grusin (1996). Remediation entails a form of information literacy, consisting of various competences, such as "computer literacy, critical thinking and information, skills, Information Technology (IT) literacy, learning how to learn (or lifelong learning) literacies, and library or digital media literacies" (Loucky 2008: 281–282). This technological turn in training reveals a shift from a static way of teaching, largely based on a transmissionist, teacher-centred approach, to a more dynamic, proactive way of learning, which involves a learner-generated approach, in that students "become responsible for their own learning and the learning of others. The teacher is no more the authority who determines what is studied and assesses the quality of students' work" (Atan 2012: 2).

Technology also favours collaborative work, information exchange, exploratory attitudes and inquiry-based learning: these activities take place in an authentic, real-world context (Sessom 2008). Crucially, these are the same tenets underpinning Donald Kiraly's socio-constructivist approach to translator training, whereby "individuals have no choice but to create or construct meanings and knowledge through participation in the interpersonal, intersubjective interaction" (Kiraly 2000: 4). Hence, students are active builders of their knowledge, they monitor and are responsible for their education process. Also, they do not act in isolation but are part of a community in which each individual is involved in a collaborative process. This pedagogic approach entails a change in the power relations between teachers and students. Students are empowered to become decisive actors in designing and planning translation activities and syllabi as well as corpus use.

The learners' design and construction of corpora, especially of specialized corpora for translation purposes, contribute to creating and enhancing their "encyclopaedic knowledge" (Evans and Green 2006). This term refers to the meaningful knowledge of specialized domains and students' awareness not only of terms, as isolated units, but also of the textual, social, cultural and pragmatic context in which these terms are used. Even though many studies of corpora in translator training are not explicitly based on cognitive linguistics, we cannot disregard the cognitive shift that is occurring in corpus-based and corpus-driven translation teaching. A case in point is Elina Symseridou's method of collecting corpora from the web through Sketch Engine with the aim to train students in healthcare translation. As she observes, "the adoption of a corpus-based teaching methodology allows for the inclusion of more specialised texts in the curriculum, even if the teacher is not acquainted with a discipline, as well as the creation of a collaborative learning environment" (Symseridou 2018: 73). Finally, and looking to the future, recent research indicates that corpus-based approaches to translator training can be further improved by incorporating other methodologies. As shown by Gaetano Falco (2014), the integration of concept maps into corpus-driven teaching methods can contribute to enhancing the trainees' cognitive processes, boosting their creativity and awareness of specialized domains, thus enabling them to acquire encyclopaedic knowledge and, accordingly, perform translation tasks successfully.

Acknowledgements The research presented in this chapter has been conducted as part of a project titled, "La svolta multilingue nella linguistica educazionale: teoria e pedagogie innovative", coordinated by Sara Laviosa and funded by the Università degli Studi di Bari Aldo Moro from 2020 to 2022. We wish to thank the editors of this volume and two anonymous referees for their valuable feedback on the topics addressed. A special word of thanks goes to Ángeles Carreres and Dominic Stewart for giving us useful background information about the educational contexts in which their textbooks are currently used. We are also grateful to Richard D.G. Braithwaite for proofreading the final version of our contribution before submitting it to the editors.

References

Alotaibi, H.M. 2017. Arabic-English parallel corpus: A new resource for translation training and language teaching. *Arab World English Journal (AWEJ)* 8 (3): 319–337.

Atan, S. 2012. Towards a collaborative learning environment through ICT: A case study. In *Proceedings of the 5th conference edition ICT for language learning*. https://www.365strangers.it/liavilfolea1975/hfjyrdkiapg-conference-proceedings-ict-for-language-129434/.

Baer, B.J. 2020. Nations and nation-building. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 361–365. London: Routledge.

Baroni, M., and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/509.pdf.

Beeby, A., P. Rodríguez-Inés, and P. Sánchez-Gijón. 2009. Introduction. In *Corpus use and translating*, ed. A. Beeby, P. Rodríguez Inés and P. Sánchez-Gijón, 1–8. Amsterdam: John Benjamins.

Bolter, J.D., and R. Grusin. 1996. Remediation. *Configurations* 4 (3): 311–358.

Bowker, L. 2015. Computer-aided translation: Translator training. In *The Routledge Encyclopedia of translation technology*, ed. S.W. Chan, 88–104. Oxon/New York: Routledge.

Calvi, M.V. 2012. Forme participative nel discorso turistico in lingua spagnola! In *Comunicare la città: Turismo culturale e comunicazione. Il caso di Brescia*, ed. M. Agorni, 19–30. Milano: FrancoAngeli.

Carreres, Á., M. Noriega-Sánchez, and C. Calduch. 2018. *Mundos en palabras: Learning advanced Spanish through translation*. London: Routledge.

Castagnoli, S. 2006. Using the WeClara Inés b as a source of LSP corpora in the terminology classroom. In *WaCky! working papers on the web as corpus*, ed. M. Baroni and S. Bernardini, 159–172. Bologna: GEDIT.

Cook, G. 2010. *Translation in language teaching: An argument for reassessment*. Oxford: Oxford University Press.

Council of Europe. 2020. *Common European framework of reference for languages: Learning, teaching, assessment. companion volume with new descriptors*. Strasbourg: Council of Europe Publishing. Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) (coe.int).

Crystal, D. 2018. *The Cambridge Encyclopedia of the English language*. Cambridge: Cambridge University Press.

Enríquez Raído, V., F. Austermühl, and M. Sánchez Torrón. 2020. Computer-assisted L2 learning and translation. In *The Routledge handbook of translation and education*, ed. S. Laviosa and M. González-Davies, 278–299. London: Routledge.

European Commission Education and Culture. 2008. *The European Qualifications Framework for Lifelong Learning (EQF).* http://relaunch.ecompetences.eu/wp-content/uploads/2013/11/EQF_broch_2008_en.pdf

Evans, V., and M.C. Green. 2006. *Cognitive linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Falco, G. 2014. *Metodi e strumenti per l'analisi linguistica dei testi economici. Dalla SFG al Web 2.0*. Bari: Edizioni dal Sud.

Frankenberg-Garcia, A. 2015. Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university. *Corpora* 10 (3): 351–380.

Frérot, C. 2016. Corpora and corpus technology for translation purposes in professional and academic environments: Major achievements and new Perspectives. In *adernos de Tradução Edição Especial—Corpus use and learning to translate, almost 20 years on*, ed. D. Gallego-Hernández and P. Rodríguez-Inés, 36 (1):36–61.

Gambier, Y., N. Gormezano, D. Gouadec, D. Kelly, H. Lee-Jahnke, N. Kocinjacic-Pokorn, and E. Tabakowska. 2009. *Competences for Professional Translators, Experts in Multilingual and Multimedia Communication*. https://ec.europa.eu/info/sites/info/files/emt_competences_translators_en.pdf

Gouadec, D. 2007. *Translation as a profession*. Amsterdam/Philadelphia: John Benjamins.

Holmes, J.S. 1988. *Translated! papers on literary translation and translation studies*. Amsterdam: Rodopi.

Liu, K. 2020. *Corpus-assisted translation teaching: Issues and challenges*. Singapore: Springer.

Kelly, D., and A. Martin. 2020. Training and education, curriculum. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 591–596. London: Routledge.

Kenny, D. 2020. Machine translation. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 305–310. London: Routledge.

Kershaw, A. 2020. Publishing landscapes. In *The Routledge Encyclopedia of Translation Studies*, ed. M. Baker and G. Saldanha, 445–449. London: Routledge.

Kiraly, D. 2000. *A social constructivist approach to translator education: Empowerment from theory practice*. Manchester: St. Jerome.

Krüger, R. 2012. Working with corpora in the translation classroom. *Studies in Second Language Learning and Teaching* 2 (4): 505–525. https://pressto.amu.edu.pl/index.php/ssllt/index

Laursen, A.L. and I. Arinas Pellón. 2012. Text corpora in translator training: A case study of the use of comparable corpora in classroom teaching. *The Interpreter and Translator Trainer* 6 (1): 1–26.

Laviosa, S. 2019. The instrumental and hermeneutic models of translation in higher education. In *Übersetzung: Über die Möglichkeit, Pädagogik anders zu denken*, ed. N. Engel and S. Köngeter, 39–55. Wiesbaden: Springer VS.

Laviosa, S. 2020. Language teaching. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 271–275. London: Routledge.

Laviosa, S., and M. González-Davies, eds. 2020. *Translation and education*. London: Routledge.

López Rodríguez, C.I. 2016. Using corpora in scientific and technical translation training: Resources to identify conventionality and promote creativity. In *Cadernos de Tradução. Edição Especial—Corpus use and learning to translate, almost 20 years on*, ed. D. Gallego-Hernández and P. Rodríguez-Inés, 36(1):88–120.

Loucky, J.P. 2008. Improving online readability and information literacy. In *Handbook of research on digital information technologies: Innovations, methods, and ethical issues*, ed. T. Hansson, 281–302. Hershey/New York: Information Science Reference.

May, S. 2014. Introducing the “multilingual turn.” In *The multilingual turn: Implications for SLA, TESOL and bilingual education*, ed. S. May, 1–6. London: Routledge.

McDonough Dolmaia, J. 2020. Crowdsourced translation. In *Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 124–129. London: Routledge.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-Based language studies: An advanced resource book*. London: Routledge.

Nattinger, J.R., and J. De Carrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Olohan, M. 2020. Technology, translation. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 574–579. London: Routledge.

Pym, A. 2012. Training translators. In *The Oxford Handbook of translation studies*, ed. K. Malmkjær and K. Windle, 475–489. Oxford: Oxford University Press.

Samson, R. 2005. Computer-assisted translation. In *Training for the new millennium*, ed. M. Tennet, 101–126. Amsterdam: John Benjamins.

Sessoms, D. 2008. Interactive instruction: Creating interactive learning environments through tomorrow's teachers. *International Journal of Technology in Teaching and Learning* 4 (2): 86–96.

Sikora, I. 2014. The need for CAT training within translator training programmes: Modern bare necessities or unnecessary fancies of translation trainers? *inTRAlinea.* Special Issue: Challenges in Translation Pedagogy. http://www.intralinea.org/specials/article/the_need_for_cat_training_within_translator_training_programmes

Stewart, D. 2018. *Italian to English translation with sketch engine: A guide to the translation of tourist texts*. Trento: Tangram Edizioni Scientifiche.

Summers, C. 2020. Authorship. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 35–39. London: Routledge.

Symseridou, E. 2018. The web as a corpus and for building corpora in the teaching of specialised translation: The example of texts in healthcare. *Fitispos International Journal* 5 (1): 60–82.

Torrés-Simon, E. and A. Pym. 2019. European Masters in translation: A comparative study. In *The evolving curriculum in interpreter and translator education stakeholder perspectives and voices*, ed. D.B. Sawyer, F. Austermühl, and V. Enríquez Raído, 75–97. Amsterdam: John Benjamins.

Toudic, D., and A. Krause. 2017. (on behalf of the EMT Board) *European Master's in Translation Competence Framework 2017*. https://ec.europa.eu/info/sites/info/files/emt_competence_fwk_2017_en_web.pdf

Venuti, L. 2017. Introduction: Translation, interpretation, and the humanities. In *Teaching translation: Programs, courses, pedagogies*, ed. L. Venuti, 1–14. London: Routledge.

Washbourne, K. 2020. Training and education, theory and practice. In *The Routledge Encyclopedia of translation studies*, ed. M. Baker and G. Saldanha, 597–602. London: Routledge.

Wong Shuk Man, C. 2015. The teaching of machine translation: The Chinese University of Hong Kong as a case study. In *The Routledge Encyclopedia of translation technology*, ed. S.W. Chan, 237–252. Oxon/New York: Routledge.

Sara Laviosa holds a BA Hons in Psychology (Open University), an MA (with Distinction) in TESOL (University of Birmingham) and a PhD in Translation Studies (University of Manchester). She is an Associate Professor in English Language and Translation Studies in the Dipartimento di Lettere, Lingue, Arti. Italianistica e culture comparate, Università degli Studi di Bari 'Aldo Moro'. Her research interests include corpus linguistics, translation studies and TESOL. She has published extensively in international journals and collected volumes and is the author of three monographs*Corpus-Based Translation Studies* (Rodopi/Brill, 2002); *Translation and Language Education* (Routledge, 2014); and *Linking Wor(l)ds* (with a digital workbook, *English Lexis, Grammar and Translation*, authored by Richard D.G. Braithwaite) (Liguori Editore, 2020). She is co-author (with Adriana Pagano, Hannu Kempannen and Meng Ji) of *Textual and Contextual Analysis in Empirical Translation Studies* (Springer, 2017). She is the founder and General Editor of the journal *Translation and Translanguaging in Multilingual Contexts* (John Benjamins). Dr. Laviosa's recent publications include *Corpus Translation Studies (CTS)* (with Meng Ji) (Pensa MultiMedia, 2019), *The Routledge Handbook of Translation and Education* (co-edited with Maria González Davies, 2020), *The Oxford Handbook of Translation and Social Practices* (co-edited with Meng Ji, 2020)*Studi empirici della traduzione basati sui corpora* (with Meng Ji) (Pensa MultiMedia, 2020), *CTS Spring-Cleaning: A Critical Reflection*, Special Issue of *MonTI* (co-edited with *María Calzada Perez, 2021), and Recent Trends in Corpus-based Translation Studies, Special Issue of Translation Quarterly (co-edited with Yanklong Liu, 2021).*

Gaetano Falco holds a PhD in Specialized Translation—Theory and Practice (Università degli Studi di Bari 'Aldo Moro'). He is a Researcher in English Language and Translation Studies at the Università degli Studi di Bari 'Aldo Moro'. His main research interests include translation studies, translation teaching, translation of LSPs, critical discourse analysis, pragmatics, cognitive linguistics and corpus linguistics. He serves as a Book Reviewer of the journal *Translation and Translanguaging in Multilingual Contexts* (John Benjamins). He is the author of the monograph, *Metodi e strumenti per l'analisi linguistica dei testi economici. Dalla SFG al Web 2.0* (Edizioni dal Sud, 2014). He is also the author of several journal articles and book chapters on translation teaching, translation of economic discourse in professional and non-professional genres (e.g. academic journals, comic books, movies), and CDA-based studies on corporate discourse. He is a member of the Associazione Italiana di Anglistica (AIA) and the European Society for Translation Studies (EST).

A Corpus-Based Examination of the Translation of the Suffix –ism into Chinese

Lily Lim

Abstract The suffix –ism is one of the most productive suffixes in English, which presents complexities and challenges in English-to-Chinese translation. This study examines corpus tools and resources that translators can have access to, aiming at revealing the potential values of corpus resources that translators can draw on to better deal with the suffix in their practice. Our investigation is focused on interrogating a parallel corpus named “Education”, which contains 450,000 pairs of English and Chinese sentences. Compared with traditional tools such as monolingual and English-Chinese (reverse) dictionaries, the parallel corpus exhibits remarkable values. It is easy to access at Sketch Engine, enabling systematic queries on decent coverage of –ism words. The English-Chinese concordance lines present a wealth of resources with which translators can tease out the senses conveyed by the –ism words, retrieve rich translation candidates in meaningful contexts that previous translators have worked with. The technical skills required appeared to be very attainable to translators with average IT competence. The merits of corpus resources are discussed in the context of translator training.

Keywords Suffix –ism · Translation equivalent · Lexicology · Parallel corpus

1 Introduction

This study on –ism words was first intrigued by the present investigator’s experience of teaching translators/interpreters at the tertiary level. When translating a passage entitled “Sukarno: A Fallen Hero’s Legacy” that contains sentence (1) into Chinese (Loveard 1994), almost every student rendered ‘nationalism’ into 民族主义 *mínzú zhŭyì* ‘national/ethical group’s ideology/ism’, a noun phrase in which 民族 *mínzú* ‘nation, ethnical group’ modifies 主义 *zhŭyì* ‘ideology or ism’. The suffix

L. Lim (✉)
School of Languages and Translation, Macao Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macau
e-mail: llim@ipm.edu.mo

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_2

–ism is therefore rendered into 主义 *zhǔyì*, an easy-to-find corresponding item in Chinese. However, this is a rather simplistic solution.

(1) Across the 5,000 km breadth of Indonesia, one name is synonymous with Indonesian **nationalism**: Sukarno, nation-builder and architect of independence (boldfaced fonts added for emphasis).

More experienced translators would consider other expressions that better express the meaning of "nationalism" in (1) because the collocation 印度尼西亚民族主义 *yìnní mínzú zhǔyì* 'Indonesian nationalism' sounds neutral or even moderately negative to Chinese ears, suggesting an ideology that entertains a narrower perspective that prioritises national interests. In fact, in Cheng's (郑宝璇 2004) textbook titled *Translation for Media*, a sample translation is provided for the passage from which (1) was extracted, in which "nationalism" is translated into 民族精神 *mínzú jīngshén* 'the spirit or essence of the nation' (ibid: 69), rather than 民族主义 *mínzú zhǔyì*. This translation sounds positive and even heroic in Chinese, in line with the sentiment of awe and admiration on Sukarno expressed in the original text.

As illustrated by example (1), more experienced translators tend to have more varied ways to render –ism words into Chinese. There is therefore a real need for student translators to expand their lexical repertoire so that they can convey the denotations and connotations entailed by the –ism words in context more aptly and effectively. In this study, we examine some of the most accessible tools and resources that translators can use and trail through the discovery processes to reveal the potential of the resources in solving translation problems surrounding –ism words. We aim at casting light on the following research questions:

(a) To what extent do traditional tools—e.g., English-Chinese dictionaries and reverse dictionaries—enable translators to understand the senses of –ism words and to obtain useful translation candidates in Chinese for rendering the words?
(b) Compared with the traditional tools, what advantages do monolingual and parallel corpora exhibit in translators' understanding of and dealing with –ism words?
(c) What level of technical skills—e.g., advanced, sophisticated, or basic—are required so that translators can carry out self-directed corpus-based studies to attain reasonable outcomes in terms of understanding the –ism words and retrieving translation candidates?

2 A Brief Literature Survey

Research over the last three decades has attested the value of using monolingual, parallel and comparative corpora in lexical studies and lexicography (e.g., Fillmore 1992; Kovář et al. 2016; Kubicka 2019; Lefever and Hoste 2014; Li 2017; Li et al.

2020b; Li and Wang 李龙兴 & 王宪 2021; Sinclair 2001; Teubert 2001; Wang and Chu-Ren 2017; Wang 2018; Wang 2021 in this volume) and also in translation practice and translation studies (e.g., Baker 1999; Chen et al. 2020; Johansson 2007; Li et al. 2020a; Mauranen 2004; Wang et al. 2020; Xiao 2010; Zanettin 2014: Chaps. 2 and 5). Recent studies on translator and interpreter training underscore the importance of fostering (trainee) translators' competence in information and communication technology (ICT) (e.g., Doval and Nieto 2019: 3–4; Laviosa and Falco 2021 in this volume; Wang and Lim 2017). A stream of the studies stresses on the notion of ICT *literacy*, which is argued to be an integrated part in the translator/interpreter's training curriculum (e.g., Laviosa & Falco 2021: Sect. 3; Lim 2020: 152), referring to the competence such as working with computer-assisted translation (CAT) tools, corpus-based tools and resources, text analysis tools, and translation memory systems. It is widely acknowledged that translators should not only consult conventional resources such as dictionaries and glossaries but also make effective use of corpus-based tools and resources in their practice.

In addition, corpora have been used in research on the translation of affixations across languages (e.g., Defrancq and Rawoens 2016; Lefer 2012; Lefer and Grabar 2015; Quah 1999). Although we were unable to find systematic research on the translation of the suffix –ism across languages, Lim's (2019) corpus-based investigation on "terror*ism*" casts light in this respect. Lim (ibid) found that "terrorism" in English and 恐怖主义 *kǒngbù zhǔyì* in Chinese differ from each other at least in two ways. First, "terrorism" is markedly more frequently used than 恐怖主义 based on corpus evidence. Second, 恐怖主义 tends to be complemented by an NP (noun phrase) much more than "terrorism" does. More precisely, 恐怖主义 occurs in complex NPs—e.g., 恐怖主义活动 *kǒngbù zhǔyì huódòng* 'terrorist act/activity' much more frequently than "terrorism" does, since "terrorism" often entails the meaning of terrorist activity in context. This points to the observation that "terrorism" spans a wider range of semantic meanings than 恐怖主义 does and therefore tends to stand alone, while, by contrast, 恐怖主义 needs to rely on an additional word such as a following NP to express more specific meanings. This further leads to the working hypothesis that the suffix –ism tends to convey wider semantic meanings than 主义 does although it needs to be studied and verified with (corpus) evidence. Given the fact that English-Chinese bilingual corpora have been steadily developed and some are easily accessible via the Internet, we can now examine some major resources in this chapter that translators can use to better deal with –ism words.

3 Conventional Tools for Understanding –ism and Identifying Its Chinese Translations

In this section, we examine a range of conventional tools and resources that translators may access for reaching a better understanding of the senses of –ism words and for retrieving translation candidates for –ism words. We examine the usefulness of monolingual and bilingual dictionaries and also monolingual corpora of both English and Chinese. From the results of this section, we will move on to explore in Sect. 4 the special contributions that a large parallel corpus can make in the hands of inquisitive translators with regard to their needs.

3.1 The Senses of –ism in Monolingual Dictionary

In terms of dictionaries, we examine a commonly used monolingual dictionary of English in this section and will explore an English-to-Chinese reverse dictionary and also large online dictionaries with English-Chinese sentence pairs as bilingual examples in the next Sect. (3.2). Of all the conventional resources, English dictionaries arguably present the most essential and traditional tools for translators to look up the definition and explanations of the suffix –ism regarding its main senses. For example, the Merriam Webster Dictionary (https://www.merriam-webster.com/dictionary/-ism) provides the definition of the noun suffix –ism in four sense groups and gives examples for each sense:

Senses	Examples	Chinese annotations[1]
1a**:** act**:** practice**:** process	// critic*ism* // plagiar*ism*	做法 *zuòfǎ* ‘practice’, 行为 *xíngwéi* ‘act’
b**:** manner of action or behavior characteristic of a (specified) person or thing	// animal*ism*	做的方式 *zuò de fāngshì* ‘manner of doing / acting’
c**:** prejudice or discrimination on the basis of a (specified) attribute	// rac*ism* // sex*ism*	歧视行为 *qíshì xíngwéi* ‘discriminatory act’
2a**:** state**:** condition**:** property	// barbarian*ism*	状态 *zhuàngtài* ‘state’, 情况 *qíngkuàng* ‘situation’, 属性 *shǔxìng* ‘property’
b**:** abnormal state or condition resulting from excess of a (specified) thing or marked by resemblance to (such) a person or thing	// alcohol*ism* // giant*ism*	病症 *bìngzhèng* ‘disease, syndrome’ …型 *xíng* ‘… shape, appearance’
3a**:** doctrine**:** theory**:** religion	// Buddh*ism*	学说 *xuéshuō* ‘doctrine’, 理论lǐlùn ‘theory’, 信仰 xìnyǎng ‘religion, faith’
b**:** adherence to a system or a class of principles	// stoic*ism*	保持系统/原则 *bǎochí xìtǒng/ yuánzé* ‘adhere to (a) system/ principle/s’
4: characteristic or peculiar feature or trait	// colloquial*ism*	语言特征等 *yǔyán tèzhēng děng* ‘language features, etc.’

Note [1]Chinese annotations (with pinyin and gloss in English) are added by the investigator for distinguishing sense groups and suggesting potential translation candidates

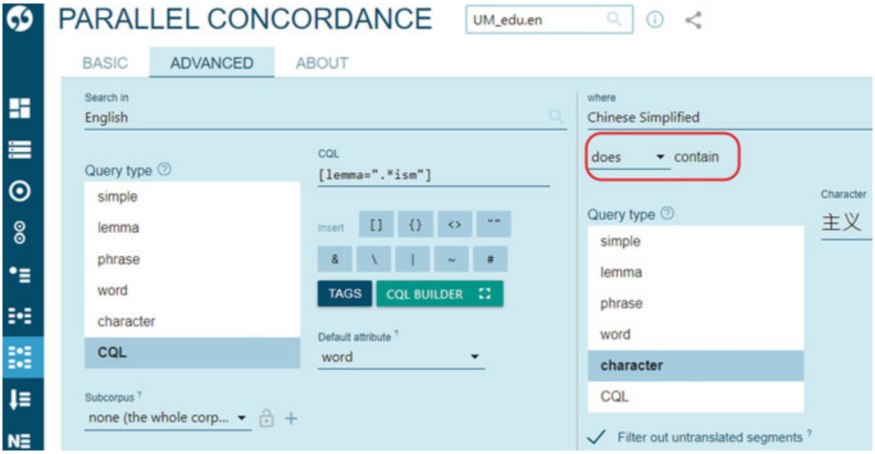
Of the four sense groups, only the sense group 3 (on doctrine and principles) is closely related to 主义 in Chinese, while the other three sense groups (1, 2 and 4) are rarely so, e.g., “criticism” of sense 1a, “colloquialism” of sense 4. There are exceptional cases under the latter sense groups though, which mainly have to do with paraphrasing translation, e.g., “sexism”, a word of sense 1c, may be paraphrased into 大男子主义 *dà nánzǐ zhǔyì* ‘male chauvinism’ in certain contexts, rather than the more commonly used translation 性别歧视 *xìngbié qíshì* ‘sex(-based) discrimination’. The –ism words of sense group 3—i.e., on a doctrine, theory, or religion and on the adherence to a system or a set of principles—tend to be rendered into 主义 in Chinese. But still, not all the words under the sense group 3 are translatable into 主义 in Chinese, e.g., Buddh*ism* is rendered as 佛教 rather than 佛主义. From the observations above, it is clear that not all –ism words are translatable into 主义 in Chinese, and, more precisely, we can (and translators in general should) reach the hypothesis that the semantic range of –ism is broader than that of the corresponding suffix 主义 in Chinese. The hypothesis needs to be tested with more bilingual language evidence, while the observations clearly point to the need for identifying translation candidates in Chinese for –ism apart from 主义.

3.2 Translation Candidates for –ism Words in English-Chinese Dictionaries

Reverse dictionaries facilitate the access of –ism words, given the fact that –ism is a suffix and all the –ism words appear in a continuous sequence in reverse dictionaries. We consulted *A Reverse English-Chinese Dictionary* by Sun (孙梅 1993), in which –ism words appear in eight consecutive pages, from “Lamaism” (ibid: 361) to “Nazism” (ibid: 368). Most of the ism words are annotated by only one or two Chinese terms, which explain the –ism word and may also serve as translation equivalents, while a few words are provided with three (e.g., “anarchism”, “nationalism”) or even four (e.g., “criticism”) Chinese correspondents. The number of the Chinese correspondent terms for each –ism word is therefore not large, which is in fact far lower than the repertoire retrievable in the parallel corpus we will investigate (see Sect. 4). However, from the Chinese expressions corresponding to the –ism words over the eight pages in Sun (ibid), observant translators can still find a considerable range of Chinese expressions ending with recurring characters such as 教 *jiào* ‘religion, sect’, 会 *huì* ‘school, association, society’, 学 *xué* ‘study, school’, 法 *fǎ* ‘method’, 论 *lùn* ‘discourse, doctrine’, 者 *zhě* ‘people, scholar’, 风气 *fēngqì* ‘trend’, 运动 *yùndòng* ‘movement’, 行为 *xíngwéi* ‘behaviour’, and 精神 *jīngshén* ‘spirit’, 性 *xìng* ‘character’, apart from 主义. These expressions are informative to the translators about the diversified translation candidates for –ism words.

In addition to reverse dictionaries and monolingual dictionaries, large English-Chinese dictionaries are undoubtedly one of the most used bilingual resources by translators. Lu Gusun's (陆谷孙 2007) 英汉大词典 (*The English-Chinese Dictionary*) is one of the most reliable English-Chinese dictionaries, which is often considered as a must-have tool by professional translators. It exhibits very conscientious work in terms of the coverage of the senses of each entry, using Chinese to finely and precisely explain the senses, which are further illustrated by examples in English with Chinese translations. Translators can look up all the information for potential translation candidates they can employ, or at least for inspiration to sort out useful translations. However, translators cannot expect that they can find the precise items immediately useful for the text they are translating, since the dictionary is more about the English language in terms of the senses of the entries (cf. Kubicka 2019: 84), rather than being designed for providing translation solutions. In other words, it is not a dictionary for English-Chinese translators.

Moreover, online English-Chinese dictionaries tend to be large, updated with free access, which come with bilingual examples of phrases and sentences that often exceed printed dictionaries in number. However, the online dictionaries do not allow translators to search all the –ism words together as they can do with reverse dictionaries, and translators can only search one –ism word at a time. Large online English-Chinese dictionaries provide examples of English-Chinese sentence pairs. However, the number and variety of the examples may still not be large enough for translators to harvest a decent range of translation candidates, although they may be good enough for language learners in general. For example, in YouDao dictionary, one of the largest online dictionaries developed in China (http://dict.youdao.com/), we retrieved only 27 English-Chinese bilingual sentence pairs for "nationalism". Of the 27 tokens of Chinese terms that correspond to "nationalism", only three different types emerged. We therefore consulted large online dictionaries with bilingual sentence pairs (e.g., YouDao dictionary) and also online bilingual corpus portals—e.g., CCL Chinese-English bilingual corpus of Peking University (http://ccl.pku.edu.cn:8080/ccl_corpus/index_bi.jsp), and BCC bilingual corpus of the BLCU (Beijing Language and Culture University) Corpus Center (although it is not accessible from 2021) (http://bcc.blcu.edu.cn/lang/bi)—to search the most frequently occurring –ism words (cf. Sect. 3.3) one at a time. We were able to retrieve various Chinese expressions other than 主义—e.g., 精神 *jingshen* 'spirit', 行为 *xingwei* 'behaviour', 制度 *zhidu* 'system', 运动 *yundong* 'movement', 性 *xing* 'character' – to render –ism words such as "professionalism", "feminism", "vandalism", and "terrorism". However, searching one –ism word at a time in these resources is laborious, and the translation equivalents identified tend to be low in both tokens and types compared to those contained in the parallel corpus "Education" (see Sect. 4), which also enables all –ism words to be retrieved by one search in the Sketch Engine platform (cf. Fig. 1).



PARALLEL CONCORDANCE UM_edu.en

Search in: English

Query type: CQL

CQL: [lemma=".*ism"]

Default attribute: word

Subcorpus: none (the whole corp...

where: Chinese Simplified

does contain

Query type: character

Character: 主义

✓ Filter out untranslated segments

Fig. 1 The “containing” query that searches both the –ism words in English and the characters 主义 in Chinese

3.3 *Monolingual Corpora of English and Chinese: Comparing –ism and -主义 Words*

Accessing monolingual English corpora enables translators to gain basic information about the frequency of occurrence of the –ism words taken together and also to identify the most frequently used ones. We used the British National Corpus (BNC) for the purpose, in view of its coverage of a variety of genres of both written and spoken texts. The results indicate that the suffix –ism is highly productive in English. The tokens of –ism words amount to 56,588 in total in BNC, which translates into a rate of 503.7 per million tokens (PMT). This frequency of occurrence is closest to that of the lemmas of “company” (57,118 tokens) and “course” (56,776), which are ranked the 21st and 22nd most commonly used nouns in BNC. Therefore, statistically, translators have to deal with various –ism words in their day-to-day practices. The top 30 most frequently used –ism words are listed in Table 1, according to the Wordlist resulting from the search of lemma ending with “ism” in BNC, a preloaded corpus accessible at Sketch Engine (SkE). In view of the much varied senses entailed by the large array of –ism words (cf. Sect. 3.1), translators should work more efficiently and effectively by gaining a better understanding of –ism words in terms of their senses and potential translation candidates.

To compare with the results of –ism words in BNC, a Chinese monolingual corpus to query主义expressions was accessed. We used the Chinese GigaWord 2 Corpus (Mainland, simplified characters), created in 2005, preloaded to SkE and tagged with the Chinese GigaWord tagset which covers newswire texts. Our Character search of主义 under the Concordance tab returned 146,816 tokens, at a rate of 587.0 per million tokens (PMT). The 30 most frequently used Chinese terms

Table 1 BNC Wordlist: lemma ending with "ism"

Rank	Lemma	Freq	PMT	Rank	Lemma	Freq	PMT
1	criticism	5806	51.68	16	scepticism	637	5.67
2	mechanism	4921	43.80	17	journalism	594	5.29
3	**capitalism**	1884	16.77	18	liberalism	504	4.49
4	organism	1793	15.96	19	professionalism	497	4.42
5	**socialism**	1644	14.63	20	metabolism	493	4.39
6	tourism	1464	13.03	21	conservatism	477	4.25
7	**racism**	1093	9.73	22	baptism	476	4.24
8	**nationalism**	1024	9.11	23	**imperialism**	420	3.74
9	**communism**	949	8.45	24	unionism	393	3.50
10	**realism**	888	7.90	25	modernism	390	3.47
11	optimism	840	7.48	26	pluralism	384	3.42
12	**Marxism**	732	6.52	27	**individualism**	380	3.38
13	feminism	693ara>	6.17	28	symbolism	378	3.36
14	**terrorism**	690	6.14	29	vandalism	338	3.01
15	fascism	654	5.82	30	Catholicism	328	2.92

Abbreviation PMT = per million tokens

are presented in Table 2, based on both the keyword in context (KWIC) list and the list of the first word on the left of 主义. The -主义 expressions in Chinese occur at a relatively higher frequency than that (503.7 PMT) of –ism words in BNC. However, we should note that the term 社会主义 *shèhuì zhǔyì* 'socialism' accounts for 54% of all the tokens of 主义 expressions in the Chinese GigaWord 2 Corpus (cf. Table 2), which reflects the tendency in lexical choices in mainland China. There are some other expressions in this Chinese corpus—e.g., 马克思主义 *mǎkèsī zhǔyì* 'Marxism' (ranked 4), 共产主义 *gòngchǎn zhǔyì* 'communism' (ranked 6), and 资本主义 *zīběn zhǔyì* 'capitalism' (ranked 9)—that occur at markedly higher frequencies than their English correspondents do in BNC. We paid particular attention to the –ism words in Table 1 that have translation equivalents in Table 2, and identified ten and highlighted them with bold-faced fonts in Tables 1 and 2. The ten –ism words tend to be rendered into -主义 expressions in Chinese, according to the intuition of proficient Chinese-English bilingual speakers. By contrast, some other words—e.g., "criticism", 'tourism", "metabolism"—are the most improbable words to be translated into主义 expressions in Chinese. However, this intuition needs to be tested with the evidence from parallel corpora (cf. Sect. 4.2).

Table 2 The most frequently used -主义 expressions in the Chinese GigaWord 2 Corpus

Rank	Chinese terms	*Pinyi*	Gloss in English	Freq	PMT
1	社会主义	*shèhuì zhǔyì*	**socialism**	78,629	314.36
2	爱国主义	*àiguó zhǔyì*	patriotism	9524	38.08
3	恐怖主义	*kǒngbù zhǔyì*	**terrorism**	8749	34.98
4	马克思主义	*mǎkèsī zhǔyì*	**Marxism**	8032	32.11
5	人道主义	*réndào zhǔyì*	humanitarian	6432	25.72
6	共产主义	*gòngchǎn zhǔyì*	**communism**	3040	12.15
7	保护主义	*bǎohù zhǔyì*	protectionism	2455	9.82
8	霸权主义	*bàquán zhǔyì*	hegemonism	2241	8.96
9	资本主义	*zīběn zhǔyì*	**capitalism**	2056	8.22
10	帝国主义	*dìguó zhǔyì*	**imperialism**	1904	7.61
11	民主主义	*mínzhǔ zhǔyì*	democracy	1802	7.20
12	马列主义	*mǎliè zhǔyì*	Marxism-Leninism	1792	7.16
13	种族主义	*zhǒngzú zhǔyì*	**racism**	1635	6.54
14	形式主义	*xíngshì zhǔyì*	formalism	1454	5.81
15	军国主义	*jūnguó zhǔyì*	militarism	1355	5.42
16	集体主义	*jítǐ zhǔyì*	collectivism	1053	4.21
17	殖民主义	*zhímín zhǔyì*	colonialism	1002	4.01
18	官僚主义	*guānliáo zhǔyì*	bureaucracy	946	3.78
19	分裂主义	*fēnliè zhǔyì*	separatism	898	3.59
20	马克思列宁主义	*mǎkèsī lièníng zhǔyì*	Marxism-Leninism	870	3.48
21	唯物主义	*wéiwù zhǔyì*	materialism	865	3.46
22	极端主义	*jíduān zhǔyì*	extremism	809	3.23
23	民族主义	*mínzú zhǔyì*	**nationalism**	637	2.55
24	拜金主义	*bàijīn zhǔyì*	money-worship-ism	545	2.18
25	英雄主义	*yīngxióng zhǔyì*	heroism	506	2.02
26	分离主义	*fēnlí zhǔyì*	separatism	382	1.53
27	个人主义	*gèrén zhǔyì*	**individualism**	296	1.18
28	享乐主义	*xiǎnglè zhǔyì*	hedonism	295	1.18
29	现实主义	*xiànshí zhǔyì*	**realism**	285	1.14
30	国际主义	*guójì zhǔyì*	internationalism	253	1.01

Abbreviation PMT = per million tokens

4 Querying a Parallel Corpus: The UM-Corpus at SkE

We selected an English-Chinese parallel corpus named “Education” and uploaded it to SkE for our queries and analysis. The “Education” corpus is one of the largest components of the UM-Corpus, a multi-domain and balanced parallel corpus constructed at the University of Macau (UM) (see Tian et al. 2014). The “Education” corpus consists of texts “acquired from online teaching materials, such as language teaching resources, and dictionaries, which can be served as language education" (ibid: 1840). We selected this corpus owing to its reasonable language and translation quality, decent size, and relevance to language. The "Education" corpus is downloaded from http://nlp2ct.cis.umac.mo/um-corpus/index.html, the website in which a 2.2 million sentence-pair version of the UM corpus is released to the community for research purposes. In addition, the "Education" corpus is one of the largest subcorpora of the released version, which consists of 450,000 pairs of sentences in English and Chinese with 8,401,095 tokens in English and 13,749,570 tokens in Chinese (ibid: 1840, Table 4). The direction of translation is not specified in the released version though. Therefore, the translation candidates for –ism words we intended to identify in the "Education" corpus cannot be specified in terms of the direction of translation. That is, the words and expressions identified in Chinese can either be the translations of –ism words from English into Chinese—i.e., "translation equivalents" in Mikhailov's (2021) terminology in this volume—or the lexical items in Chinese that are rendered into –ism words in English—i.e., "translation stimuli" (ibid).

Most of the sentence pairs in the "Education" corpus sound like English-to-Chinese rather than Chinese-to-English translation, as evidenced by their subject matter, style and wording. The investigator googled various sentence pairs on the web and confirmed that in most of the pairs the English sentences were the original, taken from books, articles, Wikipedia entries, U.S. presidents' speeches, and transcripts of lectures given at Yale University. A small portion of the sentence pairs is Chinese-to-English translation, e.g., the bilingual texts extracted from the Selected Works of Mao Tse-Tung, e.g., his speeches delivered on 3 May 1937 and on 7 April 1944. Therefore, the translation candidates identified in the "Education" corpus for rendering –ism words into Chinese predominantly consist of the instances of "translation equivalents" and also a small number of "translation stimuli", according to Mikhailov's terminology (ibid). In this section, the "translation" of –ism words is taken in a broad sense.

In the following sub-sections, we explore the values of the "Education" corpus in terms of the coverage of –ism words, the richness of translation candidates available, and the potential for translators to devise queries in order to answer their questions or tease out some patterns or regularities on the translation of –ism words. We also consider the technical competence and skills needed in the translators to conduct inquisitive learning with the corpus.

4.1 The Coverage of –ism Words

We retrieved all the instances of –ism words in the "Education" corpus under the Concordance tab using the following query devised in corpus query language (CQL):

[lemma = ".*ism"],

which returned 6,724 tokens of –ism words (i.e., 716 PMT), excluding an erroneous instance that was manually identified, in which the sentence boundary is mistaken by the system. Based on the Frequency List of the keyword in context (KWIC) for the concordance lines, we manually detected the –ism words with capitalised initials and merged them to lower-case ones. This leads to a total of 615 types of the lemmas of –ism words in English.

Compared with BNC (96,134,547 words in total), which returns 56,694 tokens and 1,543 types of –ism words by the same CQL query indicated above, the English subcorpora (8,168,479 words) of the "Education" corpus exhibits a considerable coverage of the commonly-used –ism words in both types (n = 615) and tokens (n = 6,724). The Chinese sentences aligned with the English ones constitute the wealth of resources from which translation candidates can be harvested.

4.2 Differentiating Three Types of –ism Words

Once the "Education" corpus was uploaded to SkE as a Multilingual Corpus, it was split into two corpora—an English one (8,168,479 words) and a Chinese one (8,020,107 words) with simplified Chinese characters—while the two are interlinked at the sentence level. The two work together as an English-Chinese parallel corpus that allows simultaneous searches of both corpora. The combined searches are crucial for translators to query the –ism words and the –主义 expressions in the English-Chinese paired sentences. The searches lead to quantitative results, which are far more informative than those provided by conventional resources (cf. Sect. 3).

The queries lead to an immediate finding that there are far more instances of –ism words that are *not* translated into words containing the Chinese characters 主义 *zhŭyì* 'doctrine, dogma' than those that are rendered into –主义 words. The investigator reached the finding from the quantitative results returned by the "containing" query and the "not-containing" query (detailed below), while translators can easily devise similar queries with the built-in query features at SkE. The "containing" query simultaneously searches the –ism words in the English corpus and the characters 主义 in the Chinese corpus (see Fig. 1), while the two corpora are aligned at the sentence level. It therefore retrieves the sentence pairs (n = 1,833) in which an –ism word occurs in the English sentence, while a -主义 word occurs in the aligned Chinese sentence. This involves "noises" in which the -主义 word in Chinese is the translation of a word other than the –ism word in the English sentence, for example, an –ist word (e.g., "socialist") rather than an –ism word in the sentence is translated into 主义. The parallel concordance lines were manually checked through and a low number of instances of the noise were found. Exclusion of the noises will further reduce the number of 1,833 sentence pairs returned by the "containing" query, while the main finding of the trends is valid and even strengthened. We use the term "*zhuyi*-words" to refer to those –ism words in English that are either exclusively or predominantly translated into 主义 in Chinese, e.g., "socialism", "capitalism", and "Marxism".

The “not-containing” query returned a much larger number of sentence pairs (n = 4,892) than the “containing” query (n = 1,833) did. The former query can be considered a simple modification of the latter in which the search of the Chinese characters 主义 is switched from “does contain” (like in Fig. 1) to “does *not* contain”, gathering all the instances of the –ism words with no 主义 in the corresponding Chinese sentences. The result indicates that there are at least 2.67 times as many –ism words that do not correspond to 主义 in Chinese as those that do in our data. We term the former type of –ism words “non-*zhuyi*” words to indicate their tendency of not being rendered into any word of 主义. Given the predominance of the “non-*zhuyi*” words over the “*zhuyi*” ones, translators need to give particular attention to the former group. Apart from the typical “*zhuyi*” and “non-*zhuyi*” words, there are –ism words that yield notable results in both the “containing” and “not-containing” queries, which we termed “partly *zhuyi*-words”, e.g., “terrorism”, “realism”, and “activism”.

By tabulating the frequencies of occurrence of each –ism word in both the “containing” and the “not-containing” queries, its tendency emerges regarding whether or not it tends to be rendered into (or, at least, corresponds to) –主义 words in Chinese. The –ism words in English can therefore be classified into three types—i.e., the “*zhuyi*”, “non-*zhuyi*”, and “partly-*zhuyi*” words. Classifying the –ism words into the three types entails practical value to translators, who can thereby deal with the three differently.

The first type—i.e., the *zhuyi*-words—is the easiest to render into Chinese, and competent translators should be able to produce a list of these words rather easily with SkE. Once the parallel concordance is generated by the “containing query” (cf. Fig. 1), one can yield the list of *zhuyi*-words using the Frequency of keyword in context (KWIC) feature for the English corpus (on the left in the parallel-concordance display), and, similarly, producing the list of –主义 words in the Chinese corpus (on the right). The English and Chinese lists match each other closely in the most frequently occurring items, e.g., “socialism” in English to 社会主义 *shèhuì zhǔyì* ‘socialism’ in Chinese, “capitalism” to 资本主义 *zīběn zhǔyì* ‘capitalism’, “terrorism” to 恐怖主义 *kǒngbù zhǔyì* ‘terrorism’, “imperialism” to 帝国主义 *dìguó zhǔyì I* ‘imperialism’, “Marxism” 马克思主义 to *mǎkèsī zhǔyì* ‘Marxism’, and “anarchism” to 无政府主义 *wú zhèngfǔ zhǔyì* ‘anarchism’. Once the “noises” that contain mismatches between –ism and –主义 words were cleaned, we found that these words are almost exclusively rendered into –主义 words in Chinese. Translators can handle them straightaway without spending their time considering alternative translations.

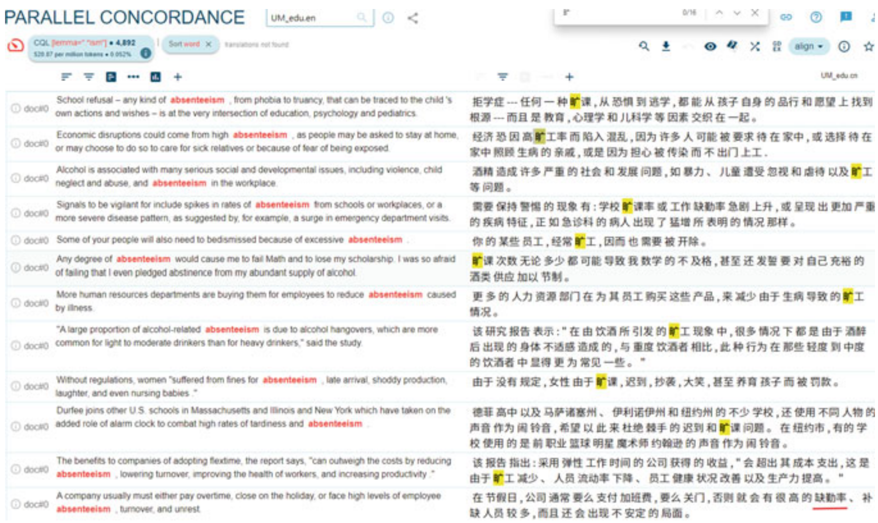
Similarly, we can retrieve the list of non-*zhuyi* words using the concordance lines in English generated by the “not-containing” query. Top on the list by descending frequency order are the words including “mechanism”, “criticism”, “organism”, “tourism”, “metabolism”, “Buddhism”, “autism”, and “baptism”. These words are rarely rendered into –主义 in our data. This appears to be useful information to translators, although native speakers of Chinese should intuitively know this. Most of the words have nothing to do with a doctrine, a system of beliefs or an ideology, and therefore are understandably non-*zhuyi* words. In addition, some words—e.g., “Taoism” and “Confucianism”—that definitely pertain to some belief systems or doctrines are ranked high on the non-*zhuyi* list as well. To sort out the ways and regularities for rendering the –ism words, parallel corpus presents primary resources that translators can turn to for information and inspiration.

In terms of the third type—the partly *zhuyi*-words—their meaning spans from the sense of *zhuyi* (doctrine, dogma, ideology) to other senses, such as the practice of a certain kind or an institution or a system established under the guidance of a certain dogma or ideology. Their most frequently occurring correspondent items in Chinese serve as important indicators of the range of the senses expressed by these –ism words.

4.3 Translation Candidates in Chinese: The Repertoire and Retrieval Methods

Since *zhuyi*-words are the easiest to translate into Chinese (cf. Sect. 4.2), we focus on the non-*zhuyi* and partly-*zhuyi* words and their potential translations in this section. We found the parallel corpus particularly useful for providing the translation candidates of individual –ism words as well as for differentiating major types of –ism words that tend to be rendered into some commonly used Chinese expressions other than 主义 (cf. Sect. 4.4).

The parallel concordance lines retrieved by the “not-containing” query display the English and Chinese sentences in juxtaposition, which present a quick reference for diversified translations of each –ism word. The English sentences can be alphabetically sorted by the –ism words they contain, which are the KWIC in the English concordance lines on the left, giving a display of all the sentences that contain each “non-*zhuyi*” –ism word (see Fig. 2), beginning with “absenteeism”, “academism”, and “activism” onward. Browsing the Chinese correspondents (on the right) to the concordances of “absenteeism” (on the left) immediately reveals the markedly frequent occurrences of both 旷工 *kuànggōng* ‘absent from work’ and 旷课 *kuàngkè* ‘absent from school/class’. Translators can use the “Find on Page” or the “Find in This Page” feature in the commonly used web browsers (e.g., Firefox, Google Chrome, Microsoft Edge) to highlight *all* the instances of the character旷 *kuàng* on the page, as illustrated in Fig. 2. In so doing, they can quickly scan through the Chinese expressions with highlighted instances of 旷 and then draw their attention to the lines in which 旷 is absent to discover other translation candidates for “absenteeism”. For instance, 缺勤率 *quēqín lǜ* ‘absence/absenteeism rate’ that occurs in the second last line of the Chinese concordances (underlined by the investigator) in Fig. 2 is a case in point. In the same way, translators can search the individual –ism words they need to study for their interest or for their translation tasks and look into the Chinese concordances for a wealth of useful translation candidates.



doc#0	School refusal – any kind of **absenteeism** , from phobia to truancy, that can be traced to the child 's own actions and wishes – is at the very intersection of education, psychology and pediatrics.	拒学症 --- 任何 一 种 旷课 , 从 恐惧 到 逃学 , 都 能 从 孩子 自身 的 品行 和 愿望 上 找到 根源 --- 而且 是 教育 , 心理学 和 儿科学 等 因素 交织 在 一起 。
doc#0	Economic disruptions could come from high **absenteeism** , as people may be asked to stay at home, or may choose to do so to care for sick relatives or because of fear of being exposed.	经济 恐 因 高 旷工率 而 陷入 混乱 , 因为 许多 人 可能 被 要求 待 在 家中 , 或 选择 待 在 家中 照顾 生病 的 亲戚 , 或是 因为 担心 被 传染 而 不 出门 上工 .
doc#0	Alcohol is associated with many serious social and developmental issues, including violence, child neglect and abuse, and **absenteeism** in the workplace.	酒精 造成 许多 严重 的 社会 和 发展 问题 , 如 暴力 、 儿童 遭受 忽视 和 虐待 以及 旷工 等 问题 。
doc#0	Signals to be vigilant for include spikes in rates of **absenteeism** from schools or workplaces, or a more severe disease pattern, as suggested by, for example, a surge in emergency department visits.	需要 保持 警惕 的 现象 有 : 学校 旷课率 或 工作 缺勤率 急剧 上升 , 或 呈现 出 更加 严重 的 疾病 特征 , 正 如 急诊科 的 病人 出现 了 猛增 所 表明 的 情况 那样 。
doc#0	Some of your people will also need to bedismissed because of excessive **absenteeism** .	你 的 某些 员工 , 经常 旷工 , 因而 也 需要 被 开除 。
doc#0	Any degree of **absenteeism** would cause me to fail Math and to lose my scholarship. I was so afraid of failing that I even pledged abstinence from my abundant supply of alcohol.	旷课 次数 无论 多少 都 可能 导致 我 数学 的 不 及格 , 甚至 还 发誓 要 对 自己 充裕 的 酒类 供应 加以 节制 。
doc#0	More human resources departments are buying them for employees to reduce **absenteeism** caused by illness.	更 多 的 人力 资源 部门 在 为 其 员工 购买 这些 产品 , 来 减少 由于 生病 导致 的 旷工 情况 。
doc#0	"A large proportion of alcohol-related **absenteeism** is due to alcohol hangovers, which are more common for light to moderate drinkers than for heavy drinkers," said the study.	该 研究 报告 表示 : " 在 由 饮酒 所 引发 的 旷工 现象 中 , 很多 情况 下 都 是 由于 酒醉 后 出现 的 身体 不适感 造成 的 , 与 重度 饮酒者 相比 , 此 种 行为 在 那些 轻度 到 中度 的 饮酒者 中 显得 更 为 常见 一些 。 "
doc#0	Without regulations, women "suffered from fines for **absenteeism** , late arrival, shoddy production, laughter, and even nursing babies ."	由于 没有 规定 , 女性 由于 旷课 , 迟到 , 抄袭 , 大笑 , 甚至 养育 孩子 而 被 罚款 。
doc#0	Durfee joins other U.S. schools in Massachusetts and Illinois and New York which have taken on the added role of alarm clock to combat high rates of tardiness and **absenteeism** .	德菲 高中 以及 马萨诸塞州 、 伊利诺伊州 和 纽约州 的 不少 学校 , 还 使用 不同 人物 的 声音 作为 闹 铃音 , 希望 以 此 来 杜绝 棘手 的 迟到 和 旷课 问题 。 在 纽约市 , 有的 学校 使用 的 是 前 职业 篮球 明星 魔术师 约翰逊 的 声音 作为 闹 铃音 。
doc#0	The benefits to companies of adopting flextime, the report says, "can outweigh the costs by reducing **absenteeism** , lowering turnover, improving the health of workers, and increasing productivity ."	该 报告 指出 : 采用 弹性 工作 时间 的 公司 获得 的 收益 , " 会 超出 其 成本 支出 , 这 是 由于 旷工 减少 、 人员 流动率 下降 、 员工 健康 状况 改善 以及 生产力 提高 。 "
doc#0	A company usually must either pay overtime, close on the holiday, or face high levels of employee **absenteeism** , turnover, and unrest.	在 节假日 , 公司 通常 要么 支付 加班费 , 要么 关门 , 否则 就 会 有 很 高 的 缺勤率 、 补 缺 人员 较 多 , 而且 还 会 出现 不 安定 的 局面 。

Fig. 2 Parallel concordance of “non-*zhuyi*” –ism words with highlighted characters in the Chinese concordances

Those frequently occurring –ism words lead to a large number of concordance lines, e.g., “mechanism” has 1,321 lines and “criticism” 334. Browsing through the concordance lines takes a considerable amount of time and efforts, although this allows translators to thoroughly identify all the translation candidates in the corpus. We were able to manually collect a large repertoire of translation candidates for “mechanism”, of which the frequently occurring ones are presented in Table 3. Meanwhile, there are also other much less frequently used ones such as 病机 *bìngjī*, ‘pathogenesis, 心理 *xīnlǐ* ‘psychology’, 方法 *fāngfǎ* ‘method, means’ and so on, which are not included in Table 3 though.

To save the labour of manually picking through the concordance lines, translators can create a subcorpus of the Chinese concordances of the specific –ism word/s they want to focus on. Using the subcorpus they can generate the lists of keywords and key terms in which the frequently used Chinese item/s corresponding to the –ism word under investigation should emerge. We take “mechanism” as an example again, since it is by far the most frequently used –ism word in our sample, with 1,321 tokens in total and taking both singular and plural forms. It corresponds to a wide range of terms in Chinese, as previously discussed. One can perform parallel-text concordancing by searching the lemma of “mechanism”, using the English corpus as the primary corpus and the Chinese one as the secondary. From the query results, one can create a subcorpus of the Chinese concordance lines displayed on the right hand of the parallel-concordance interface and switch to the Chinese corpus (from using the English corpus) so as to have access to the subcorpus of “mechanism”. The subcorpus can then be processed under the keywords tab to extract keywords (termed “single-words” in SkE) and key terms (termed

Table 3 The list of Chinese translation correspondents of “mechanism”

Chinese terms with *Pinyin*	Gloss in English
机制 *jīzhì*	‘mechanism’
机理 *jīlǐ*	‘principle, (reaction) mechanism”
作用 *zuòyòng*	‘effect, function’
机构 *jīgòu*,	‘device, mechanism’
机械 *jīxiè*	‘mechanical (device)’
结构 *jiégòu*	‘structure’
装置 *zhuāngzhì*	‘device’
原理 *yuánlǐ*	‘principle, theory’
系统 *xìtǒng*	‘system”
机器 *jīqì*	‘machine’

“multi-words terms” in SkE), using Chinese Web (enTenTen17, simplified characters) as the reference corpus, which comes as the default. The top translation candidates for “mechanism” we manually identified above also appear in the keyword list—e.g., 机制 (n = 103), 机理 (n = 41), 作用 (n = 28), 机构 (n = 20), 装置 (n = 9) (cf. Table 3). The key-term list further indicates that some of the keywords tend to join each other to form collocations, e.g., 作用机制*zuòyòng jīzhì* ‘mechanism of action’, 作用机理 *zuòyòng jīlǐ* ‘mechanism or principle of action’, 机械装置 *jīxiè zhuāngzhì* ‘mechanical device’. The results not only attest to the significance of the translation candidates but also provide useful collocational patterns of the items. To reiterate, creating and using the subcorpus enables translators to harvest the frequently occurring Chinese translation candidates rather efficiently and effectively at SkE.

Once the list of the translation candidates has been produced, inquisitive translators would then need to sort out the (typical) situations in which the major translation candidates tend to be used. In the example of “mechanism”, the list of the Chinese translation candidates contains items relating to a variety of technical fields and domains, of which translators may not have specific knowledge. The translators, however, can always draw on the subcorpus that consists of the English and Chinese concordances as described above to better understand the translation candidates in terms of their context of use. With this in mind, we now examine five frequently occurring translation candidates 机械, 机理, 作用, 机构 (cf. Table 3) and also 病机, which is a truncated form.

The term 机械 *jīxiè* ‘mechanical, machinery’ emerges in our data with close relevance to mechanical engineering. Turning on “Find on page” in the web browser to highlight all the instances of 机械 in the parallel concordances of the subcorpus on “mechanism”, translators can quickly locate the typical co-texts in which the word “mechanism” occurs in English while the term 机械 is used in Chinese. The examples include “good mechanism performance” in English, which is aligned to 优良的机械物理性能 *yōuliáng de jīxiè wùlǐ xìngnéng* ‘excellent mechanical (and) physical properties/performance' in Chinese, and "[a] mechanism (for launching aircraft)" in English corresponding to (发射航空器的)一种机械装置 '(*fāshè hángkōngqì de*) *yī zhǒng jīxiè zhuāngzhì*' 'a mechanical device (for launching an aircraft)' in Chinese. From the contexts of use, we can observe that 机械, when used as a translation correspondent for "mechanism", denotes mechanical devices, machines, or appliances. Similarly, by "finding" the term 机理 on the concordance page, one can soon realise that it tends to refer to the scientific or theoretical understanding, basis, or principles that explain certain phenomena or observations in natural sciences as well as in pathology. The parallel sentences show that 机理 can refer to the mechanism for a gravitational action in physics, that for crops breeding in biology, that for the ageing of materials in chemistry, and that for a virus infection or brain's reactions to drugs in pathology. In addition, searching the translation candidate 作用 on the concordance page shows that it tends to form some frequently occurring collocations such as 作用机制 or 作用机理 (cf. Table 3). Such collocations also appear in the multi-term list at SkE, which refer to the reason for which, or the underlying processes by which, a (pharmaceutical or pathological) effect or outcome takes place. These Chinese collocations closely correspond to the English expression "mechanism of action" in terms of word-for-word meaning—i.e., "action" for 作用 and "mechanism" for 机理 or 机制 (cf. Table 3). However, the phrase "mechanism of action" occurs infrequently in our English data, while the single word "mechanism" tends to be commonly used in context to refer to the action of drugs. For example, "mechanism" in the English concordance line matches a longer expression (药物) 作用机制 *(yàowùǐ) zuòyòng jīzhì* '(drug's) action/effect mechanism' in the Chinese line. There are also situations in which 作用 is used alone to translate "mechanism", e.g., "the anti-inflammatory mechanisms (may be related to…)" in English is rendered into 抗炎作用 (可能与…有关) *kàng yán zuòyòng (kěnéng yǔ…yǒuguān)* 'anti-inflammatory effect (may be related to…)" in Chinese. This shows that the collocation 作用机制 in Chinese is simplified into 作用, while the meaning of 机制 can be still derivable from the context. Moreover, translators can easily note from the concordance lines that, when 机构 is used to render "mechanism", it refers to a *mechanical* device, tool, or structure rather than an institution or (administrative, educational) establishment. Furthermore, our concordance data on "mechanism" contains both (a) the compound expressions such as 发病机制 *fābìng jīzhì*, 致病机理 *zhì bìng jīlǐ*, and 疾病机制 *jíbìng jīzhì*, which all have the meaning of 'pathogenesis' and (b) the shortened form 病机*bìng jī*. Sharp-eyed translators should quickly figure out that the latter is a truncated form of any the three compounds.

From a wide range of translation candidates for "mechanism", we can observe that each of the expressions in Chinese tends to point to a certain sense of "mechanism" in English. Sorting out different translation candidates turns out to be a process of laying out various senses that the word "mechanism" entails. This learning process allows translators to understand "mechanism" in terms of its fine-grained shadings of meaning by its translation correspondents.

4.4 From Chinese Translations to the Senses of –ism Words

The parallel corpus not only allows translators to conduct self-directed learning of the translation of –ism words from English into Chinese (cf. Sect. 4.3), and it can also be used for queries in the other direction—i.e., from Chinese to English—translators can investigate the commonly used Chinese expressions and seek out the (type of) –ism words that tend to be rendered into these expressions. For example, 论 *lùn* 'doctrine, discourse, theory' is a very notable character on the keyword list of the subcorpus of the Chinese concordances corresponding to –ism words in English, which appears in various expressions in Chinese—e.g., 论 in 不可知论 *bùkězhī lùn* 'agnosticism', 无神论 *wúshén lùn* 'atheism', 建构论 *jiàngòu lùn* 'constructivism'. Translators may further sort out the (types of) –ism words that tend to be rendered into the …论 expressions in Chinese. An exhaustive search can be conducted by a query similar to the "containing" query in Fig. 1, in which 论 replaces 主义 to be the character contained in the Chinese corpus. The query returns 749 parallel concordance lines in the "Education" corpus, while the English concordances present the resources in which the –ism words corresponding to 论 are to be retrieved. From the frequency list of KWIC lemmas of the English concordances, translators can spot and then confirm with examples in the concordances the –ism words that correspond to …论 terms (see Table 4).

Table 4 The –ism words that correspond to …论 terms in Chinese

The –ism words in English	The corresponding Chinese terms and *pinyin*	The gloss of Chinese terms
aestheticism	美学理论 *měixué lǐlùn*	'aesthetic theory or doctrine'
agnosticism	不可知论 *bù kězhī lùn*	'agnostic doctrine'
animism	泛灵论 *fàn líng lùn*	'animist, or pan-spirit, doctrine'
anthropomorphism	神人同性论 *shén rén tóngxìng lùn*, 拟人论 *nǐrén lùn*	'anthropophuistic doctrine', 'personification doctrine'
atheism	无神论 *wúshén lùn*	'atheist doctrine'
bloodline-ism	血统论 *xuètǒng lùn*	'blood-line doctrine, pedigree'
cognosciblism	可知论 *kězhī lùn*	'gnostic doctrine'
constructionism	建构论 *jiàngòu lùn*	'constructivist doctrine'
dogmatism	教条论 *jiàotiáo lùn*	'dogma(tist) doctrine'
electromagnetism	电磁理论 *diàncí lǐlùn*	'electromagnetic theory'
esotericism	隐微理论 *yǐn wēi lǐlùn*	'esoteric or implicit theory'
functionalism	德国功能派翻译理论 *déguó gōngnéng pài fānyì lǐlùn*	'German functionalist translation theory'
Marxism	马克思主义理论 *mǎkèsī zhǔyì lǐlùn*	'Marxist theory'
nihilism	国家主权虚无理论 *guójiā zhǔquán xūwú lǐlùn*	'national sovereignty nihilist theory'
syllogism	三段论 *sānduàn lùn*	'three-statements theory'

The –ism words in Table 4 strongly suggest that the …论 expressions correspond to those –ism words that denote schools of thoughts, scholarly or ideological positions and theories. The sense expressed is very close to that of 主义, and indeed several –ism words in Table 4 correspond to both 论 and 主义. For example, “aestheticism” can be also rendered into 唯美主义*wéiměi zhǔyì*, “dogmatism” into 教条主义 *jiàotiáo zhǔyì*, “nihilism” into 虚无主义 *xūwú zhǔyì*, and “functionalism” into 功能主义 *gōngnéng zhǔyì*. These words can be considered “partly–*zhuyi*” words, for which both 主义 and 论 are major translation candidates. In addition, the word “Marxism” can be rendered into 马克思 主义理论 in some contexts (cf. Table 4), in which 主义 and 论 co-occur to form a Chinese word cluster. Translators can certainly hold that “Marxism” is a *zhuyi*-word, given the fact that its Chinese correspondents almost always contain 主义 (cf. Sect. 4.2), though truncated forms can be used too. Moreover, from the corpus evidence, translators can understand that the Chinese correspondents of “Marxism” can be expanded from 马克思主义 to 马克思主义 理论 (cf. Table 4) and 马克思主义思想 *mǎkèsī zhǔyì sīxiǎng* ‘Marxist thoughts or ideology’. Furthermore, unlike the *zhuyi*- and partly-*zhuyi* words, there are words in Table 4 that are rendered into …论 but *not* …主义 expressions in our data, e.g., “agnosticism”, “electromagnetism”, and “syllogism”, which are clearly non-*zhuyi* words.

Using the same methods as described for querying 论 above, translators can investigate other frequently occurring Chinese expressions and their corresponding –ism words in English. Some main findings of our searches are presented in Table 5, although, given the space of this chapter, we are unable to enlist all the notable recurring Chinese expressions that emerged in our data. From the Chinese expressions, we can observe that the –ism words entail at least three major senses—i.e., (a) action, (b) doctrine, and (c) characteristics or particularities of language, disease, and so on—which we can now illustrate with examples.

The first sense pertains to action/s or behaviour/s of some kinds. For example, 无意识行为 *wúyìshí xíngwéi* ‘non-conscious act’ corresponds to “automatism”, which denotes actions performed unconsciously. The actions can be performed collectively in the form of a campaign or a social movement as well—e.g. 反恐运动 *fǎnkǒng yùndòng* ‘antiterrorist campaign’ that corresponds to “counterterrorism” in English. In addition, the action(s) can be denoted by the actors or performers engaged in the action, and, for this reason, we can put this particular sense to the first sense of –ism words. For example, the word “activism” in (2) is rendered into 活动家 *huódòng jiā* ‘activists’, while “futurism” in (3) is rendered into 未来学家 *wèiláixué jiā* ‘futurology scholar/s’, which sounds appropriate and idiomatic in context. In fact, translators can discover that the “Education” corpus contains rich examples in which the Chinese expressions with …家 *jiā* ‘master, scholar’ or …者 *zhě* ‘practitioner, participant’ correspond to –ism words in English. For example, 儒家 *rú jiā* ‘Confucius followers or doctrine’ corresponds to “Confucianism”, referring to both the followers and the school of Confucianism. Similarly, 不可知论者 *bù kězhīlùn zhě* ‘not-knowable-doctrine believer’ corresponds to “agnosticism” in context.

Table 5 Chinese expressions and the corresponding –ism words in major sense (groups)

Chinese expressions, *pinyin* and gloss	Chinese Examples, *pinyin* and gloss in English	The corresponding –ism words
(a) **Action**		
…行为 *xíngwéi* ‘act, action, behaviour’	无意识行为 *wúyìshí xíngwéi* ‘non-conscious act’	automatism
	破坏 (公共财物) 行为 *pòhuài (gōnggòng cáiwù) xíngwéi* ‘damaging (public assets) acts or behaviours’	vandalism
	反犹太行为 *fǎn yóutài xíngwéi* ‘anti-Semitic act’	anti-Semitism
	恐怖主义行为 ‘terrorism act’	terrorism
	利他行为 *lìtā xíngwéi* ‘benefitting others act’	altruism
(a1) **Actor or practitioner**		
…家 *jiā* ‘a master, expert, or scholar of a doctrine, or the school of thoughts promoted by the practitioner’	活动家 *huódòng jiā* ‘activist’	activism
	儒家 *rú jiā* ‘Confucius followers or the school’	Confucianism
	道家 *dào jiā* ‘Daoist person or doctrine’	Daoism
	佛家 *fó jiā* ‘Buddhist person or doctrine’	Buddhism
	法家 ‘legalist person or doctrine’	Legalism
	批评家 *pīpíng jiā* ‘criticising person, critics’	criticism
	未来学家 *wèiláixué jiā* ‘futurology scholar’	futurism
(b) **Doctrine, school of thoughts; mentality**		
…学 *xué* ‘study, doctrine’	孔子学说 *kǒngzǐ xuéshuō*, 儒学 *rúxué* ‘Confucius doctrine or studies’	Confucianism
	汉学 *hànxué* ‘Han or Chinese studies’	sinologism
	行为学派 *xíngwéi xuépài* ‘the behaviourist doctrine or school’	behaviourism
	(生物)电磁学 *(shēngwù) diàncí xué* ‘(bio-) electromagnetic studies’	(bio)electromagnetism
	东方学 *dōngfāng xué* ‘Oriental studies’	Orientalism

(continued)

Table 5 (continued)

Chinese expressions, *pinyin* and gloss	Chinese Examples, *pinyin* and gloss in English	The corresponding –ism words
…精神 *jīngshén* 'spirit, mentality, or principle'	体育精神 *tǐyù jīngshén* 'athletic spirit'	athleticism
	乐观精神 *lèguān jīngshén* 'optimistic spirit'	optimism
	爱国精神 *àiguó jīngshén* 'patriotic spirit'	patriotism
	人道主义精神 *réndào zhǔyì jīngshén* 'humanitarian spirit'	humanitarianism
	敬业精神 *jìngyè jīngshén* 'respect-profession spirit'	professionalism
…思想 *sīxiǎng* 'thoughts, ideas, ideology'	儒家思想 *rújiā sīxiǎng*, 孔子思想 *kǒngzǐ sīxiǎng* 'Confucius thoughts'	Confucianism
	宪政思想 *xiànzhèng sīxiǎng* 'constitutional thoughts'	constitutionalism
	君权思想 *jūnquán sīxiǎng* 'kingship thoughts, ideology'	royalism
(b1) **Systems or institutions established under certain doctrines or ideologies**		
…制 *zhì* 'system, institution, or establishment'	宪政体制 *xiànzhèng tǐzhì* 'constitutional system'	constitutionalism
	工会制度 'union system'	unionism
	集权制度 *jíquán zhìdù* 'centralised-power system'	authoritarianism
	联邦制度 *liánbāng zhìdù* 'federal system'	federalism
	封建制度 *fēngjiàn zhìdù* 'feudal system'	feudalism
(c) **Characteristics of language, pathology…**		
…语 *yǔ* 'language, word/s, discourse'	双语 *shuāngyǔ* 'bilingual'	bilingualism
	警语 *jǐng yǔ* 'warning or pithy words'	aphorism
…症 *zhèng* 'disease'	自闭症 *zì bì zhèng* 'self-closed disease'	autism
	酒精中毒症 *jiǔjīng zhòngdú zhèng* 'alcohol poisoning disease'	alcoholism
	梦游症 *mèngyóu zhèng* 'dream-travel disease'	somnambulism

(2)
English FAIR’s Media Activism Resources
Chinese 众多媒体活动家资源
Zhòngduō méitǐ huódòng jiā zīyuán
‘Zhongduo media activist resources’

(3)
English Futurism rejected all traditions
Chinese 未来学家拒绝一切传统
Wèilái xué jiā jùjué yīqiè chuántǒng
‘Futurology scholars rejected all traditions”

The second sense denotes certain doctrines, thoughts or ideologies. The expressions with 学 *xué* ‘studies, doctrine’ are commonly used, e.g., 汉学 *hànxué* ‘Han or Chinese studies’ corresponds to “Sinologism” in English. Also broadly under the second sense, we can find expressions denoting people’s personality traits, mentalities, convictions and the principles they subscribe to, e.g., 爱国精神 *àiguó jīngshén* ‘patriotic spirit, mind’ in (4) corresponds to “patriotism” in English. In addition, the systems or institutions established under certain doctrines or ideologies can be put under the second sense. For example, in (5), “constitutionalism” in English is rendered into 宪政制度 *xiànzhèng zhìdù* ‘constitutional system’ in Chinese. More examples can be found in Table 5.

(4)
English We’ve seen patriotism slide into jingoism
Chinese 我们看到爱国精神不知不觉地陷入侵略主义
Wǒmen kàn dào àiguó jīngshén bùzhī bùjué de xiànrù qīnlüè zhǔyì
‘We see patriotic spirit unnoticeably falling into invad-ism’

(5)
English Impeachment system is one of the important parts of western constitutionalism
Chinese 弹劾制度是西方宪政制度中的一项重要制度 。
Tánhé zhìdù shì xīfāng xiànzhèng zhìdù zhōng de yī xiàng zhòngyào zhìdù
‘Impeachment system is an important system in the western constitutional system”

Finally, the third sense of –ism words denotes (the characteristics, style, or features of) language varieties, diseases, and pathological conditions and so on. The typical examples are “colloquialism”, “autism”, and “alcoholism”, which are usually rendered into Chinese expressions with 语 *yǔ* ‘language, discourse’, 言 *yán* ‘words, language’ and 症 *zhèng* ‘disease, syndrome of certain disease’. The examples include 口语 *kǒuyǔ* ‘spoken language (style)’ for “colloquialism”, 自闭症 *zì bì zhèng* ‘self-closed disease’ for “autism”, and 酒精中毒症 *jiǔjīng zhòngdú zhèng* ‘alcohol poisoning disease’ for “alcoholism”.

From the Chinese lexical correspondents of –ism words, we can identify three major sense groups of –ism words. The “Education” corpus therefore exhibits much potential for translators to tease out the senses of –ism words and harvest useful translation candidates. The skills and competence for conducting corpus-based queries as demonstrated in Sect. 4 are not technically advanced or highly sophisticated. As a result, we believe that self-directed inquisitive translators are able to reach discoveries in their interested areas insomuch as they are guided by their questions and are ready to make efforts for seeking out their answers in the corpus.

5 Discussion

We can now discuss the results of our study in relation to the research questions outlined in Sect. 1.

5.1 Traditional Tools Versus Parallel Corpora for Translators

From our examination of the conventional tools for translators to deal with –ism words, monolingual English dictionaries are valuable in laying out the major senses of the suffix, while reverse English-Chinese dictionaries are particularly useful for providing key translation candidates for a collection of –ism words presented altogether. These tools and resources assist translators to gain an overview of –ism words in terms of their senses and allow them to understand that a variety of Chinese translations tends to correspond to different –ism words. However, large parallel corpora have unique merits for translators to conduct (exhaustive) searches that yield quantitative results. The quantitative results lead to the emergence of the patterns and distributions of –ism words and their Chinese translations, while the parallel concordances allow –ism words and their translations to be systematically gathered, analysed, and examined in textual contexts. Large parallel corpora therefore extend a rather solid basis for teasing out the senses of –ism words, classifying the words into major types in relation to their Chinese translations, and harvesting translation candidates. More importantly, inquisitive translators can devise their own queries, in particular, with corpus query language (CQL) at the SkE interface and sort out their interested issues. The values of parallel corpus in these respects, especially the flexibility provided to translators for in-depth investigations, clearly exceed those that traditional resources can offer.

In terms of the denotations entailed by –ism words and their connotations invoked in context, dictionaries tend to give definitions with emphasis placed on the former, while large parallel corpora tend to capture the latter by both the –ism words used in various contexts and their Chinese correspondents. Translators certainly need to gain a sound understanding of the denotative meanings, but still, the connotative meanings brought to light by the Chinese correspondents of –ism words are of practical use to them as well. The translation candidates strongly suggest that there are different alternatives that previous translators have utilised in the meaningful contexts in which they have worked. The alternatives bring to life authentic translation situations, to which translators can relate and be more resourceful, imaginative, and inspired.

5.2 Parallel Corpus for Translators: Drawbacks and Merits

There have been scholarly debates in the literature on the drawbacks and merits of parallel corpora for lexicography and for linguistic studies, e.g., an early debate arose between Teubert (1996) and Mauranen (2002), both leading researchers of the field. Parallel corpora tend to be much more modest in size compared to both monolingual corpora and comparable corpora, with a narrower scope of genres and text types available. Automatic alignment between the source and the target texts tends to involve a scope of inaccuracy, while manual checking of the aligned texts, especially at sentence level, is laborious and time-consuming. In addition, the issue of the direction of translation has long been raised, while a recent discussion can be found in Mikhailov (2021 in this volume). The direction of translation indeed should be specified at the stage of the construction of the corpus so that the translation equivalents can be specified accordingly, and both the frequency of occurrence and the distributions of the array of translation equivalents can make real sense.

Having said this, however, parallel corpora with unspecified direction of translation like the "UM-Education corpus" we explored in this study can still be valuable to translators. Most translators use corpora not for carrying out translation studies as translation scholars do—e.g., systematically describing translation equivalents or translation stimuli. Neither are they aiming at conducting contrastive linguistics analysis for generalisable results as attempted by linguists. Their primary need is to find their way to translation solutions, in particular, gathering useful translation candidates for their translation assignments. The parallel corpora that provide them with words used in the contexts similar to the ones they need to deal with can be potentially useful resources, with which they can select the most useful translation candidates. If the parallel corpora come with rich meta-data on the direction of translation and information on the institutional or the technical context in which the text is produced, that will certainly be context-rich information that translators can make sense of and derive insights from.

5.3 Technological Skills for Translators

Using monolingual corpora to generate wordlists with SkE (cf. Sect. 3.3) only requires basic skills. Querying parallel corpus at the interface of Parallel

Concordance is very attainable by translators with average computer skills. SkE-based access to the aligned bilingual texts is rather user-friendly, and the setting for various built-in query methods (esp. CQL) gives translators flexibility to devise and carry out queries that answer their questions surrounding the translation of –ism words. Self-motivated translators can acquire the skills and benefit a great deal by switching back and forth between different interfaces and resources at SkE for answering the questions they want to pursue. For example, by shifting between parallel concordance and the subcorpora that can be created from the parallel concordance lines, translators can produce keyword and key-term lists to spot recurrent Chinese expressions that correspond to the –ism words they seek to investigate (cf. Sect. 4.3). Translators' competence in making effective use of corpus tools and resources give them clear advantages in their professional practices, compared with those who rely only on traditional resources such as dictionaries and glossaries.

6 Conclusion

This study examines conventional and more recent corpus tools and resources for translators to gain practical knowledge of –ism words in English and retrieve their translation candidates in Chinese. We discovered that the traditional tools such as monolingual and bilingual dictionaries, especially reverse English-Chinese dictionaries, present translators the major senses of the suffix –ism and diversified translations of –ism words. In addition, large monolingual corpora of both English and Chinese, when accessed by translators with basic corpus skills, provide the lists of both –ism words and –主义 words in English and Chinese respectively, sortable by frequency to enable comparing and contrasting between the two. Moreover, large-scale parallel corpora exhibited unique value in terms of providing a large repertoire of translation candidates for a decent coverage of –ism words in a variety of contexts. The parallel concordance lines present a wealth of resources with which inquisitive translators can devise their queries to reach revealing results. The findings would allow translators to tease out their practical classification of different types of –ism words, appreciate the different senses denoted by the –ism words, and identify their typical translations into Chinese in certain situations or cotexts. We would argue that the ability to utilise corpus tools and resources, in particular, large-scale or subject-specific parallel corpora, constitutes an important and integral part of professional translators' competence in the IT age, and this competence needs to be fostered in the translator training curriculum. Subsequent studies can draw closer attention to (trainee) translators' actual use of the corpus-based tools and the ways they develop their skills and competence in this respect.

Acknowledgements The author thanks the anonymous reviewers of this chapter for their valuable comments and suggestions and acknowledges the research grant (RP/ESLT-01/2016) from the Macao Polytechnic Institute.

References

Baker, M. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4: 281–298. https://doi.org/10.1075/ijcl.4.2.05bak.

Chen, X., V.X. Wang, and C.R. Huang. 2020. Sketching the English translations of Kumārajīva's The Diamond Sutra: A comparison of individual translators and translation teams. In *Proceedings of the 34th pacific asia conference on language, information and computation*, ed. L.M. Nguyen, C.M. Luong, and S. Song, 30–41.

Defrancq, B., and G. Rawoens. 2016. Assessing morphologically motivated transfer in parallel corpora. *Target-International Journal of Translation Studies* 28 (3): 372–398. https://doi.org/10.1075/target.28.3.02def.

Doval, I., and M.T.S. Nieto. 2019. Parallel corpora in focus: An account of current achievements and challenges. In *Parallel corpora for contrastive and translation studies: New resources and applications*, ed. I. Doval and M.T.S. Nieto, vol. 90, 1–15. John Benjamins.

Fillmore, C.J. 1992. Corpus Linguistics or Computer-aided armchair linguistics. In *Directions in corpus linguistics: Proceedings of Nobel Symposium*, ed. Jan Svartvik, vol. 82, 35–60. Stockholm: Walter de Gruyter.

Johansson, S. 2007. *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. John Benjamins.

Kovář, V., V. Baisa, and M. Jakubíček. 2016. Sketch engine for bilingual lexicography. *International Journal of Lexicography*, 29 (3): 339–352. https://doi.org/10.1093/ijl/ecw029

Kubicka, E. 2019. So-called dictionary equivalents confronted with parallel corpora (and the consequences for bilingual lexicography). *Glottodidactica. An International Journal of Applied Linguistics*, 46 (2): 75–89.

Laviosa, S., and G. Falco. 2021. Using corpora in translation pedagogy. In *New perspectives on corpus translation studies*, ed. V.X. Wang, L. Lim, and D. Li, 3–27. https://doi.org/10.1007/978-981-16-4918-9_1.

Lefer, M.A. 2012. Word-formation in translated language: The impact of language-pair specific features and genre variation.*Across Languages and Cultures* 13 (2): 145–172. https://doi.org/10.1556/Acr.13.2012.2.2.

Lefer, M.A., and N. Grabar. 2015. Super-creative and over-bureaucratic : A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures* 16 (2): 187–206. https://doi.org/10.1556/084.2015.16.2.3.

Lefever, Els, and Véronique Hoste. 2014. Parallel corpora make sense: Bypassing the knowledge acquisition bottleneck for Word Sense Disambiguation. *International Journal of Corpus Linguistics* 19 (3): 333–367.

Li, D. 2017. Translator style a corpus-assisted approach. In *Corpus methodologies explained an empirical approach to translation studies*, ed. M. Ji, M. Oakes, D. Li, and L. Hareide, 103–136. Routledge.

Li, L., S. Dong, and V.X. Wang. 2020a. Gaige and reform: A Chinese-English comparative keywords study. In *From minimal contrast to meaning construct: Corpus-based, near synonym driven approaches to Chinese lexical semantics*, ed. Q. Su, and W. Zhan, 321–332. Springer/Peking University Press.

Li, L., C.R. Huang, and V.X. Wang. 2020b. Lexical variations and human behavior changes: A corpus-assisted Investigation of GAMBLING and GAMING in the Past Centuries. *SAGE Open* 10 (3): 1–14. https://doi.org/10.1177/2158244020951272

Lim, L. 2019. Are terrorism and *kongbu zhuyi* translation equivalents? A corpus-based investigation of meaning, structure and alternative translations. In *Proceedings of the 33rd pacific asia conference on language, information and computation*, ed. R. Otoguro, M. Komachi, and T. Ohkuma, 516–523.

Lim, L. 2020. Interpreting training in China: Past, present and future, In *Key issues in translation studies in China: Reflections and new insights*, ed. L. Lim, and D. Li, 143–160. https://doi.org/10.1007/978-981-15-5865-8_7

Loveard, K. 1994. Sukarno: A fallen hero's legacy. *Asiaweek*, 36.

Mauranen, A. 2002. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2: 161–185.

Mauranen, A. 2004. Contrasting languages and varieties with translational corpora. *Languages in Contrast* 5 (1): 73–93.

Mikhailov, M. 2021. Mind the source data! Translation equivalents and translation stimuli from parallel corpora. In *New perspectives on corpus translation studies*, ed. V.X. Wang, L. Lim, and D. Li, 259–279. Springer. https://doi.org/10.1007/978-981-16-4918-9_10.

Quah, C.K. 1999. Issues in the translation of English affixes into Malay. *Meta* 44 (4): 604–616. https://doi.org/10.7202/003881ar.

Sinclair, J. 2001. Data-derived multilingual lexicons. *International Journal of Corpus Linguistics*, 6 (Special Issue): 79–94.

Teubert, W. 1996. Comparable or parallel corpora? *International Journal of Lexicography* 9 (3): 238–264. https://doi.org/10.1093/ijl/9.3.238.

Teubert, W. 2001. Corpus linguistics and lexicography. *International Journal of Corpus Linguistics*, 6 (SI): 125–153. https://doi.org/10.1075/ijcl.6.si.11teu

Tian, L., D.F. Wong, L.S. Chao, P. Quaresma, F. Oliveira, and L. Yi. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, ed. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, 1837–1842. European Language Resources Association (ELRA).

Wang, Shan, and Huang Chu-Ren. 2017. Word sketch lexicography: New perspectives on lexicographic studies of Chinese near synonyms. *Lingua Sinica* 3 (1): 1–22.

Wang, V.X. 2018. Sketching a Chinese writer's vocabulary profile in English: The case of Ha Jin. In *Proceedings of the 32nd pacific asia conference on language, information and computation*, ed. S. Politzer-Ahles, Y.Y. Hsu, C.R. Huang, and Y. Yao, 722–728.

Wang, V.X. 2021. Examining prefix de- with its translations in Chinese: Making sense with parallel corpora. In *New perspectives on corpus translation studies*, ed. L. Lim, D. Li, and V.X. Wang, 299–318. Springer. https://doi.org/10.1007/978-981-16-4918-9_12.

Wang, V.X., and L. Lim. 2017. How do translators use web resources? Evidence from the performance of English-Chinese translators. In *Human issues in translation technology*, ed. D. Kelly, 63–79. Routledge.

Wang, V.X., X. Chen, S. Quan, and C.R. Huang. 2020. A parallel corpus-driven approach to bilingual oenology term banks: How culture differences influence wine tasting terms. In *Proceedings of the 34th pacific asia conference on language, information and computation*, ed. L.M. Nguyen, C.M. Luong, and S. Song, 318–328.

Xiao, R. 2010. How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics* 15 (1): 5–35.

Zanettin, F. 2014. *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Routledge.

References in Chinese

陆谷孙. 2007. 英汉大词典 (2nd ed.). 上海译文出版社.

李龙兴, & 王宪. 2021. 应急语言服务视角下的新冠肺炎医学英语专题术语表开发 (The development of COVID-19 word list from the perspective of emergency language services). 中国科技术语 (*China Terminology*), *23*(2), 32–41.

孙梅. 1993. 英汉倒排词典 (*A reverse English-Chinese dictionary*). 广西教育出版社.
荀恩东, 饶高琦, 肖晓悦, & 臧娇娇. 2016. 大数据背景下 BCC 语料库的研制. 语料库语言学 (*Corpus Linguistics*), *3*(1), 93–118.
郑宝璇. 2004. 传媒翻译. 香港城市大学出版社.

Lily Lim holds a PhD in Applied Linguistics (University of Queensland), a Master's Degree in Software Engineering (University of Macau), Certificate of Training Techniques (Escolas da Armada, Portugal), and Certificate of Chinese-Portuguese Conference Interpreting (Comissão Europeia). She has been both a practising interpreter and trainer for conference interpreters for nearly two decades. She is currently an Associate Professor and Coordinator of the Chinese-English Translation Program at the School of Languages and Translation, Macao Polytechnic Institute. Her recent research covers computer-assisted interpreter and translator training, and corpus-based language studies. She has published papers in *ReCALLBabel*, and *The Interpreter and Translator Trainer*; book chapters with Rodopi, Springer, Cambridge Scholars Publishing, and Routledge; and a monograph with Bookman.

Translation Norms and Styles

New Trends in Corpus-Based Translator's Style Studies

Libo Huang

Abstract Since (Baker, M. 2000. Towards a methodology for investigating the style of a literary translator. *Target* 12(2): 241–266) makes the proposal of investigating translator's style with the support of corpus, there has been a variety of discussions on this topic. Methodologically, the previous research can be categorised into two types: S-type (source text type) and T-type (target text type) translator's style investigations based on parallel and comparable corpora, respectively. The former refers to the regularities manifested in the distinctive strategies adopted by a translator in coping with specific source language phenomena in a single translated text, while the latter focuses on the habitual linguistic behaviour of individual translators in all of his or her translations. Nevertheless, some problems still remain unsolved. On the one hand, the studies based on the parallel model decide the translator's style with the evidence from one individual text, the results of which are not so convincing. On the other hand, the studies based on the comparable model rely chiefly on the formal parameters derived from the studies of translation universals, which makes the findings more similar to those of translation universals. An overview of the studies of translator's style in recent years shows that there are some new trends in this field: firstly, some sub-topics, such as interpreter's style, group translators' style, self-translating style, and the diachronic variation in translating style of specific translator, are more noticed; secondly, some new perspectives are being explored including sociological, quantitative linguistic, and multi-dimensional/factorial perspectives. The issues to be tackled in this area include the following: firstly, a more comprehensive methodology has to be designed to replace the one relying on the studies of translation universals; secondly, more attention could be attached to interdisciplinarity in terms of description and interpretation of the findings.

Keywords Translator's style · Corpus · New trends

L. Huang (✉)
Xi'An International Studies University, Xi'An, Shaanxi, China

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_3

1 Introduction

Translation Studies, generally speaking, focuses on transferring the meaning (or content) and style (or form) from the source text to the target text. While the two aspects are inseparable in the same text of one specific language, they often conflict with each other in the interlingual transferring process. In other words, either of the two has to be sacrificed to some extent for retaining the other in translation in accordance with the source text genre. Moreover, the content is more likely to be put in the first place by translators. For instance, Nida and Taber (1969: 12) maintain that "translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style". In comparison with meaning, style is supposed to be in a secondary position. Nida's proposal is based on his practice of Bible translation which attaches more importance to the content of the authoritative source text. According to Boase-Beier (2006: 5), style in translation can be approached from four aspects: "the style of the source text as an expression of its author's choices; the style of the source text in its effects on the reader (and on the translator as reader); the style of the target text as an expression of choices made by its author (who is the translator); the style of the target text in its effects on the reader". It indicates while the style of the original text is ascribed to the original author, the style of translated text should be ascribed to the translator's linguistic choices. One part of the linguistic choices made by the translator is his or her conscious responses to the original text, while the other part results from the translator's subconscious linguistic options. The latter is what Baker (2000) discusses in her paper.

Since Baker (2000) made the proposal of a corpus-based methodology for investigating the style of a literary translator, a large number of scholars have shown great interests in the topic—translator's style. In recent years, the study of translator's style actually hit a bottleneck due to ambiguous definition as well as the flawed methodology. Nevertheless, some new trends are witnessed in the past decade or so and manifested, in particular, in the development of new sub-topics and the further improvement of the methodology. The present paper attempts to discuss the future trend in the field of translator's style studies in terms of both research topics/perspectives and methodological issues.

2 A Review and the State of the Art

2.1 *The Pre-Corpus Period*

Translator's style, as a matter of fact, is not a newly developed technical term or research topic in the field of Translation Studies. It had been discussed long before in a series of literature. In the early discussion of translator's style, there are not only definitions but also debates about whether a translator should have his or her own style and the relationship between the translator's style and the original author's style.

Among the 12 translation principles put forward by Savory (1957), the fifth and the sixth are as follows:

> A translation should reflect the style of the original.
>
> A translation should possess the style of the translator. (See Venuti 2000: 393).

Concerning the two potential principles, at least two questions can be raised: (1) Should a translator have his or her peculiar style which is different from that of the original text or the author? (2) Is there any conflict between translator's style and the original author's style? The two questions are partly ethical in nature. In translation practice, however, they are the dilemma confronting translators. In response to the belief that translator's style denies the style of the original text, Liu (劉隆惠 1961) holds that translator's style will not affect the conveyance of the original text style in translation. Yuan (袁洪庚 1988: 111–113) maintains that translator's style can be independent of the original text and a smart translator can transfer various authors' styles in his or her different translations; nevertheless, translator's style should be subordinate to author's style and a translator should highlight the author's style as much as possible apart from displaying his or her own one. Zhang (張今 1987: 93–94) advocates that a translator should have his or her own translating style; otherwise, the translated works can never be assimilated as a part of the native literature canons in the target culture. As for the relationship between translator's style and author's style, Zhang (ibid.) believes that author's style should be conveyed through translator's peculiar style, while translator's style should be incorporated into the author's style. Hermans (1996: 27–28) puts forward the notion of "translator's voice" which is regarded as "an index of the Translator's discursive presence" and may present itself in the forms of some paratextual information such as translator's notes, prefaces, and postscripts.

The above-mentioned discussions about translator's style apparently are confined to literary translations. They belong to the traditional approach to Translation Studies and are prescriptive in nature. Nevertheless, they not only notice the phenomenon but also have already demonstrated a profound knowledge about it. For instance, Zhang (張今 1987: 94) claims: "if we observe a specific translation in isolation, it seems that we can only detect the author's style rather than the translator's style; but if we examine a couple of translations by the same translator at the same time, we will discern not only the author's style but the style of the translator." This idea of translator's style is almost identical with Baker (2000: 245) in her conception of the study of translator's "preferred or recurring patterns of linguistic behaviour, rather than individual or one-off instances of intervention".

During the pre-corpus period of Translation Studies, it was acknowledged by a number of scholars that translator's style does exist. Researchers embarked on the studies of translator's style which is more prescriptive in nature and more source text-oriented. In addition, there was a lack of quantitative investigations.

2.2 The Corpus-Based Period

Baker (2000) makes the proposal of a corpus-based investigation into translator's style. The idea has its origin in the application of Stylometry in authorship attribution. Stylometry is a linguistic discipline which evaluates an author's style based on statistical methods (Holmes 1998), with its applications ranging from authorship attribution to author profiling, forensic issues, author clustering, and so on (Savoy 2020: 9–17). Drawing on Stylometry, quantitative descriptive analysis is applied in differentiating one translator from another in terms of the patterns shown in their linguistic behaviours.

Methodologically, corpus-based studies into translator's style can be categorised into two types: T-type (target text type) and S-type (source text type). T-type studies (e.g. Baker 2000; Olohan 2004; Saldanha 2011a, b) are based on comparable corpora and focus on the habitual linguistic behaviour in many translations by an individual translator. Baker (2000) pioneers in employing a comparable model in which a comparison is made between translations by one specific translator and translations by another translator as a reference. Later on, more empirical studies are carried out in this comparable mode. The parameters for analysis include type-token ratio, mean word length, mean sentence length, lexical density, use of specific word forms (e.g. high-frequency words, function words, foreign words, and italics), etc.

S-type studies (e.g. Bosseaux 2001, 2004, 2007; Winters 2004a, b, 2007, 2009; Hu 胡開寶 2011: 116–121) are based on parallel corpora and take source text features into consideration. In the parallel model, a parallel corpus consisting of one source text and its different translations in another language is employed as the source of data. Within this model, both interlingual comparison between the source text and its corresponding translations and intralingual comparisons between the translated texts can be made. The target of this type of research lies in the detection of different patterns in various renderings of specific linguistic patterns in the source text by different translators. The source text is employed as a criterion in evaluating different translating styles. Besides, the translations under discussion can be compared both synchronically and diachronically: in synchronic comparison, with the time period being a constant, the research findings mainly reveal the variations between different translators whereas in diachronic comparison, the historical period may be one of the factors leading to the differences between individual translators in their styles.

Nevertheless, given the complexity of translator's style, both the comparable and parallel models appear inadequate to ensure systemic and comprehensive explorations into the issue. On the one hand, translator's style may manifest itself in different dimensions, such as statistical, linguistic, and narrative dimensions; attaching too much importance to statistical parameters, the comparable model cannot reveal all the features of a translator's style. On the other hand, translator's style should be consistent in many translations by the same translator; deciding on a translator's style with the evidence from one individual text, the parallel model does not seem to produce convincing findings. Those are the specific problems to be solved within the comparable and the parallel models.

To solve this problem, it is better to combine corpus-based approach with corpus-driven and corpus-assisted approaches[1] rather than depend merely on the statistics provided by the corpus-based approach. The triangulation based on both quantitative and qualitative analyses may guarantee the identification of translator's style. Corpus-driven approach, first of all, can help researchers to identify the linguistic patterns in all the translations by the same writer; then the corpus-based approach is employed to test the hypothesis with the help of specific parameters; based on the first two steps, researchers may focus on the text properly and analyse the results of concordances manually so as to summarise the regularities of language use in the translated texts. According to Li (2017: 111), sense-making of the statistics is the key issue in the study of translator's style and it involves "what do these numbers say about the process and product of translation? How did the translation come about the way it did? Why did the translation come out the way it did? What social, cultural and political effects did the translation produce on the TL/TC?".

The above-mentioned literature provides an overview of the translator's style studies in the past decade. Criticism is made and certain problems are pointed out. Recently, some new trends have been manifested in the development of the topic and research perspectives.

3 New Trends in the Study of Translator's Style

In recent years, the research methodology in corpus-based Translation Studies is increasingly characterised by interdisciplinarity. The existing research topics have been further explored. The topic of translator's style is no exception. A large number of review articles about it have been published (e.g. Ran et al. 冉詩洋, 張繼光, 魯偉 2016; Hou and Lian 侯羽 and 廉張軍 2017; Hu and Xie 胡開寶 and 謝麗欣 2017; Lǘ and Wang 呂奇 and 王樹槐 2018, 2019). According to Ran et al. (冉詩洋, 張繼光 and 魯偉 2016), the study of translator's style should attach more importance to the consistency in the concept, identification of indicators, and relevance between case studies. Based on an overview of the status quo in translator's

[1] The differences among the three types of approaches lie in the corpus function: corpus-based approaches are top-down ones where the corpus is used "mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli 2001: 65); corpus-driven approaches are bottom-up ones where the corpus "itself is the data and the patterns in it are noted as a way of expressing regularities (and exceptions) in language" (Baker et al. 2006: 49), making minimal a priori presumptions about linguistic features (Biber 2010: 162); corpus-assisted approaches are characterized by the integration between corpus methods and non-corpus methods, stressing the interplay of intuition, data-observation, and introspection (Taylor 2008: 183).

style research, Hou and Lian (侯羽 and 廉張軍 2017) advocate more genres be involved in the discussion and descriptions be extended to the semantic, pragmatic, and textual levels. Hu and Xie (胡開寶 and 謝麗欣 2017) reiterate the idea put forward by Baker (2000) that translator's style be approached from both linguistic and non-linguistic perspectives such as the translation of culture-specific terms. Based on the database of Web of Science and supported by CiteSpace, Lű and Wang (呂奇 and 王樹槐 2018) make a visualised bibliometric analysis of international studies on translator's style (2002–2016). Lű and Wang (呂奇 and 王樹槐 2019) make a similar analysis of translator's style studies in China based on CNKI database during the 15 years from 2002 to 2016. The two literature reviews indicate, while a shift in the approach from linguistic perspectives to interdisciplinary perspectives is detected, there is a dilemma for this type of study in both research models and research methodologies.

3.1 Research Topics

While Baker (2000) focuses more on the formulation of the methodology for investigating the style of literary translator(s), the methodology under discussion is not confined to the literary genre. As an independent topic, the study of translator's style can be further categorised into some sub-topics such as interpreter's style (or interpretese), group translators' style, self-translating style, and the diachronic shift in translating style of specific translators.

3.1.1 Interpreter's Style

Pan and Hu (潘峰 and 胡開寶 2015: 59) analyse the reasons for the limited investigation of interpreter's style from two aspects. For one thing, the principle of loyalty in interpretating hinders the presence of interpreter's style, so that researchers usually tend to explore the style of the speech maker. For another, more importantly, the under-development of specific corpus of interpretese and technological means constrain the study of interpreter's style.

From the interpreter's perspective, Yagi (2000) proposes a series of parameters for the quantification of the style in simultaneous interpretation including fluency, chunking, delay (also lag or ear-voice span), linear discourse development, etc. Among them, specific indicators for fluency consist of pause, false-start, hesitation, incomplete sentence, extended delays, etc.; chunking refers to the strategy employed by the interpreter to "to divide up TL long stretches of discourse into chunks of manageable size" and it can help researchers to find out the interpreter's reformulation strategy; and delay refers to the time span between the start of the speech maker's chunk and the interpreter's chunk. It is usually held that the longer the time span is, the more time the interpreter has for information restructuring; the investigation into the linear discourse development will reveal the patterns in interpreter's fluency, synchronicity, rhythm, and time handling (Yagi 2000: 522–527). The diversification of parameters for examining the interpreter's style contributes to the improvement of the research methodology in particular for interpreter's style studies.

Van Besien and Meuleman (2008) put forward the notion of "interpreter style" and analyse the stylistic differences between two professional interpreters in their simultaneous interpreting from Dutch into English. The focus is put on both global and local strategies employed by the interpreters. The former category involves presentation (including speed of delivery, lag, diversity of vocabulary, intonation, non-verbal behaviour, etc.), additions (including cohesion, appropriateness repairs, clarifications, identifications, etc.), and omissions (of redundancies, of meta-messages, modulating expressions, irrelevant stretches, etc.), while the latter category consists of transcoding, backtracking, anticipation, and pause (Van Besien and Meuleman 2008: 138). The research findings demonstrate the global strategies are more effective in differentiating the stylistic differences between interpreters who have produced abundant and lean interpreting products, respectively. According to Van Besien and Meuleman (2008), this type of study may shed some light on the exploration of interpreting norms.

3.1.2 Group Translators' Style

When advocating the corpus-based approach in investigating the individual literary translator's style, Baker (2000: 244) also mentions the examination of group translators' style which is rarely noticed then. Wang and Huang (王瑞 and 黄立波 2015) divide the English translations of Jia Pingwa, a contemporary novel writer in China, into two categories on the basis of translation direction: translations into one's mother tongue (also known as direct translation) and translations into a foreign language (also known as inverse translation). A comparison is made between the two categories of texts in terms of formal statistics, textual presentation mode, and translation strategy. The research findings show direct translation and inverse translation present some significant differences in three parameters. According to Huang (黄立波 2018: 79), translators can be categorised into different groups according to various criteria: mother-tongue translators, foreign language translators, and translators in collaboration based on the direction of translation; research-oriented translators, professional translators, and Sinologist translators based on the professional background; independent translators and translators in collaboration based on the mode of translation, etc. With compound corpora of different categories of translations, group translators' style can be approached from different perspectives. Hou and Hu (侯羽 and 胡開寶 2019) make a corpus-based investigation into the collaborative translators' style—the style of Howard Goldblatt and his Chinese wife in their English translations of contemporary Chinese novels. The research findings indicate, compared with independent translations, collaborative translations exhibit some linguistic patterns such as low lexical diversity, more concise use of expressions, shorter sentences, more structural imitation of the ST and more use of the strategy of foreignization, etc.

3.1.3 Self-Translating Style

The study of self-translating style in China mainly focuses on Eileen Chang's self-translation which is more representative. With readability as an indicator of translator's style, Huang (黄立波 2012) explores the translating style of Eileen Chang in her self-translations. The comparison is made between three types of texts including Chang's English self-translations, her original English writings, and English translations of her works by other translators. The focus is on the relationship between Chang's English self-translations and her original English writings and the relationship between her English self-translations and English translations of her works by other translators. The corpus-based study shows Chang's self-translations are very close to her English writings in style and the translations by Kingsbury (who is an expert on Eileen Chang) are stylistically closer to Chang's self-translations in many aspects. Huang (黄立波 2012) makes the proposal that methodologically, the study of translator's style should not be confined to the comparable model or the integration of parallel and comparable models; instead, a multi-complex one, in which all types of texts are compared with each other, is needed to triangulate the research results. Focusing on *Shame, Amah*, one of Eileen Chang's short stories, and its English translations, Li's (黎昌抱 2015) multi-complex comparative study demonstrate that Chang's self-translations and translations by the other translator share some peculiarities. Li (ibid.) points out the interaction between "translating" and "writing" in the two types of translations: in self-translations, while "translating" is in a primary position, "writing" is more prominent; in translations by the other translator, "translating" is dominant.

Based on the corpus of different English translations of Eileen Chang's *Shame, Amah*, Gan (甘媚 2018) focuses on the translating style of Chang's self-translation. The investigation is conducted in terms of lexical diversity, syntactical diversity, and translation strategies. The investigation shows that Chang maintains a rich diversity in her self-translations. Chinese culture-loaded items are mainly literally translated with Chinese pinyin for the purpose of introducing the cultural background information to the target language readers. Zhan (詹潇潇 2018) makes a comparison between Eileen Chang's English self-translations and her English writings and finds that while there is no significant difference between the two categories of texts, Chang uses more short sentences and connectives in her self-translations than in her English writings.

3.1.4 Diachronic Shift in Translating Style of Specific Translators

Translator's style, essentially, is the result of the linguistic choice personally and regularly made by the translator in all his or her translations. To be more specific, it is a kind of regularity consistently maintained in translations and remains stable within a specific period of time. That is to say, the stability is not permanent but relative. Diachronically, the translator's style might present some shifts. In investigating the translating style of Howard Goldblatt in his translations of 17 modern and contemporary Chinese novels from 1979 to 2010, Huang and Zhu (黃立波 and 朱志瑜 2012) detect there is a significant diachronic change in Goldblatt's style in his translations in terms of standardised type/token ratio (STTR) and mean sentence length (MSL). In terms of the STTR, the discrepancy between the lowest one in *Tales of Hulan River* (40.65) and the highest one in *Blood Red Sunset* (47.77) is 7.12. As far as the MSL is concerned, the scope of change is more than 10 words with the longest MSL of 23.38 words in *Tales of Hulan River* and the shortest one of 11.81 words in *Black Snow*. *Tales of Hulan River* was the first Chinese novel translated by Goldblatt in 1979. *Blood Red Sunset* came out in 1995 and *Black Snow* in 1993, which are near the middle part of the overall publication time span of around 30 years. It is therefore inferred that significant changes may also take place in a translator's lexical variety and complexity of syntactic structures over time.

Based on a parallel corpus of *Hong Lou Meng* (*The Dream of the Red Chamber*), Zhang and Liu (張丹丹 and 劉澤權 2014) make an comparison between the first 24 chapters and the last 32 chapters of the English translation of the Chinese classic by Bencraft Joly, a British vice-consulate to Macao in the nineteenth century. It is found that there exist significant differences between the two parts in terms of poetic prosodies, sentence patterns, and contextual appropriateness. The researchers infer that Joly's English version of *Hong Lou Meng* might have been accomplished by different translators. The research applies the methods of translatorship attribution which provides another perspective in the study of diachronic shift in translator's style. With the help of a corpus consisting of *The Old Man and the Sea* by Ernest Hemingway and its six different Chinese translations spanning a time period of about 60 years, Liu and Wang (劉澤權 and 王夢瑤 2018) analyse the stylistic differences among the translations by different translators at different times. The corpus-based description manifests to some extent the diachronic changes in translators in different historic periods.

3.2 *Potential Perspectives*

3.2.1 Sociology of Translation

According to Baker (2000: 258), "identifying linguistic habits and stylistic patterns is not an end in itself: it is only worthwhile if it tells us something about the cultural and ideological positioning of the translator, or of translators in general, or about the cognitive processes and mechanisms that contribute to shaping our translational behavior." That is to say, quantitative description of translator's style based on corpus statistics is only the point of departure. It is more significant to make a reasonable interpretation of the motivation behind the statistics. Baker (2000) makes an interpretation of the differences between Peter Bush and Peter Clark in their translator's style in terms of some extra-textual factors, such as texts selected for translation, the living and language surroundings of translators, and the distance between the source language and the target language in cultural and literary norms. It is a socio-cultural analysis of the motivation behind the stylistic differences. This type of descriptive-explanatory framework confirms the empirical nature of corpus-based study of translator's style.

In interpreting the translator's style, apart from the consideration of translator's obedience to or violation of the translation norms of the source language or target language, we can also focus on the translator's habitus which determines the decision-making of the translator covertly but profoundly. In discussing the application of Bourdieu's notion of "habitus" to Translation Studies, Semioni (1998: 19) indicates there is a type of mapping between the differences in social habitus of a group of people and the differences in choices made by them in a specific field. That is to say, the optional behaviours of individuals are constrained or affected by their social habitus. As far as the study of translator's style is concerned, Semioni (ibid: 21) claims, "a programme of research which could be embarked on profitably in this socio-translational framework is one that would investigate … whether the differential of stylistic choices distinguishing different translators can be shown to be a function of the differences in the specialized habitus." In other words, translator's style is closely related to translatorial habitus and the former is the behaviouristic externalisation of the latter. Although Semioni's (1998) interpretation of translator's style is 2 years earlier than the proposal made by Baker (2000), they can arrive at a consensus in the notion of Descriptive Translation Studies.

3.2.2 Quantitative Linguistic Analysis

Quantitative linguistic methods are also employed by some researchers in examining the phenomenon of translator's style. In quantitative linguistics, mathematical quantitative methods are adopted to analyse "various linguistic phenomena, linguistic structures, nature of those structures, and the relationship between them", and "precise measurement, observation, imitation, modeling and interpretation are made to find out the mathematic regularities behind the linguistic phenomena so as to reveal the inherent causes for formation of those phenomena and explore the self-adaptation mechanism of linguistic system and the motivation for language evolution" (Liu and Huang 劉海濤 and 黃偉 2012: 179). According to this conception, language can be studied in a mathematical way. Actually, this tradition dates back to Herman's (1966) idea of "language as choice and chance". The advantage of this approach lies in the quantitative description of linguistic features and the relationship between them. Dong (董琇 2014) proposes a three-sphere model, consisting of the target readers' familiarised style, the original author's writing style, and the translator's writing style, for investigating the specific factors which affect the translation style. Based on the dimension-reduction methods of subjective categorization and principal component analysis, Dong (ibid.) provides a quantitative description of the translator's style of Pearl S. Buck in her translation of *Shui Hu Zhuan* (as *All Men Are Brothers* in English), a Chinese classic, compared with five categories of comparable texts. Mathematical modelling is employed in this study, which is an attempt to apply quantitative linguistic methods to the exploration of translator's style.

Based on a corpus consisting of two Chinese translations of *Pride and Prejudice*, Zhan and Jiang (詹菊紅 and 蔣躍 2016) explore the topic of translatorship attribution by dividing the corpus into two sub-corpora, respectively: training translation texts (TTT1 and TTT2) and experimental translation texts (ETT1 and ETT2) whose translators were assumed unknown. Among the 14 linguistic properties used for investigation, five of them are proved to be effective in differentiating the translations by two different translators. The research has provided some direct insight into the inquiry of translator's style. Zhan and Jiang (詹菊紅 and 蔣躍 2017) make use of the learning model of support vector machine (SVM) to differentiate distinct translations or different translators' styles, and the research findings suggest that SVM classifier is highly effective in discriminating translation styles.

3.2.3 Multi-Dimensional/Factorial Perspectives Analysis

One of the problems in the methodology for investigating translator's style lies in the isolatedness of each individual parameter in specific research. In other words, it is not sensible or convincing to determine the significant differences between two translators in their translating styles only in terms of a couple of parameters such as standardised type-token ratio, mean sentence length, use of contracted forms, italics, and foreign words. In recent years, a significant development in this area is the synthesis of the parameters for investigation which aims at the identification of the dominant factors in shaping the translator's style and the minor ones affecting it. De Sutter, Lefer and Delaere (2017: 1) advocate, in corpus-based Translation Studies, multivariate statistical techniques including multi-dimensional scaling, hierarchical cluster analysis, mixed-effect models, etc. be employed "to visualize, describe, explain and predict patterns of variation within translations and between translations and non-translations". This type of investigation will be able to "find out which factors simultaneously affect linguistic behavior in translations compared to non-translations" (ibid.). With regard to the study of translator's style, the ultimate goal of the application of multi-dimensional/factorial analysis is to detect the various causes or motivations, linguistic or non-linguistic, behind the language choices made by different translators. Han et al. (韓紅建, 蔣躍 and 袁小陸 2019) indicate, against the background of "big data", corpus-based studies of translator's style should adopt the notions of systematicity, holisticness, relevance, and the linguistic outlook which are unique to quantitative linguistics so as to ensure a shift from the static and causal approach to a dynamic and comprehensive one and a shift from the local style to the global style of the translated texts. It is also proposed that methodologies from such areas as data mining, computational linguistics, quantitative linguistics, and stylometry studies be employed and parameters for investigation including entropy, repeat rate, syntactic dependence, affective traces, effectiveness of affective expressions translation, etc. be incorporated in the studies of translator's style (ibid.).

The advantage of the multi-dimensional/factorial analysis lies in the establishment of a link between linguistic analysis and socio-cultural studies in the field of corpus-based Translation Studies.

4 Concluding Remarks

Since Baker (2000) makes the proposal of a corpus-based approach to the study of translator's style, this research topic has been explored persistently with a variety of modified methods. Although the existing methodology is far from being perfect, the set of conceptual tools facilitate our understanding of the nature of translation. Researchers in the future are to be confronted with a series of missions: firstly, the considerable reliance on the methods for investigating translation universals should be reduced to a certain extent and an independent comprehensive methodology designed particularly for the study of translator's style could be formulated, so as to further improve the research rationale, object of study, scope, contents, approaches and so on; secondly, based on the principle of relevance, more multi- or interdisciplinary approaches could be devised in the hope of integrating linguistic explorations with extra-linguistic studies such as socio-cultural, historical, and ideological studies, which will lead to more sensible interpretations from various perspectives; last but not least, all the approaches and methodologies proposed should serve the primary purpose of Translation Studies proper and avoid plausible statistics and complicated diagrams.

References

Baker, M. 2000. Towards a methodology for investigating the style of a literary translator. *Target* 12 (2): 241–266.

Baker, P., A. Hardie, and T. McEnery. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Biber, D. 2010. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford handbook of linguistic analysis*, ed. B. Heine and H. Narrog, 159–191. New York: Oxford University Press.

Boase-Beier, J. 2006. *Stylistic approaches to translation*. Manchester: St. Jerome.

Bosseaux, C. 2001. A study of the translator's voice and style in the French translation of Virginia Woolf's. In *The waves CTIS occasional papers 1*, ed. M. Olohan, 55–75. Manchester: Centre for Translation and Intercultural Studies, UMIST.

Bosseaux, C. 2004. Point of view in translation: A corpus-based study of French translations of Virginia Woolf's to the lighthouse. *Across Languages and Cultures* 1: 107–122.

Bosseaux, C. 2007. *How does it feel? Point of view in translation: The case of Virginia Woolf into French*. Amsterdam: Rodopi.

De Sutter, G., M. Lefer, and I. Delaere, eds. 2017. *Empirical translation studies: New methodological and theoretical traditions*. Berlin: De Gruyter.

Hermans, T. 1996. The translator's voice in translated narrative. *Target* 8: 23–48.

Holmes, D.I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing* 13 (3): 111–117.

Li, D. 2017. Translator style—A corpus-assisted approach. In *Corpus methodologies explained: An empirical approach to translation studies*, ed. M. Ji, M. Oakes, D. Li, and L. Hareide, 103–136. London: Routledge.

Nida, E.A., and C. Taber. 1969. *The theory and practice of translation*. Leiden: E. J. Brill.

Olohan, M. 2004. *Introducing corpora in translation studies*. London: Routledge.

Saldanha, G. 2011. Style of translation: The use of foreign words in translations by Margaret Jull Costa and Peter Bush. In *Corpus-based translation studies: Research and applications*, ed. A. Kruger, K. Wallmach, and J. Munday, 237–258. London and New York: Continuum.

Saldanha, G. 2011. Translator style: Methodological considerations. *The Translator* 1: 25–50.

Savory, T.H. 1957. *The art of translation*. London: Jonathan Cape.

Savoy, J. 2020. *Machine learning methods for stylometry: Authorship attribution and author profiling*. Cham: Springer International Publishing.

Semioni, D. 1998. The pivotal status of the translator's habitus. *Target* 10 (1): 1–40.

Taylor, C. 2008. What is corpus linguistics? What the data says. *ICAME Journal* 32: 179–200.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Van Besien, F., and C. Meuleman. 2008. Style differences among simultaneous interpreters: A pilot study. *The Translator* 14 (1): 135–155.

Venuti, L., ed. 2000. *Translation studies reader*. London: Routledge.

Winters, M. 2004a. German translations of F. Scott Fitzgerald's *The Beautiful and Damned*: A corpus-based study of modal particles as features of translators' style. In *Using corpora and databases in translation*, ed. I. Kemble, 71–88. Portsmouth: University of Portsmouth.

Winters, M.F. 2004b. Scott Fitzgerald's *Die Schönen und Verdammten*: A corpus-based study of loan words and code switches as features of translators' style. *Language Matters, Studies in the Languages of Africa* 1: 248–258.

Winters, M.F. 2007. Scott Fitzgerald's *Die Schönen und Verdammten*: A corpus-based study of speech-act report verbs as a feature of translators' style. *Meta* 3: 412–425.

Winters, M. 2009. Modal particles explained: How modal particles creep into translations and reveal translators' styles. *Target* 1: 74–97.

Yagi, S.M. 2000. Studying style in simultaneous interpretation. *Meta* 45 (3): 520–547.

References in Chinese

Dong, X. 董琇. 2014. "基於降維法的譯者風格研究 (A study of the translator's style based on the dimension-reduction methods)." 外語教學與研究 (*Foreign Language Teaching and Research*) 46(2): 282–293.

Gan, M. 甘媚. 2018. "基於語料庫的張愛玲自譯風格研究 (A corpus-based study of Eileen Chang's style of self-translation)." 未發表的碩士論文 (Unpublished M.A. thesis), 華南理工大學 (South China Universtiy of Technology).

Han, H., Y. Jiang, and X. Yuan. 韓紅建, 蔣躍, 袁小陸. 2019. "大數據時代的語料庫譯者風格研究 (Corpus-based study of translator's style in the Big Data Era)." 外語教學 (*Foreign Language Education*) 40(2): 88–93.

Huang, L. 黃立波. 2012. "張愛玲譯者風格的語料庫考察 (A corpus-based study of Eileen Chang's translator's style)." 全國第二屆語料庫翻譯學研討會宣讀論文 (Paper presented at the 2nd national symposium on Corpus Translation Studies), 曲阜師範大學 (Qufu Normal University), 山東曲阜 (Qufu, Shandong), 31st Mar–1st Apr 2012.

Huang, L. 黃立波. 2018. “語料庫譯者風格研究反思 (Reflections on corpus-based studies of translator’s style).” 外語教學 (*Foreign Language Education*) 39(1): 77–81.

Huang, L., and Z. Zhu. 黃立波, 朱志瑜. 2012. “譯者風格的語料庫考察——以葛浩文英譯現當代小說為例 (A corpus-based study of translator’s style—A case study of Goldblatt’s English translations of Chinese modern and contemporary fictions).” 外語研究 (*Foreign Languages Research*) (5): 64–71.

Hou, Y., and K. Hu. 侯羽, 胡開寶. 2019. 基於語料庫的葛浩文夫婦合譯風格分析——以劉震雲小說英譯本為例 (A corpus-based study of co-translation style of Howard Goldblatt and Lin Li-chun: With reference to their English version of Liu Zhenyun’s novels). 燕山大學學報(哲學社會科學版) (*Journal of Yanshan University (Philosophy and Social Science Edition)*) 20(1): 32–41.

Hou, Y., and Z. Lian. 侯羽, 廉張軍. 2017. 國內外語料庫譯者風格研究現狀分析 (Review of contemporary corpus-based translator style research at home and abroad). 燕山大學學報(哲學社會科學版) (*Journal of Yanshan University (Philosophy and Social Science Edition)*) 18(6): 63–70.

Hu, K. 胡開寶. 2011. “語料庫翻譯學概論 (*A general introduction to Corpus Translation Studies*).” 上海 (Shanghai): 上海交通大學出版社 (Shanghai Jiaotong University Press).

Hu, K., and L. Xie. 胡開寶, 謝麗欣. 2017. “基於語料庫的譯者風格研究: 內涵與路徑 (Towards a corpus-based study of translator’s style).” 中國翻譯 (*Chinese Translators Journal*) 38(2): 12–18.

Li, C. 黎昌抱. 2015. “基於語料庫的自譯與他譯比較研究 (A Corpus-based comparative study between self-translation and conventional translation).” 外國語 (*Journal of Foreign Languages*) 38(2): 57–64.

Liu, H., and W. Huang. 劉海濤, 黃偉. 2012. “計量語言學的現狀、理論與方法 (Quantitative Linguistics: State of the art, theories and methods).” 浙江大學學報(人文社會科學版) (*Journal of Zhejiang University (Humanities and Social Sciences)*) (2): 178–192.

Liu, L. 劉隆惠. 1961. “談談文藝作品風格的翻譯問題 (On the translation of style in literary works).” 學術月刊 (*Academic Monthly*) (11): 51–54.

Liu, Z., and M. Wang. 劉澤權, 王夢瑤. 2018.《老人與海》六譯本的對比分析——基於名著重譯視角的考察 (A contrastive analysis of six Chinese translations of The old man and the sea—from the perspective of the re-translation of masterpieces). 中國翻譯 (*Chinese Translators Journal*) 39(6): 86–90.

Lű, Q., and S. Wang. 呂奇, 王樹槐. 2018. 國際譯者風格研究可視化文獻計量分析 (2002–2016) (A visualized bibliometric analysis of international studies on translator’s style (2002–2016)). 外語學刊 (*Foreign Language Research*) (2): 82–89.

Lű, Q., and S. Wang. 呂奇, 王樹槐. 2019. “國內語料庫譯者風格研究十五年(2002–2016)——CiteSpace輔助的可視化文獻計量分析 (A visualized bibliometric analysis of corpus-assisted translator’s style studies in China (2002–2016)).” 燕山大學學報(哲學社會科學版) (*Journal of Yanshan University (Philosophy and Social Science Edition)*) 20(1): 42–49.

Pan, F., and K. Hu. 潘峰、胡開寶. 2015. “語料庫口譯研究: 問題與前景 (Corpus-based Interpreting Studies: Problems and prospects).” 語言與翻譯 (*Language and Translation*) (2): 55–61.

Ran, S., J. Zhang, and W. Lu. 冉詩洋, 張繼光, 魯偉. 2016. “國內譯者風格研究的CiteSpace科學知識圖譜分析 (A scientific CiteSpace analysis of the research on translators’ style in China).” 外國語文 (*Foreign Language and Literature*) 32(4): 133–137.

Wang, R., and L. Huang. 王瑞, 黃立波. 2015. “賈平凹小說譯入與譯出風格的語料庫考察 (A corpus-based investigation of the style between direct and inverse translations of Jia Pingwa’s novels).” 中國外語 (*Foreign Languages in China*) 12(4): 97–105.

Yuan, H. 袁洪庚. 1988. “試論文學翻譯中的作者風格與譯者風格 (On author’s style and translator’s style in literary translation).” 蘭州大學學報(社会科学版) (*Journal of Lanzhou University (Social Sciences)*) (2): 109–116.

Zhan, J., and Y. Jiang. 詹菊紅, 蔣躍. 2016. "語言計量特徵在譯者身份判定中的應用——以《傲慢與偏見》的兩個譯本為例 (Application of Quantitative Linguistic Properties to Translatorship Recognition)." 外語學刊 (*Foreign Language Research*) (3): 95–101.

Zhan, J., and Y. Jiang. 詹菊紅, 蔣躍. 2017. 機器學習算法在翻譯風格研究中的應用 (On Using Machine Learning Methods to Discriminate Translation Styles). 外語教學 (*Foreign Language Education*) 38(5): 80–85.

Zhan, X. 詹瀟瀟. 2018. "基於語料庫的張愛玲自譯風格研究 (A corpus-based stylistic study on Eileen Chang's English self-translations)." 未發表的碩士論文 (Unpublished M.A. thesis), 北京外國語大學 (Beijing Foreign Studies University).

Zhang, D., and Z. Liu. 張丹丹, 劉澤權. 2014. 《紅樓夢》喬利譯本是一人所為否?——基於語料庫的譯者風格考察 (Is this English translation of Hong Lou Meng by Joly himself?—A corpus-based investigation of translator style). 中國外語 (*Foreign Languages in China*) 11(1): 85–93.

Zhang, J. 張今. 1987. "文學翻譯原理 (Principles of literary translation)." 開封 (Kaifeng): 河南大學出版社 (Henan University Press).

Libo Huang is a professor of Translation Studies at the Foreign Language and Literature Institute, Xi'an International Studies University, Xi'an, China. He obtained his Ph.D. at the National Research Center for Foreign Language Education, Beijing Foreign Studies University, and worked as a postdoctoral fellow at the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University from 2010 to 2012. From 2017 to 2018, he worked as a Fulbright Visiting Scholar at the English Department of UC Berkeley in the US. His research interests include translation studies, corpus linguistics, and cultural history of translation. In recent years, his focus of research is mainly on the parallel corpus-based studies of styles in translation, English translations of modern and contemporary Chinese novels, English translations of Chairman Mao Zedong's works, and traditional Chinese translation theories.

Building a New-Generation Corpus for Empirical Translation Studies: The Dutch Parallel Corpus 2.0

Ryan Reynaert, Lieve Macken, Arda Tezcan, and Gert De Sutter

Abstract This chapter introduces a new, updated version of the Dutch Parallel Corpus, a bidirectional parallel corpus of expert translations for Dutch><English and Dutch><French language pairs. This revisited version of the corpus, which we dub Dutch Parallel Corpus 2.0, is dynamic in nature, and contains 2.75 million words at the time of writing. The corpus is sentence-aligned, lemmatized and POS-tagged using the state-of-the-art natural language processing toolkit Stanza. Compared to its predecessor, the Dutch Parallel Corpus 2.0 contains more metadata about the translators (e.g. gender, education, experience) and the translation projects (e.g. L1/L2 translation, software used, degree and type of revision), next to the traditional metadata about the texts themselves (e.g. source and target language, intended audience, intended goal, register). The availability of an extensive set of metadata is considered the main asset of this corpus, together with a more principled and flexible register classification, thus stimulating corpus-based translation scholars to answer more refined research questions about the linguistic and contextual factors that shape translated texts, and ultimately fostering ideas and theories about the social and cognitive processes involved in translation performance. The corpus is freely available for research purposes via https://www.dpc2.ugent.be/.

R. Reynaert · G. De Sutter (✉)
Empirical and Quantitative Translation and Interpreting Studies (EQTIS),
Department of Translation, Interpreting and Communication, Ghent University,
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
e-mail: Gert.DeSutter@ugent.be

R. Reynaert
e-mail: Ryan.Reynaert@ugent.be

L. Macken · A. Tezcan
Language and Translation Technology Team (LT3), Department of Translation,
Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45,
9000 Ghent, Belgium
e-mail: Lieve.Macken@ugent.be

A. Tezcan
e-mail: Arda.Tezcan@ugent.be

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_4

Keywords Parallel corpus · Corpus compilation · Linguistic annotation · Stanza · Universal PoS · Rich metadata · Register classification

1 Introduction

Over the last 30 years, the availability of linguistic data collected in parallel corpora has propelled the empirical study of translated texts in relation to their source texts and comparable texts in the target language, thereby gradually uncovering the complex linguistic identity of translated texts as communicative products produced in specific sociocognitive circumstances. Following Kruger and van Rooy (2016) and Kotze (2020), these (varying) circumstances can be characterized along five dimensions: language activation (monolingual vs. bilingual), modality (spoken, written vs. multimodal) and register, text production (mediated vs. unmediated), proficiency (proficient vs. learner) and task expertise (expert vs. non-expert). Taken together, these circumstances shape the specific linguistic make-up of translations, which in turn shows traces of these social and cognitive circumstances. It goes without saying that the availability of translational corpora, which dates back to Mona Baker's ground-breaking work in the 1990s (1993, 1996, 2004), has contributed significantly to a better understanding of the specific linguistic features of translated text in comparison to their source texts and comparable non-translated texts, in that it has enabled fine-grained comparative descriptive research (see collective volumes and monographs for an overview, e.g. Laviosa 2002; Olohan 2004, Oakes and Ji 2012; Fantinuoli and Zanettin 2015; Xiao and Hu 2015; Corpas Pastor and Seghiri 2016; Ji 2016; De Sutter et al. 2017; Malamatidou 2018; Vandevoorde et al. 2020; see also De Sutter and Lefer 2019 for a critical overview). Moreover, corpora have also stimulated the use of a wide range of descriptive and inferential statistics, thereby elucidating and validating the patterns found in corpora (see e.g. Oakes and Ji 2012; Mellinger and Hanson 2016; De Sutter et al. 2017). The subtle patterns that emerge from such multivariate corpus-based research is now also slowly developing in more encompassing theoretical models, such as Halverson's *gravitational pull* model (2017), which is a (partial) cognitive-linguistic theory that models the selection of meanings, words and constructions in a translational context (Halverson 2013, 2017), or the constrained-communication model, which aims at understanding the different dimensions that affect translated language use and tries to capture the similarities and differences of different types of constrained language varieties, such as written translations, audiovisual translation, interpreting, second-language learner production, etc. (Kruger and van Rooy 2016; Kotze 2020).

Despite the obvious advances linguistic corpora have brought, it has also become clear in recent years that the traditional corpora which were developed in the 1990s and early 2000s are not fit anymore to answer the more specific research questions that need to be addressed now in order to get an extended, more accurate understanding of the (cognitive and social) mechanisms that shape the language used in translated texts. Rather than sheer corpus size, corpus-based translation studies in particular needs more qualitative data, i.e. every (translated) text in a corpus should be accompanied by meta-information about the exact circumstances under which texts and translations were produced: 'corpora will need to be much more carefully designed to take consideration of translators' backgrounds and the circumstances of text production' (Kotze 2020: 356; see also Halverson 2015; De Sutter et al. 2012; De Sutter and Lefer 2019; Lefer 2020). The current chapter responds to this invitation to design new-generation corpora by introducing the new version of the Dutch Parallel Corpus, which we dub Dutch Parallel Corpus 2.0. This new corpus adopts the main compilation and design principles of its predecessor, which was released in 2010 (Macken et al. 2011): it is a sentence-aligned, linguistically enriched and stylistically and regionally stratified bidirectional parallel corpus of Dutch, English and French, with Dutch as the central language. For the Dutch Parallel Corpus 2.0, new source texts and translations were collected, as well as considerably more metadata which characterize translation, translator, translational context and translational task more accurately. Furthermore, a new (flexible) register classification procedure was developed. All translations in the corpus were produced by professional translators who mainly work for (Belgian) media institutions, governmental institutions and publishers, among others.

In the remainder of this chapter, the Dutch Parallel Corpus 2.0 (DPC 2.0) is presented in detail, including the challenges we encountered while collecting and processing the data and the ensuing limitations of the present corpus. Section 2 is devoted to the data collection process, Sect. 3 discusses the way the data was processed and Sect. 4 presents the metadata we collected for each of the translations. Finally, Sects. 5 and 6 discuss the new (flexible) register classification system that was developed for our corpus, as well as the full potential of its exploitation.

2 Data Collection

DPC 2.0 was compiled using the same design principles as its predecessor while at the same time solving the main problems of the first Dutch Parallel Corpus (Macken et al. 2011). This means that we collected (interlingual) translations and their source texts from French into Dutch and vice versa, and from English into Dutch and vice versa. We focused on the Belgian–Dutch translation market and aimed for a large variety of text types. Texts were included in DPC 2.0 if the text providers and translators were willing to provide us with a (broad) set of metadata concerning the source and target texts, the translator(s), the translation project and its context. Additionally, source texts had to be *proper* source texts, i.e. not translated from yet another source text. We invited all text providers who contributed to the first DPC and invited new text providers. Although the compilation of DPC 2.0 is still

Table 1 General data overview per translation direction in DPC 2.0

Translation direction	Word count[1]	Number of source texts	Number of text providers	Number of translators[22]
English > Dutch	398,774	110	10	13 (10)
Dutch > English	430,094	105	20	22 (16)
French > Dutch	1,029,739	153	15	20 (14)
Dutch > French	925,002	176	20	22 (19)
Total	**2,783,609**	**544**	**65**	**77 (59)**

ongoing and new texts are continuously being added, the corpus comprises approximately 2.75 million words at the time of writing.

The preliminary overview presented in Table 1 demonstrates that significantly more Dutch >< French texts are included in the corpus. Taking into account that both French and Dutch constitute the two major official languages in Belgium, translations in both directions are more frequently required for this language pair, particularly in institutional contexts. English translations were nevertheless easily obtained from companies which use English for commercial purposes.

Obtaining data from different text providers obviously poses several challenges for corpus compilers. A lack of financial compensation for the provision of data, for instance, in many cases resulted in a restricted amount of texts. Moreover, the acquisition of metadata related to the global translational context depended on translator's willingness to collaborate and caused an extra hurdle to overcome in the negotiations with text providers and/or translators themselves. In the case of companies relying on freelance translators, each translator had to be contacted separately in order to obtain additional information. Negotiations with in-house translation departments, on the other hand, were usually held with the head of the translation department, and conveniently enabled us to receive completed questionnaires of multiple employees at once. In the end, both in-house and freelance translators were usually inclined to provide us with additional information on the translation process and, hitherto, resulted in the representation of a myriad of translator profiles.

An additional issue was encountered when it came to obtaining permission for the non-commercial use of each individual text. In line with De Clercq and Montero Perez (2010), we confirm that negotiations concerning copyright clearance are known 'to drag on for months and exceptionally even years' (p. 3384). In fact, the

[1] The word count was calculated after the initial cleaning process of all texts (cf. Section 3). The eventual word count may deviate somewhat from this preliminary calculation.

[2] Translation agencies which were not fully able to provide us with specific details on their employees were counted as a single translator, although more translators may have been involved in the translation process. This is clearly marked in the corpus. The numbers between parentheses refer to the amount of individual translators whose profile could be determined on the basis of *all* available metadata.

recent introduction of the EU-regulation 2016/679 concerning general data protection (GDPR) seemed to complicate the negotiation process even further. On top of an already existing preoccupation with the eventual goal(s) of the corpus and its (online) access, text providers and translators expressed their concerns with regard to the anonymization and storage of personal data. This urged us to come up with a specific licence agreement in order to reassure our text providers. With the help of the legal department of Ghent University, a new licence agreement was drafted which underscores the non-commercial aim of DPC 2.0, guarantees restricted access to the corpus and ensures the protection of personal data. This shift from both commercial and non-commercial applications of texts in the original DPC to an exclusively non-commercial research project significantly seemed to lower the threshold for an agreement. Whereas the first DPC required four different types of agreements, recent negotiations only involved the above mentioned, shorter licence agreement and/or an e-mail containing explicit permission of our text providers. With the exception of some major text providers, the shorter version of the licence agreement was considered too formal by other providers and, as a result, permission was mainly given explicitly through mail by a representative of the involved organization.

Acquiring literary texts for inclusion in language corpora is known to be especially problematic. Next to the issues mentioned above, ‘negotiations with publishers are slow and time-consuming’ (Geyken 2007, p. 32), partly because of the amount of parties involved (the original author, the translator and the publishers of both the original book and its translation), but also because commercial publishers fear undue competition, as text production constitutes their core business. Despite the non-commercial aims of DPC 2.0, this preoccupation for unfair competition prevailed with commercial publishers. Direct negotiations with commercial publishers were therefore avoided, and we turned our attention to negotiations with authors and translators. We furthermore decided to broaden our scope and include literary classics which belong to the public domain but are nevertheless still frequently translated.[3] In such cases, permission of the translator and the foreign publisher was sufficient, which significantly reduced the duration of the negotiation processes.

Hitherto, no attempt was made to reintegrate texts which were present in the original DPC. Since these texts were produced more than a decade ago, such attempt would in fact require a retrospective, and thus time-consuming retrieval of metadata. Considering the emerging patterns in DPC 2.0, however, literary texts are still underrepresented, and a limited amount of overlapping literature may therefore be taken into consideration at further stages of the corpus’ development.[4] In such exceptional case(s), metadata can be obtained directly from individual book translators, thereby reducing the effort involved in extending the original metadata files.

[3] Such literary texts were retrieved from Project Gutenberg, an online library of free eBooks: https://www.gutenberg.org/.

[4] In contrast with texts of the original DPC-project, which are primarily outdated a decade after being produced, literary texts remain relevant to a higher extent. As such, they are better suited for reintegration in DPC 2.0.

3 Data Processing

Text providers usually sent us human-readable texts in Word or pdf, containing both textual and non-textual information such as graphs, illustrations and footnotes, among others. In order to facilitate the data-processing of DPC 2.0, each text was first converted to a raw, plain-text format. This partially manual cleaning process resulted in the deletion of non-textual information as well as the conversion of each text to a machine-readable format, more specifically XML and TMX. As soon as a DPC text had been converted to a standardized format, it could be subjected to further data processes involving sentence alignment and linguistic annotation through PoS-tagging and lemmatization.

Although these intermediary steps were equally present in the compilation of the original DPC, 10 years after its completion, the tools and software applied in the data-processing have been surpassed by more innovative alternatives which are presented in the following subsections. For an in-depth comparison with the alignment and annotation process of the original DPC, however, we refer to Macken et al. (2011).

3.1 Sentence Alignment

Sentence alignment constitutes a crucial step in the processing of parallel texts. It makes it possible to establish connections between each sentence in a source text and its equivalent in the corresponding target text. Despite the relevance of aligning parallel texts on a sentence level, Santos (2011) warns that 'splitting a text into sentences is not a trivial task because the formal definition of what is a sentence [sic] is a problem that has eluded linguistic research for quite a while' (p. 124). In the interest of uniformity, we decided to define a sentence in line with the original DPC, thereby focusing on endings with terminal punctuation (a period, question mark or exclamation mark) or paragraph breaks. In doing so, we also leave open the option of reintegrating original DPC-texts into the revisited version at a later stage, although this is not a priority for DPC 2.0.

Unlike the alignments in the original DPC, in which the output of multiple alignment tools[5] was combined and manually checked, text pairs in the new DPC were aligned with a single tool, viz. AlignFactory Light.[6] This alignment tool enables its users to automatically split and align sentences in numerous language pairs, including English, French and Dutch. Additionally, AlignFactory Light comes with an editor which allows for an immediate manual verification. Finally, the tool also provides a possibility of exporting, for instance, TMX-files, thus allowing direct integration into any translation memory.

[5] Vanilla Aligner (Danielsson and Ridings 1997), Geometric Mapping and Alignment tool (Melamed 1997) and Microsoft Bilingual Aligner (Moore 2002).

[6] AlignFactory Light was developed by the software company Terminotix: http://www.terminotix.com/.

Table 2 Example of sentence alignments in DPC 2.0[7]

Source text in Dutch	Alignment type	Target text in English
Bovendien zijn ze vrij toegankelijk voor al wie hun stilte respecteert	1: 1	Feel free to enter these premises, but don't forget to respect their perfect tranquility
Eeuwen geleden werden de godshuizen voor het eerst opgetrokken uit liefdadigheid, vandaag zijn ze met hun pittoreske tuintjes, witgeschilderde gevels en heerlijke stilte dé rustplekken van de stad	1: many	The almshouses were first founded centuries ago for charitable purposes Today, with their picturesque gardens, their white-painted gables and their perfect peace and quiet, they are amongst the most tranquil places in Bruges
Vandaag zijn zowat alle godshuizen gerestaureerd en gemoderniseerd en wonen er nog steeds bejaarden in Met hun pittoreske tuintjes en witgeschilderde gevels zijn het de plekken bij uitstek om even tot rust te komen	**many: 1**	Practically, all of the almshouses have been carefully restored and modernized and offer cosy living to today's elderly, whilst their small yet picturesque gardens and white-painted façades offer welcoming peace and quiet to the present-day visitor
‘Wie de museumshop van het Groeningemuseum binnenstapt, gaat er gegarandeerd buiten met een origineel souvenir Je neemt er je favoriete kunstschatten mee in de vorm van een fraai geïllustreerd boek, een reproductie of een prentkaart	**many: many**	‘Whoever enters the museum shop of the Groeninge Museum will leave with some wonderful memories, that I can assure you Perhaps you will take home your favourite art treasures in the shape of a handsomely illustrated book or a reproduction on a poster maybe, or depicted on a few picture postcards And why don't you surprise yourself with an original souvenir?
(Op de stadsplattegrond staan alle godshuizen aangeduid.)	**1: Ø**	[*no corresponding segment in target text*]
[*no corresponding segment in source text*]	**Ø: 1**	The interviews have been transcribed in full

AlignFactory Light generates highly accurate alignment links between two language pairs. Nevertheless, all alignment links in DPC 2.0 were manually checked and, in the case of inconsistencies, corrected. We found that two alignment categories were particularly error-prone: (i) many-to-many alignments, which are the result of restructured paragraphs in the target-language text and (ii) null alignments, which point towards a deletion or addition in the target-language text. The resulting alignment types correspond to the types identified in Macken et al. (2011). Table 2 provides a comprehensive overview of all possible alignment types in the corpus.

[7] With the exception of the last example (Ø: 1), which was taken from dpc2-img-000453-NL_EN, all other alignment types were extracted from dpc2-vbr-000244-NL_EN, which is a tourist brochure on the city of Bruges.

In DPC 2.0, 81% of the alignment types are one-to-one alignments (1: 1). 8% of all alignments connect a single source-text sentence with at least two segments in the target text (1: many), whereas 6% of the alignments connect at least two source-text sentences with just a single sentence in the translated text (many: 1). Additionally, many-to-many alignments constitute 4% of the retrieved alignment types. Segments which could not be linked to a corresponding segment in the translation (1: Ø) or source text (Ø: 1) were only exceptionally encountered. These null alignments represent just 1% of all alignment types.

3.2 Linguistic Annotation: PoS-Tagging and Lemmatization

To facilitate the linguistic exploration of DPC 2.0, we used Stanza (Qi et al. 2020) to generate Part-of-Speech (PoS) information and additional morphological features for all words and to perform lemmatization. Stanza is a recently introduced state-of-the-art natural language processing toolkit supporting 66 languages. Prior to adding these linguistic annotations, we used Stanza to tokenize the text and perform sentence segmentation. During tokenization, a sentence is split into sequences of words and punctuation marks.

One of the main advantages of using Stanza is that it labels all tokens with their universal PoS tags[8] (Upos) and universal morphological features (Ufeats).[9] The universal set of PoS tags and morphological features are part of the Universal Dependencies framework,[10] whose major objective is to provide a consistent annotation of grammar across different languages to facilitate the development of multilingual part-of-speech taggers and parsers. We believe that the usage of universal PoS tags and morphological features is particularly useful for translation studies and multilingual corpus analysis as it will make it easier to formulate cross-lingual queries.

In Table 3, we give an example of a Dutch source sentence translated into English and French. For each token, the table contains, respectively, its lemma (which is the base form of the word), the universal PoS tag and the universal morphological features. The default Stanza models for Dutch and English also generate the language-specific PoS tags (Xpos), which is for Dutch the CGN part-of-speech tag set (Van Eynde, Zavrel et al. 2000) and for English the Penn Treebank tag set (Marcus, Santorini et al. 1993).[11] These tag sets are identical to the ones used in the original Dutch Parallel Corpus.

[8] https://universaldependencies.org/u/pos/.

[9] https://universaldependencies.org/u/feat/index.html.

[10] https://universaldependencies.org/.

[11] For English, Dutch and French, the following language models were used, respectively: UD_English-EWT, UD_Dutch-Alpino and UD_French-GSD.

Table 3 Example of a tokenized sentence enriched with linguistic annotations in Dutch, English, and French

	Token	Lemma	Upos	Ufeats	Xpos
NL	financiering	financiering	NOUN	Gender = Com|Number = Sing	N|soort|ev|basis|zijd|stan
	werd	worden	AUX	Number = Sing|Tense = Past|VerbForm = Fin	WW|pv|verl|ev
	verkregen	verkrijgen	VERB	VerbForm = Part	WW|vd|vrij|zonder
	via	via	ADP	/	VZ|init
	verschillende	verschillend	ADJ	Degree = Pos	ADJ|prenom|basis|met-e|stan
	bronnen	bron	NOUN	Number = Plur	N|soort|mv|basis
			PUNCT	/	LET
EN	funding	funding	NOUN	Number = Sing	NN
	was	be	AUX	Mood = Ind|Number = Sing|Person = 3|Tense = Past|VerbForm = Fin	VBD
	obtained	obtain	VERB	Tense = Past|VerbForm = Part|Voice = Pass	VBN
	from	from	ADP	/	IN
	various	various	ADJ	Degree = Pos	JJ
	sources	source	NOUN	Number = Plur	NNS
			PUNCT	/	.
FR	un	un	DET	Definite = Ind|Gender = Masc|Number = Sing|PronType = Art	
	financement	financement	NOUN	Gender = Masc|Number = Sing	
	a	avoir	AUX	Mood = Ind|Number = Sing|Person = 3|Tense = Pres|VerbForm = Fin	
	été	être	AUX	Gender = Masc|Number = Sing|Tense = Past|VerbForm = Part	
	obtenu	obtenir	VERB	Gender = Masc|Number = Sing|Tense = Past|VerbForm = Part	
	de	un	DET	Definite = Ind|Gender = Fem|Number = Plur|PronType = Art	
	diverses	divers	DET	Gender = Fem|Number = Plur	
	sources	source	NOUN	Gender = Fem|Number = Plur	
			PUNCT	/	

4 Metadata

In their project overview of the original DPC, Macken et al. (2011) rightfully state that 'rich metadata is an essential prerequisite to the optimal use of any corpus' (p. 6). In the first DPC release, 10 years ago, each original DPC-file was provided with a metadata file including, among others, text goal, translation direction, domain and text type. Over the past decade, numerous corpus-based translation studies (e.g. Halverson 2017; De Sutter and Lefer 2019; Kruger and Van Rooy 2012; Kotze 2020) have continued to advocate the compilation of corpora which include larger amounts of metadata in order to be able to answer more specific research questions concerning the (simultaneous) effect of the translation's profile, the translator's profile and the translation project on linguistic choices in translated texts. The set of metadata that was considered crucial for the new DPC 2.0 is inspired by recent research in (product-oriented) translation studies, more particularly, the constrained-communication framework mentioned above (Kotze 2020). When applying the five dimensions in this framework to the data collected for DPC 2.0, the dimensions *language activation* and *proficiency* are not relevant for the current project, since we only collected *professional* translations (there are no student translations in the corpus) and *interlingual* translations which by definition involve two languages being activated during the translation process. In sum, there is no variation in DPC 2.0 with respect to these dimensions. However, the dimensions *modality and register* (spoken, written, written to be spoken, written transcripts of spoken text), *text production* (self-revision, other-revision) and *task expertise* (domain expert, non-expert) are relevant and are hence incorporated in the list of metadata (see below for a full overview). Additionally, the list of metadata was inspired by the metadata that were collected for the MUST-project (Granger and Lefer 2020), such as *translation directionality*, which distinguishes between L1 and L2 translation. Finally, many suggestions of essential metadata mentioned in empirical studies were a source of inspiration for the metadata included in this new version of the corpus (De Sutter et al. 2012; De Sutter and Lefer 2019). Thus, DPC 2.0 particularly distinguishes itself for its amount of added information related to the translation project and context and the involved translator(s). Sections 4.1 and 4.2 present an overview of all parameters included in the metadata files of DPC 2.0 and give a detailed description of how this information was retrieved.

4.1 Translation- and Translator-Related Metadata

The inclusion of metadata for each translation and translator came with an additional step in the data-gathering process. Table 4 was sent to all text providers or directly to the translators in the form of a questionnaire (see Appendix), providing them with a list of predetermined answers from which they could choose. In some exceptional cases, translators signalled the absence of the most appropriate answer,

Table 4 Overview of translation- and translator-related metadata in DPC 2.0

Translation- and translator-related metadata		Proportions[31]
1. Version	Source text	50%
	Translation	50%
2. Source text language (if translation)	NL	52%
	EN	20%
	FR	28%
3. Pivot language[14]	Yes (which language?)	1%
	No	98%
		unspecified: 1%
4. Translation tool or memory	Yes (which tool or software?)[15]	49%
	No (manual translation)	40%
		unspecified: 11%
5. Post-editing	Yes (which tool or software?)	8%
	No	81%
		unspecified: 11%
6. Collaborative translation	Yes	2%
	No	95%
		unspecified: 3%
7. Translation directionality	L1-translation	82%
	L2 translation	13%
		unspecified: 5%
8. Translator gender	F M	68% 13%
	X	0%
		unspecified: 19%

(continued)

which led us to apply some minor changes to the answer options in the questionnaire.[12] Table 4 provides a comprehensive overview of the responses for each metadatum.

[12] Initially, we did not include translators with a degree in interpreting or occasional translators, for instance.

[13] Translator-specific criteria, such as age or gender, were often left unspecified in the questionnaire, since we regularly obtained a general overview of a translation department instead of a unique questionnaire for each translator.

[14] As mentioned in Sect. 2, source texts had to be *proper* source texts, i.e. not translated from yet another source text. Nevertheless, DPC 2.0 contains four source texts which are translations themselves. In contrast with the original DPC, however, the inclusion of these texts was only accepted when the language of the original source text was known.

[15] CAT-tools that were mentioned by the text providers are MemoQ, SDL Trados Studio, Déjà Vu X3 Professional, XTM and Wordfast. Post-edited texts were generated by either DeepL or Google Translate.

Table 4 (continued)

Translation- and translator-related metadata		Proportions[34]
9. Translator age	20–30	9%
	31–40	46%
	41–50	10%
	51–60	4%
	61–70	6%
		unspecified: 25%
10. Translator experience	None (occasional translator)	1%
	0–5 years	8%
	6–10 years	41%
	11–20 years	16%
	20 + years	9%
		unspecified: 25%
11. Translator degree	No specific language degree	12%
	Translation MA	36%
	Translation BA	4%
	Language and literature	23%
	Interpreting	1%
		unspecified: 24%
12. Translator status	Freelance	55%
	In-house	42%
		unspecified: 3%
13. Revision	Monolingual (only translation)	33%
	Bilingual (source text and translation)	26%
	None	13%
		unspecified: 28%
14. Style guides	In-house guidelines	9%
	In-house glossary	5%
	Both	14%
	None	44%
		unspecified: 28%
15. Domain expertise[16]	Expert	65%
	Non-expert	11%
		unspecified: 25%

Hitherto, DPC 2.0 contains approximately 1100 texts of which over 70% contain complete metadata files. In the exceptional case of collaborative translations, we decided to provide a general overview of the translation process, and excluded specific translator-related information such as translator's age and translator's experience, among others.

[16] With domain expertise, we refer to translators' subjective estimation of their expertise regarding a particular translation task and its topic(s).

Whether or not *all* translation-related metadata are available for a text largely depends on the size of the translation agency or department within a company. A small-sized translation department within a company seemed to enhance the feasibility of the data-gathering. With these companies, we experienced an augmented response ratio, since text providers put us directly into contact with their individual translator(s). This allowed for a more personal approach and an unmediated way of providing instructions on how to fill out the form, including a clear description of the project's aims. On the other hand, as soon as too many translators are employed within a company, the collection of only a few texts per translator quickly became time-consuming and usually caused text providers to abandon the project, even more so because no financial compensation was provided for their participation.

4.2 Text-Related Metadata (Translated Texts and Source Texts)

The second set of metadata characterizes the translated texts and their source texts (see Table 5). This selection was mainly based on the previous version of DPC as well the framework of situational characteristics provided by Biber and Conrad (2009) which, at a further stage, led us to rethink the register classification of our texts (Sect. 5). In contrast with the metadata presented in the previous section, which were determined on the basis of completed questionnaires, no translators were involved in the identification of text-related metadata. Instead, these criteria were attributed by the corpus' main annotators and, at a further stage, by multiple student raters.

The first parameter, *filename*, refers to the unique name of each of the texts; the next two parameters, *language* and *numbers of words/tokens*, further characterize the texts with respect to the language used ((Belgian-)Dutch, (American-)English or (Belgian-)French) and the number of words it contains. As for the *text provider* metadatum, our subdivision into five categories is based on Delaere (2015):

- Commercial companies (e.g. ArcelorMittal, a multinational steel manufacturing corporation)
- Public services (e.g. Unia, the centre for equal opportunities and opposition against racism)
- Public enterprises (e.g. Visit Bruges, a tourist information centre offering both commercial and non-commercial services)
- Media institutions or publishers (e.g. Knack, a weekly news magazine)
- Research and Development (e.g. BIRA, the Belgian federal scientific research institute)

As can be seen, a clear distinction was made between commercial and non-commercial texts. As such, *commercial companies* were defined as private

Table 5 Overview of text-related metadata in DPC 2.0[17]

Text-related metadata		Proportions
1. Filename	dpc2-xxx-000123-nl/en/fr	
2. Language	NL, EN, FR	
3. Number of words/tokens	X	
4. Text provider	Commercial company	25%
	Media institution (or publisher)	31%
	Research & Development	5%
	Public service (non-commercial)	33%
	Public enterprise (commercial)	6%
5. Channel	Written to be read	77%
	Written to be spoken	0.5%
	Written reproduction of speech	7%
	Multimodal	0.5%
		Unspecified: 15%
6. Intended audience	Broad external	74%
	Specialized external (or internal)	26%
	In between specialized and broad external	[*this additional distinction is yet to be integrated in DPC 2.0*]
7. Communicative purpose	To inform	51%
	To instruct	17%
	To persuade	7%
	To form an opinion	22%
	To narrate	3%
8. Topic/ keyword	Corporate, Culture, Economy, Education, History, Law, Leisure, Nature, Politics, Science, Sports, Tourism, Transport	[*multiple topics and/or free topic choice were allowed per text*]
9. Register	Manuals for a general audience	4%
	Manuals for specialists	8%
	(Popular) science	8%
	Journalistic texts	28%
	Commercial communication	20%
	Public service communication	24%
	Political speeches	2%
	Literature	1%
	Touristic texts	5%

companies which aim at selling products and services; commercial companies were contacted in the search for advertising material and product descriptions, among

[17] These preliminary calculations were made on the basis of the main annotator's initial labelling throughout the text-collection phase and do not account for doubtful cases, nor for hybrid contexts. The results of the interannotator agreement are expected to generate subtle modifications for the metadata *channel, intended audience, communicative purpose* and *topic*.

others. *Public services* are those government-owned institutions which do not pursue any commercial goal whatsoever. *Public enterprises* were defined as government-owned institutions which do engage to some extent in commercial activities (such as Belgian Railways NMBS/SNCB) and were expected to show similarities with private commercial companies. We expected this distinction between commercially oriented companies and institutions belonging to the public service sector to be reflected in the ways in which texts are adapted to their audience. Since we aimed to include a certain amount of scientific texts, *Research and development* was introduced as a separate category; in this category, text providers are research institutes which do not pursue financial profit. Finally, as in the original DPC, texts from *media institutions* or *publishers* were collected as well; these text providers are private, non-governmental organizations which publish newspapers and magazines (either online or in print). The classification of texts into one of these different text provider categories was obvious and did not cause any ambiguities.

Whereas assigning any of the above-mentioned metadata was fairly unproblematic, all other text-related metadata added to each text implied a textual and contextual analysis, i.e. each text needed to be read, and information about the text provider needed to be gathered in order to classify texts in one (or more) of the categories for *channel*, *intended audience*, *communicative purpose* and *topic*. Classifying texts in one of these categories is no easy task, as this partly depends on the interpretation of the annotator, and hence, unavoidably entails some degree of subjectivity. For instance, some texts are expected to fulfil more than one main communicative purpose. Moreover, it can be challenging to distinguish between texts which are intended for a broad external audience or a specialist external audience. In order to overcome the subjectivity concerning the *channel, intended audience, communicative purpose* and *topic* metadata, we decided that each text had to be annotated by at least two annotators, which allows for a post-hoc evaluation of the annotation procedure by means of an interrater agreement score per text. All texts were rated by students on the verge of obtaining their master's degree in Translation, Interpreting or Multilingual Communication, and, as a result, all have reached a high level of Dutch, English or French (or a combination of these). All Dutch texts were annotated by four independent annotators, the English texts by three annotators and the French texts by at least two raters.

In order to determine a text's intended audience (addressee), our raters were asked to make a distinction between *broad external*, *specialist* and *in between broad and specialist*, thereby referring to the amount of prior knowledge an addressee needs to grasp the intended message of a text.[18] While classifying each of the texts in one of the categories, annotators were asked to provide extra

[18] In addition to texts which were produced for an external audience, we were able to gather texts which are written for an *internal target audience*, in which organization-internal information is provided to a very specific, internal target audience. Texts which were produced for an internal audience are automatically classified as *specialist*.

information about the criteria they used to classify a text (e.g. the complexity of a text's content, grammar and vocabulary). By collecting this information, we aim to take into consideration which criteria played a role in the rating process and, as a result, keep this in mind when interpreting texts which are somewhere *in between broad and specialist*.

Channel or *mode* distinguishes between texts which are *written to be read, written to be spoken* or *written reproduction of spoken language*. With the exception of transcribed interviews or speeches, however, texts in DPC are mainly written to be read. Nevertheless, even in those written-to-be-read texts, quotes or citations are frequently included, thereby leaving traces of spoken language. In order to make a distinction between written-to-be-read texts which partially contain traces of spoken language and those texts which are exclusively written, our annotators were asked to mark the presence of frequent quotations in a text. This is indicated as an additional metadata field. In line with the bottom-up classification of web registers developed by Egbert, Biber and Davies (2015), we did not expect frequent quotations to emerge as a characteristic of any register in particular, but instead merely aimed to raise awareness for the traces of originally spoken language across all texts.

We approach a text's communicative purpose as multi-layered and acknowledge that texts can combine more than one communicative goal at once.[19] In order to determine the particular goal(s) of a single text, we readopted the undermentioned guidelines by Delaere (2015) and allowed our annotators to select one or two main communicative purpose(s):

– To inform: The text provides objective information on a particular service or product. No personal opinion is involved (e.g. press releases, yearly reports).

> Since we assumed that many texts fulfil some sort of informative purpose, we asked our annotators to select an additional goal when appropriate. In line with Delaere (2015), *inform* was only chosen exclusively when no other communicative purpose was present.

– To instruct: The text provides readers with a step-by-step guide on how to achieve a particular result. Often, but not always, some sort of physical act is required (e.g. recipe, manual).

> Additionally, this category also includes, for instance, guidelines and best practices which are less strictly presented as a step-by-step guide but nevertheless serve to instruct their readers.

– To persuade/to activate: The text provides arguments which aim to convince readers and change their point of view. As a result, readers may be encouraged to

[19] The calculations in Table 5 were based on each text's main communicative purpose. However, the register classification in DPC 2.0 (cf. Section 5) takes into consideration the presence of additional communicative goals within a single text.

take action. It involves both subjective and objective argumentation of the author (e.g. advertising).

> In contrast with Delaere (2015), these two communicative purposes are combined, since DPC 2.0 contains just a small amount of activating texts. We argue that the eventual aims of persuasive and activating texts are closely interrelated.

– To form an opinion: The text provides a personal opinion which aims to make readers reflect on a particular (societal) theme and stimulates readers to form an opinion of their own (e.g. reviews, opinion piece).

– To narrate: Involves story-telling which mainly aims to entertain readers (e.g. tales, short stories).

Finally, all annotators also determined the *topic* of each text (what the text is about). They could choose the name of the topic from a list which partially corresponds to the original DPC-topics (corporate, economy, culture, tourism, history, nature, education, leisure, politics, law, sports, transport, science) or choose another one (they could also pick more than one topic). In line with Biber and Conrad (2009), we argue that the identification of different topics may enhance transparency within a defined register category. As will be illustrated in the following section on the register classification of our corpus, the specification of topic enabled us, for instance, to single out *touristic texts* as a particular instantiation of *commercial communication*.

The annotation of the above-mentioned metadata categories by multiple raters adds another layer of information that can be employed by end-users of the corpus to verify, for instance, to what extent a certain text is annotated identically across different raters. At the time of writing, we have not conducted an extensive in-depth analysis of the annotation behaviour of the raters yet, but we did calculate some general interannotator agreement scores in Table 6, using Krippendorff's Alpha, which is particularly useful in the context of more than two raters (Hayes and Krippendorff 2007; De Swert 2012).

As can be seen in Table 6, reasonable interannotator agreement scores (>0.65) are obtained only for the annotation of channel-related metadata. The scores for intended audience and communicative purpose are clearly lower; not quite surprisingly, the annotation of an informative purpose yielded the most heterogeneous set of responses. This might be due to the very general nature of this communicative purpose, and one could claim that simply *all* texts are informative in nature. It is thus possible that raters did not always act consistently (within raters and across raters) when indicating whether a text is informative in nature or not. An in-depth analysis is needed in order to clarify this.

Table 6 Interannotator agreement scores for the text-related metadata (Krippendorff's Alpha)

Metadata category	Subcategory	Krippendorff's Alpha
Communicative purpose	To inform: y/n	0.29
	To persuade:y/n	0.51
	To form opinion: y/n	0.55
	To instruct: y/n	0.59
	To narrate: y/n	0.59
Channel/mode	Dialogue: y/n	0.66
	Quotations: y/n	0.76
Intended audience	-	0.40

5 Classifying Texts into Registers

The classification of texts into different registers, text types, genres is both a crucial and a difficult task for corpus compilers (cf. Neumann 2013). As has become clear in Delaere (2015) and Delaere and De Sutter (2017), the register classification in the first release of the DPC was to some extent problematic, since (i) some of the labels used were vague (e.g. external communication, administrative texts), (ii) some of the register categories contained very heterogeneous texts (recipes, governmental guidelines and commercial manuals were all classified as *instructive texts*) and (iii) some texts which are identical appeared in different register categories: some yearly reports ended up being labelled as *external communication*, others as *administrative texts*. Similarly, specific technical texts appeared as both *instructive texts* and *administrative texts*. This led Delaere (2015) and Delaere and De Sutter (2017) to partly reclassify the original DPC, using the situational parameters *text provider, intended audience, channel* and *communicative purpose* (see Sect. 4.2 for a discussion of these parameters) in new registers such as *tourist information, broad commercial texts* and *journalistic texts*.[20] By doing so, they followed Biber and Conrad's (2009) general register framework, in which registers are defined by language-external, situational characteristics, which in turn are associated with certain linguistic features that are used more or less frequently. In other words, 'core linguistic features like pronouns and verbs are functional, and, as a result, particular features are commonly used in association with the communicative purposes and situational context of texts [i.e. a specific register]' (p. 2) (see also Egbert et al. 2015).

Ideally, the combination of situational characteristics leads to a specific, unambiguous register category, which is at least in one respect different from other register categories. In reality, however, the combination of one or more situational characteristics results in a multitude of different registers, which do not clearly match with traditional register concepts. What usually happens is that researchers build their own register classification system based on ad-hoc combinations of

[20] Text provider and intended audience, respectively, refer to addressor and addressee mentioned in Biber & Conrad (2009).

situational parameters. In Delaere (2015), for instance, the narrowing down process of DPC registers ultimately resulted in seven register varieties. A distinction was made between, for instance, *specialized communication* and *broad commercial texts*, which are defined on the basis of a single criterion, namely, the intended audience. Two register varieties are defined on the basis of various situational criteria and, as a result, inherently cluster texts at a higher level of specificity. In fact, *legal texts* are determined by both a text's purpose (persuade-activate) as well as its addressor (public service), whereas *tourist information* is defined by a text's purpose (activate or inform/persuade), its addressor (public enterprise) *and* its addressee (broad external audience). Delaere and De Sutter (2017) warn that this reclassification procedure into a limited amount of register varieties is unavoidably 'idiosyncratic, and hence could be re-considered in future research' (p. 13).

The narrowing down of register categories indeed constitutes a difficult balancing task for corpus compilers since any combination of characteristics constitutes a potential register variety at a different level of generality. Researchers selecting only one situational characteristic for classifying texts inherently decide to define registers at a high level of generality and are thus expected to ascertain a wider range of linguistic variability in their registers. *Broad commercial texts* in the classification of Delaere (2015), for instance, still contain a large amount of internal variation. At the other extreme, the addition of too many situational characteristics might generate highly specific registers containing an insufficient amount of texts.

In analogy with the study of Delaere (2015), texts in DPC 2.0 were clustered on the basis of four situational characteristics: *text provider, intended audience, channel* and *communicative purpose*.[21]

This bottom-up approach resulted in nine clearly distinguishable register categories which, to some extent, coincide with the labels applied by Delaere (2015). Despite these similarities, however, particular text types in DPC 2.0 required an alternative, more specific register label. Table 7 presents an overview of the register categories which were included in DPC 2.0, as well as their associated situational characteristics.

In order to deal with the expected amount of internal variation within broadly defined registers, such as *commercial communication*, further specification on the basis of a text's topic leads to more concretely defined subregisters. As a result, we were able to identify *touristic texts* as a concrete subvariety within this overarching register.[22]

[21] As we mentioned in the previous section, these criteria were determined on the basis of the main annotator's initial labelling, in anticipation of the in-depth analysis of the students' ratings. All ratings will be added to the final corpus in order to allow for a more nuanced, fine-grained interpretation of (hybrid) situational criteria, depending on the specific aim(s) of each research project.

[22] In order to retrieve literature or journalistic texts which discuss a touristic topic, we invite end-users of DPC 2.0 to further subdivide all registers according to this particular topic. Additionally, the flexibility of our approach equally allows for a topic-based classification of texts, regardless of their predefined situational characteristics.

Table 7 Register classification in DPC 2.0

Register	Situational characteristics	Content
Manuals for a general audience	1. Audience (*broad external*) 2. Communicative purpose (*to instruct*)	e.g. manuals, DIY-guides, how-to-procedures for a broad audience
Manuals for specialists	1. Audience (*specialist*) 2. Communicative purpose (*to instruct*)	e.g. manuals, DIY-guides, how-to-procedures for a specialist or internal audience
(Popular) science	1. Text provider (*Research & Development*)	e.g. (popular) scientific texts, press releases
Journalistic texts	1. Text provider (*media & publisher*) 2. Communicative purpose *(to inform/ to form an opinion*)	e.g. news articles, opinion articles, columns for (non-) specialists
Commercial communication	1. Text provider (*commercial company/public enterprise*) 2. Communicative purpose (*to inform/to persuade*)	e.g. promotion or advertising material, commercial brochures and websites, e-mails, press releases, yearly reports
Public service communication	1. Text provider (*public service*) 2. Communicative purpose (*to inform/to persuade*)	e.g. product or service information, press releases, yearly reports, contracts, museum texts
Political speeches	1. Mode (*written to be spoken/ written reproduction of spoken language*)	e.g. official speeches, proceedings of parliamentary debates
Literature	1. Communicative purpose (*to narrate*)	e.g. books, book chapters, short stories, tales
Touristic texts	1. Text provider (*public enterprise*) 2. Audience (*broad external*) 3. Communicative purpose (*to inform/to persuade*) 4. Topic (*tourism*)	e.g. touristic brochures, promotion or advertising material

The proposed register classification in Table 7 was determined on the basis of the unique combination of one or more situational characteristics and by no means aims to be exhaustive. Instead, end-users of DPC 2.0 are invited to explore the potential of defining register varieties on the basis of shared situational characteristics in their own research project(s). Besides its apparent relevance for both translation scholars and contrastive linguists, DPC 2.0 equally appeals to linguists dealing with registers and/or register variation on a more general note. In particular, by including the results of all student annotations for each text in the corpus, we stimulate researchers to observe texts which caused disagreement between annotators, since this ambiguity suggests the possibility of flexible, multi-layered, registers.

6 Corpus Exploitation

The way in which texts are collected, processed and stored constitutes a crucial step for the eventual exploitation of a corpus. At the outset of their project(s), compilers of (parallel) corpora should therefore carefully consider how their data will eventually be made available and which format best suits those purposes. As mentioned in Sect. 3 on data-processing, the annotated text files of DPC 2.0 were stored in a transparent machine-readable format. As in the first version of DPC, each filename is named in a consistent way and, as such, refers to a unique content. However, whereas the first version of DPC used three separate XML-files to store one source-target text pair—with monolingual files for each annotated source text or translation, bilingual files for each sentence-aligned language pair and metadata files for each source text or translation—DPC 2.0 shows an increased efficiency and exclusively makes use of one extended TMX-file for each pair. As a result, metadata files are automatically added in the header of each processed text pair.

The re-birth of the Dutch Parallel Corpus, which we described in the previous sections, allows researchers from different disciplines to carry out linguistic research on translations and their source texts in which the relationship with the extra-linguistic context plays an important role. The output of each search query can be filtered according to a large variety of text-related, translation-related and translator-related criteria. The bottom-up register classification presented in Table 7, for instance, enables translation scholars and contrastive linguists to analyse linguistic phenomena in texts sharing a specific context of use. In analogy with the topic-based subclassification of *touristic texts* (cf. Section 5), DPC 2.0 also provides numerous additional opportunities to cluster texts on different levels of generality. Moreover, the significant amount of translator- and translation-related metadata encourages translation scholars to search and compare a great variety of translator profiles. DPC 2.0 provides the possibility to analyse texts which were for instance translated by i) female translators ii) with the assistance of a CAT-tool, etc. Depending on the amount of specified metadata, queries can thus generate both general as well as highly specific data sets. Finally, as mentioned in the previous section, we aim to include the rating results of our student annotators in order to facilitate the identification of multi-layered texts in the corpus. Whereas the register classification of DPC 2.0 already takes into account texts pursuing more than one communicative goal, the rating results point towards additional layers of complexity within a text which require further attention, such as written texts containing a considerable amount of originally spoken language or texts whose intended audience was labelled as *in between specialist and broad external*.

7 Conclusion

A decade after the completion of the original Dutch Parallel Corpus (DPC), we presented an updated version of the corpus which aims to respond to the recent need for more qualitative data in corpus-based translation studies. Our motivation to re-create the original DPC was primarily inspired by two issues which over the years prevailed with end-users of the original corpus. First and foremost, the limited amount of metadata included in the first version of DPC offered insufficient information with regard to a text's production circumstances and the translator(s) who produced the translation, making it challenging to distinguish, for instance, between L1 and L2 translation and between 'original' source texts and pivot source texts, which were originally translated from yet another source text. Second, for some of the metadata available in the original DPC, it was not always clear how these were gathered and organized; the corpus contained, for instance, an ambiguous register classification, which was reflected in the application of vague register labels, viz. *administrative texts*, and their sometimes heterogeneous content. In order to overcome its predecessor's main issues, the Dutch Parallel Corpus 2.0 is compiled with specific attention for the inclusion of a wide range of additional and verified metadata regarding textual properties, translator properties and translation project properties. As such, DPC 2.0 distinguishes itself for its large amount of contextual information related to the source texts, the translations and the translators, thereby enabling a detailed analysis of similarities and differences between translator profiles and a text's production circumstances, and the effects these have on their linguistic choices. The specification of this extensive amount of contextual metadata in DPC 2.0 eventually led us to integrate a more flexible register classification on the basis of their shared situational characteristics. This bottom-up approach, which was previously put into practice by Delaere (2015), allows to generate various context(s) of use, and ultimately encourages researchers to renounce to the traditional register labels which often do not fully seem to account for linguistic variation within a register, nor for linguistic similarities across registers.

Appendix

Questionnaire for translators

1. Documents or websites translated (please mention the title of each text):
2. Translation direction:
3. Collaborative translation:
    - yes
    - no
4. Translator's gender:
    - m
    - f
    - x
5. Translator's degree:
    - no specific language degree
    - translation Master
    - translation Bachelor
    - language and literature
    - interpreting
6. Experience as a translator (in years):
7. Translator's year of birth:
8. Translation tools or memory involved:
    - none, manual translation
    - CAT-tool, i.e. ________________________________
    - post-editing—machine translation, i.e. ____________________________
9. Translation directionality:
    - (L1 (first language)
    - (L2 (foreign language)
10. Translator's status
    - freelance
    - in-house
    - both
11. Use of style guides:
    - in-house guidelines
    - in-house glossary
    - both
    - none
12. Domain expertise (regarding the text's topic)
    - expert
    - non-expert

13. External revision
    - monolingual (only translation)
    - bilingual (source text and translation)
    - no revision

References

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and technology: In honour of John Sinclair*, ed. M. Baker, G. Francis, and E. Tognini-Bonelli, 233–250. Philadelphia, PA/Amsterdam: Benjamins.

Baker, M. 1996. Corpus-based translation studies: the challenges that lie ahead. In *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*, ed. H. Somers, 175–186. Amsterdam/Philadelphia: John Benjamins.

Baker, M. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9 (2): 167–193.

Biber, D., and S. Conrad. 2009. *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Corpas Pastor, G., and M. Seghiri. 2016. *Corpus-based approaches to translation and interpreting: From theory to applications*. Frankfurt am Main [etc]: Lang.

Danielsson, P., and D. Ridings. 1997. Practical presentation of a "vanilla aligner‟. In *Proceedings of the TELRI workshop on alignment and exploitation of texts,* Ljubljana.

De Clercq, O., and M. Montero Perez. 2010. Data collection and IPR in multilingual parallel corpora: Dutch parallel corpus. In *LREC 2010 : Seventh conference on international language resources and evaluation*, ed. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias, 3383–3388. Paris, France: European Language Resources Association (ELRA).

Delaere, I. 2015. *Do translations walk the line?: Visually exploring translated and non-translated texts in search of norm conformity*. Faculty of Arts and Philosophy, Ghent, Belgium: Ghent University.

Delaere, I., and G. De Sutter. 2017. Variability of English loanword use in Belgian Dutch translations : Measuring the effect of source language, register, and editorial intervention. In *Empirical translation studies: New methodological and theoretical traditions*, vol. 300, ed. G. De Sutter, M.-A. Lefer, and I. Delaere, 81–112. Berlin/Boston: De Gruyter Mouton.

De Sutter, G., P. Goethals, T. Leuschner, and S. Vandepitte. 2012. Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures* 13 (2): 137–143.

De Sutter, G., M.-A. Lefer, and I. Delaere (eds.). 2017. *Empirical translation studies: New methodological and theoretical traditions*. Berlin, Boston: De Gruyter.

De Sutter, G., and M.-A. Lefer. 2019. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives-Studies in Translation Theory and Practice* 28 (1): 1–23.

De Swert, K. 2012. *Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha*. Unpublished manuscript University of Amsterdam. https://www.polcomm.org/wp-content/uploads/ICR01022012.pdf.

Egbert, J., D. Biber, and M. Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66 (9): 1817–1831.

Fantinuoli, C., and F. Zanettin. 2015. *New directions in corpus-based translation studies (Translation and Multilingual Natural Language Processing 1)*. Berlin: Language Science Press.

Geyken, A. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, ed. C. Fellbaum, 23–41. Continuum Press.

Granger, S., and M.-A. Lefer. 2020. The multilingual student translation corpus: A resource for translation teaching and research. *Language Resources & Evaluation* 54: 1183–1199. https://doi.org/10.1007/s10579-020-09485-6.

Halverson, S.L. 2013. Implications of cognitive linguistics for translation studies. In *Cognitive linguistics and translation: Advances in some theoretical models and applications*, ed. A. Rojo, and I. Ibarretxe-Antuñano, 33–74. Berlin/Boston: Mouton de Gruyter.

Halverson, S.L. 2015. Cognitive translation studies and the merging of empirical paradigms. The case of 'literal Translation'. *Translation Spaces* 4(2): 310–40.

Halverson, S.L. 2017. Gravitational pull in translation: Testing a revised model. In *Empirical translation studies: New methodological and theoretical traditions*, ed. G. De Sutter, M.-A. Lefer, and I. Delaere, 9–46. Berlin/Boston: Mouton De Gruyter.

Hayes, A.F., and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1 (1): 77–89.

Ji, M. 2016. *Empirical translation studies. Interdisciplinary methodologies explored*. Sheffield, UK: Equinox.

Kotze, H. 2020. Converging what and how to find out why: An outlook on empirical translation studies. In *New Empirical Perspectives on Translation and Interpreting*, ed. L. Vandevoorde, J. Daems, and B. Defranq, 333–371. Routledge.

Kruger, H., and G. De Sutter. 2018. Alternations in contact and non-contact varieties. Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior* 1 (2): 251–290.

Kruger, H., and B. Van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13 (1): 33–65.

Kruger, H., and B. Van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and non-native indigenised varieties of English. *English World-Wide* 37 (1): 26–57.

Laviosa, S. 2002. *Corpus-based translation studies. Theory, findings, applications.* Amsterdam/New York: Rodopi.

Lefer, M.-A. 2020. Parallel corpora. In *A practical handbook of corpus linguistics*, ed. M. Paquot and S. Th. Gries, 257–282. Springer.

Macken, L., O. De Clercq, and H. Paulussen. 2011. Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta* 56 (2): 374–390.

Malamatidou, S. 2018. *Corpus triangulation: Combining data and methods in corpus-based translation studies*. London: Routledge.

Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2): 313–330.

Melamed, D.I. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th annual meeting of the association of computational linguistics (ACL)*, 305–312. Madrid, Spain.

Mellinger, C.D., and T.A. Hanson. 2016. *Quantitative research methods in translation and interpreting studies*. Abingdon, Oxon: Routledge.

Moore, R.C. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th conference of the association for machine translation in the Americas*, 135–244. Tiburon, California.

Neumann, S. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: de Gruyter.

Oakes, M.P., and M. Ji. 2012. *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research.* Philadelphia, PA/Amsterdam: Benjamins.

Olohan, M. 2004. *Introducing Corpora in translation studies*. London: Routledge.

Paulussen, H., L. Macken, W. Vandeweghe, and P. Desmet. 2013. Dutch parallel corpus: A balanced parallel corpus for Dutch-English and Dutch-French. In *Essential speech and language technology for Dutch: Results by the STEVIN-programme*, ed. P. Spyns, and J. Odijk, 185–199. Berlin, Germany: Springer.

Qi, P., Zhang Yuhao, Zhang Yuhui, J. Bolton, and C.D. Manning. 2020. A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*, 101–108. Online.

Vandevoorde, L., J. Daems, and B. Defrancq, eds. 2020. *New empirical perspectives on translation and interpreting.* Routledge.

Van Eynde F., J. Zavrel, and W. Daelemans. 2000. Part of speech tagging and lemmatisation for the spoken Dutch Corpus. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, ed. M. Gavrilidou et al., 1427–1433. Athens, Greece.

Xiao, R., and X. Hu. 2015. *Corpus-based studies of translational Chinese in English-Chinese Translation*. Springer.

Ryan Reynaert is a research assistant at the Department of Translation, Interpreting and Communication at Ghent University (Belgium).

Lieve Macken is Associate Professor at the Department of Translation, Interpreting and Communication at Ghent University (Belgium). She has strong expertise in multilingual natural language processing. Her main research focus is translation technology, translation quality assessment, and translation difficulty. She has been involved in several corpus creation projects, amongst which the original Dutch Parallel Corpus.

Arda Tezcan is a post-doctoral researcher at the Department of Translation, Interpreting and Communication at Ghent University (Belgium) with a background in mathematics and artificial intelligence. His research interests include natural language processing, machine translation and human-machine interaction in the context of translation studies.

Gert De Sutter is Associate Professor at the Department of Translation, Interpreting and Communication at Ghent University (Belgium). His research can be broadly characterised as empirical (corpus-based) research of linguistic variation in translated and non-translated texts, and aims to get a better insight into the complex interaction of social norms and bilingual cognition in translators.

Probing a Two-Way Parallel T&I Corpus for the Lexical Choices of Translators and Interpreters

Oi Yee Kwong

Abstract Working under greater time pressure, interpreters often necessarily produce less refined renditions than translators do. At the lexical level, some studies have hypothesised that interpreters could only access their most active vocabulary containing more frequent words, while others have suggested that words used by interpreters tend to be less formal and accurate than those used by translators. The connection between such intuitions and observations remains to be investigated empirically and thoroughly. In this study, we made use of a two-way parallel translation and interpreting (T&I) corpus to examine and compare the lexical choices of translators and interpreters. The frequency effect was tested and concrete examples were analysed, to understand the linguistic behaviour of translators and interpreters in relation to the nature and demand of their tasks, respectively.

Keywords Lexical choices · Frequency effect · Simultaneous interpreting · Two-way parallel corpus · Translators and interpreters

1 Introduction

Although translation and interpreting are often perceived as very closely related activities given their apparently similar communicative functions, that is, to take what has been expressed in one language and re-express it in another language, it is nevertheless over-simplified to only characterise them by their modality as written and oral translations, respectively.

The original version of this chapter was revised: In abstract section spell error corrections have been incorporated. The correction to this chapter is available at https://doi.org/10.1007/978-981-16-4918-9_13

O. Y. Kwong (✉)
Formerly The Chinese University of Hong Kong, Hong Kong, China
e-mail: oykwong@cantab.net

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_5

Modality aside, it is well noted that translation and interpreting are in essence fairly distinct activities in various regards, which collectively lead to very different manifestations with certain readily observable features. A major factor is undoubtedly the time allowed for translators and interpreters to finish their tasks under normal circumstances. While interpreters have to come up with a rendition almost instantly and move on right away, translators can often think it over and afford extra time to polish their products. They tend to work for different purposes, in different situational contexts, and with a different set of skills. Thus although both translation and interpreting (simultaneous interpreting, in particular) are cognitively demanding, involving very complicated mental operations for language comprehension and production, the management of the cognitive efforts is different for the two activities, and interpreting is often considered a more difficult exercise. According to the Effort Model (Gile 2009), simultaneous interpreting (SI) is depicted as consisting of three Efforts, namely the Listening and Analysis Effort (L), the Short-term memory Effort (M), and the Speech production Effort (P), which are coordinated by the Coordination Effort (C):

$$\boldsymbol{SI = L + P + M + C}$$

As one's total processing capacity is fixed, the success of SI heavily relies on the effective allocation of the limited available capacity to the various Efforts which may be concurrently active and operating at any given time.

In contrast, written translation is modelled much more simply as:

$$\boldsymbol{Translation = Reading\ Effort + Writing\ Effort}$$

where there is no competition between the two Efforts, as the processing capacity can be devoted to reading and writing alternately, without overlap or clash along the time line.

The potential competition for limited cognitive resources at any one time in simultaneous interpreting thus makes it a more difficult and demanding task than written translation. Consequently, expectations of translation and interpreting are different, as reflected in their performance evaluation criteria, and it seems natural that interpreting often warrants more tolerance of imperfections like errors and omissions. On the other hand, to ensure smooth operation with acceptable results and professional performance, a common aim in interpreter training is to make the best of one's working memory capacity. In addition to short-term memory drills, it also involves ways to automate the speech production process with the use of habitual expressions and other direct correspondences available for a specific language pair, so as to free up as much capacity as possible.

As far as speech production is concerned, lexical access is obviously one of the first hurdles. For written translations, it is quite clear that '[a]lmost every word used by the translator is the result of a choice (not necessarily conscious, of course)' (Santos 2004: 21). Interpreters must have also undergone such steps, except that their choices have to be made swiftly and may as a result be crude at times (e.g. Shlesinger 2008). It has been suggested in some models that under the time pressure, interpreters often only access their active vocabulary which often contains more frequent words (e.g. Gile 2009). While such a hypothesis is intuitively plausible, it remains to be better supported by empirical evidence, on top of casually observed isolated instances. In other words, does the frequency assumption always hold in practice? Are the lexical choices by translators and interpreters demonstrably different? If so, how exactly are they different? How do these differences (or similarities) found empirically enhance our understanding of the nature and demand of translation and interpreting in general, and cross-lingual lexical access in particular?

In this study, we will address these questions based on the empirical evidence obtained from a two-way parallel translation and interpreting (T&I) corpus. Our data consist of written translations and interpreted speeches, from English to Chinese/Cantonese, produced by professional translators and interpreters, for essentially the same sources in their written and spoken forms, respectively. We focus on the phenomena at the lexical level, fully realising that they may not always be completely detachable from syntactic and discourse considerations. The findings will allow us to better understand the material differences between the lexical choices made by translators and interpreters, if any, and guide future research as well as training approaches.

2 Characterising Translationese and Interpretese

The rise of corpus-based studies has notably allowed more systematic and objective characterisation and comparison of text genres by means of a collection of linguistic features and measures (e.g. Biber 1988). The quantitative analysis enabled by corpora has been applied in translation studies, for stylistic comparisons among and across translators (e.g. Baker 1995). It has also led to the investigation into translation universals, like simplification, explicitation and normalisation (Laviosa 2012). Such patterns or tendency are considered salient indicators to differentiate translations from non-translations. For instance, translated texts are often found to show a relatively lower proportion of lexical words to grammatical words, and a relatively higher proportion of high-frequency to low-frequency words. Repetitions are abundant in translated texts, thus the lexical variety therein is lower than that in non-translated texts.

While corpus-based translation studies have started to grow and flourish since 1990s as pioneered by Mona Baker (e.g. Baker 1993), corpus-based interpreting studies obviously started some years later, mostly signified by Miriam Shlesinger's seminal paper describing it as an offshoot of the former (Shlesinger 1998). The lag for corpus-based interpreting studies is to a certain extent attributable to the limited relevant corpus resources available to start with, and adding to the scarcity issue is the relative difficulty of manually constructing them. Spoken corpora have all along been more expensive to build than written ones (Kennedy 1998). Naturally occurring interpreting materials are less readily obtained, and large interpreting corpora may only be specifically constructed in artificial settings for pedagogical and other purposes (e.g. Tohyama and Matsubara 2006). From time to time, studies would have to rely on interpreting data elicited for experimental purposes (e.g.

Shlesinger 2008). Even when there are accessible recordings from naturalistic data, the interpreting directions and the amount can sometimes be quite skewed. For instance, the European Parliament Interpreting Corpus (EPIC) (Bendazzoli and Sandrelli 2005) contains source speeches made in the European Parliament in Italian, English and Spanish, and the simultaneously interpreted speeches in all possible combinations and directions. However, the sizes of the sub-corpora with English as the source are much larger than those with the other two languages as the source. For other SI data from official sources, they are often SI from the official language into English, such as the Chinese-English Conference Interpreting Corpus (CECIC) based on the interpreting at the press conferences of the State Council of the People's Republic of China (Hu and Tao 2010). In addition to the data availability problem, there are considerable issues in the transcription and annotation of paralinguistic information in interpreting corpora to provide significant indicators for particular characteristics of interpreted speech (Zou and Wang 2014).

Interpreting studies relying on corpora could be corpus-based, corpus-driven or corpus-informed (Zhang 2013). On a macro level, corpus-based interpreting studies have by and large followed a similar trajectory as its translation precedents (and counterparts) in terms of research questions and methodology. For instance, Shlesinger (2008) found that the lexical variety is even smaller for interpreting, which also exhibits somewhat different part-of-speech distributions from translation. On a micro level, however, they did develop quite autonomously given the very difference between interpreting and translation as distinct activities. As far as the language pair English-Chinese is concerned, previous corpus-based interpreting studies tend to focus on interpreting into English as a second language, and more often on the features at the sentence or discourse level. For example, Hu and Xie (2014) found that Chinese-English interpreters tend to use more relative clauses in conference interpreting than translators do when they render Chinese government work reports from Chinese to English. Pan (2014) found that interpreters tend to add more hedges when they work into English than translators do in written translation, although it is still significantly less than English speakers do in original English speech. Wang and Zou (2018) examined how long and complex attribute structures in Chinese are interpreted into English.

Comparing and contrasting written translations with oral translations has long been considered an important research agenda for reinforcing the links between translation studies and interpreting studies (e.g. Chesterman 2004). To this end, one does not only need an adequate amount of data for both modalities, that is, translated texts and interpreted speeches, but also need them to be comparable in order to establish a reasonable connection between what is found for translationese and interpretese individually. Shlesinger and Ordan (2012) compared English-Hebrew simultaneous interpreting with written translation into Hebrew and original Hebrew speech. They found more similarities between SI and original speech than between SI and written translation, which consistently demonstrates SI's tendency towards orality (that is, interpretese is more spoken than translated). Meanwhile, comparison by other parameters like type-token ratio also suggests that interpreting is after all an extreme case of translation which also exhibits certain universals pertaining to translation in general.

In addition to lexical variety and part-of-speech distribution, as mentioned earlier, Shlesinger's (2008) study also found remarkably different lexical choices between interpreting and translation, especially in terms of high-register and low-register words. Only two examples for closed-class words were given as 'a mere sampling of the striking differences that were found in terms of lexical choices, with the participants showing a clear preference for the unmarked form when interpreting but a clear preference for a formal, marked alternative when translating' (Shlesinger 2008: 248–249). While the formality of the words chosen in interpreting, as compared to translation, may be a result of the typical characteristics of spoken and written languages respectively, it could as well be caused by the different requirements and cognitive demands of the two modes of translation in practice. Unfortunately, the study did not go on to further analyse the semantic properties of individual lexemes, and simply assumed that paradigmatic choices among available patterns, as a key indicator of register in spontaneous speech, should apply to translations as well (Shlesinger 2008). Hence, a more systematic comparison and analysis of the lexical choices between translators and interpreters along this line is definitely necessary.

Similarly, An and Zhang (2014) compared the translated and interpreted English versions of former Chinese Premier Zhu Rongji's press conferences, with respect to language style, lexical choice, sentence structure and discourse markers. It was only generally claimed that the written translations tend to be more formal, with more accurate words and more complex sentences, and informationally dense, while the oral translations are colloquial, with repetitive word choices, simple sentences and sparse information. The analysis of lexical choice only briefly touched on formality and precision, showing a handful of casual examples and is far from being systematic and thorough enough.

3 Corpus Data as Manifestation of Cognitive Processes

The linguistic phenomena observed from naturalistic data can be taken as the manifestation of the underlying cognitive process, although many studies tend to tap the cognitive process involved in translation and interpreting in an experimental setting (e.g. Gile 1999) or with psycholinguistic methods. They often found performance differences in simultaneous interpreting between those who have received specific training and those who did not, despite similar bilingual proficiency and cognitive abilities for the two groups (e.g. Christoffels et al. 2006). Often it is not only a matter of the working memory capacity (e.g. Stavrakaki 2012) but also, more importantly, the skills in utilising limited cognitive resources to manage competing tasks effectively (e.g. Liu et al. 2004). These findings, however, have to be understood with the language pairs, directionality, and task nature (relatively simple memory or lexical tasks, or genuine interpreting) taken into account.

On top of the language proficiency that one might expect of translators and interpreters in general, performance in actual interpreting situations often depends on the ‘availability’ of lexical units. The availability issues, according to Gile (2009), often arise after sound signals are passed on to the working memory for processing in speech comprehension, and at the planning stage of speech production. The implications are explained by the Gravitational Model of language availability. The language constituents, including lexical units, compositional rules of general language and rules of languages for special purposes, are perceived to be gravitating on orbits around a nucleus. Those on an orbit closer to the nucleus are supposed to have higher availability. However, such availabilities are by no means static and may vary with time, context and situation. One important factor affecting the dynamics of the system is stimulation frequency. In fact, the frequency effect is very typically supported by psycholinguistic experiments, where more frequent words are often more quickly accessed in the mental lexicon (e.g. Kwong 2016).

Under the time pressure, the more active vocabulary in one’s mental lexicon, which contains the more frequently used words, is often more directly accessed. Interpreter training also encourages the use of the high availability of translinguistic associations to free processing capacity for other tasks (e.g. Guy 2018), while such a relatively mechanical operation acceptable in oral translation is much less favourably considered for written translation. Training exercises for interpreters usually involve some elements to enhance one’s language skills in general, in addition to meeting the technical and cognitive skills specifically required for interpreting. These drills may focus at the discourse level, such as paraphrasing, and at the lexical level, such as searching for synonyms, hyponyms and hypernyms (Ilg 1978 cited in Gile 2009). The usefulness of such exercises and the application of such skills in actual work could be investigated empirically by comparing translators’ and interpreters’ lexical choice in a pair-wise manner.

Although the psycholinguistic aspects are not exactly the focus of the current study, we are interested in how the linguistic features observed from the products of translation and interpreting, at the lexical level particularly, may connect to the cognitive differences between the two activities. On the one hand, investigations by means of parameters like type-token ratio and word class distributions, as commonly done in corpus-based studies, may serve to characterise written and oral translations as different genres, but fall short in relating the actual differences to their very nature and to the cognitive causes behind. On the other hand, provided the appropriate corpus data are available, more detailed pair-wise comparisons would be necessary to illuminate the qualitative, in addition to quantitative, differences between written and oral translations attributable to the task requirements and cognitive demands on the translators and interpreters.

4 The Current Study

The current study thus aims at making use of an authentic corpus with naturalistic parallel translation and interpreting data to compare the lexical choices of translators and simultaneous interpreters in a pair-wise manner. In addition, we expect to make unique contributions in the following regards:

- The corpus, to be described in the following section, contains authentic translation and interpreting materials produced by professional translators and interpreters in Hong Kong. It provides first-hand empirical evidence for investigating the process of translation and simultaneous interpreting by means of the linguistic features found in the products, in relation to the cognitive aspects of the two activities.
- Many previous studies, particularly those on English-Chinese, tend to examine the situations where interpreters work into a B language as defined by the International Association of Conference Interpreters (AIIC 1982), that is, an active working language of the interpreter who can master it both actively and passively almost as well as a native. In this study, we look at simultaneous interpreting from English into the interpreters' native tongue, or A language, that is, Cantonese in this case.
- The study compares the lexical choices between interpreters and translators based on transcribed interpreted speeches and written translations for basically the same source presented orally and in written form respectively. The parallel data are naturalistic, in contrast to being experimental data, and from a relatively formal context. It provides convincing examples of the actual strategies employed in real practice for both teaching and research.
- Instead of macro-features like lexical density, type-token ratio, word length, sentence length, etc., in this study we pay special attention to how interpreters and translators render the same English words in Cantonese and Modern Standard Chinese, respectively. This microscopic way of exploring how their behaviour may be related to the time factor and thus the cognitive demands of interpreting and translation is rarely done before and is definitely contingent upon the availability of the appropriate corpus data.

5 A Two-Way Parallel T&I Corpus

As mentioned above, while written translations are more readily usable for corpus-based investigation, interpreted speeches are not as easily accessible especially after the events. Moreover, the lack of genuinely parallel materials makes it very difficult to compare interpreting with translation fairly. For instance, the interpreting data, translation data, and ordinary speech data in Hebrew used in Shlesinger and Ordan (2012) are only comparable, with content in approximately the same domain. Similarly, the Chinese-English Conference Interpreting Corpus and the written translation of Chinese government work reports from Chinese to English, used in Hu and Xie (2014), are also only comparable and not really parallel, despite the considerable overlap in the content of the press conferences and the government work reports. The resource in An and Zhang (2014) is possibly a rare exception, where there is also written records published for the former Chinese Premier's meetings with the press, which co-exist with the interpreted versions.

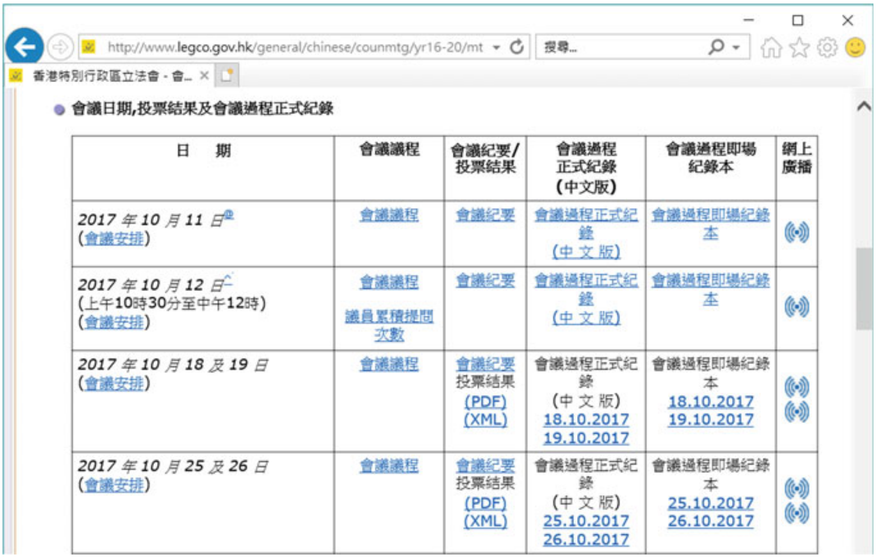
A parallel translation and interpreting (T&I) corpus with the written and oral renditions of the same source is particularly critical for the kind of investigation pursued in the current study. Although it may be feasible to compare the renditions of the same lexical items or phrases in separate translation corpora and interpreting corpora, it is in practice not possible to control for other factors. These factors, like the language proficiency of translators/interpreters, their skill levels, the type and nature of the source (formality, topic, etc.), and the context underlying the translation/interpreting (purpose, audience, formality, etc.), may interact to affect the actual translation or interpreting product.

Thus we resort to the council meeting records of the Hong Kong Legislative Council (LegCo) which are kept in both English and Chinese, the two official languages of Hong Kong. The proceedings of the meetings are recorded in writing, first in either English or Chinese as per original speech, and then translated into the other language. This bilingual written record is more often known as the Hong Kong Hansard and we take the English-to-Chinese parts for our translation corpus. The speeches delivered in the meetings, which may be in Chinese (mostly Cantonese) or English, are video-recorded, with simultaneous interpreting into English, Cantonese and Mandarin as appropriate. We select the English speeches and the corresponding Cantonese interpreted speeches, and transcribe them to form our interpreting corpus. In this way, we have a bilingual parallel T&I corpus.

We can be quite confident that the translation and interpreting are rendering the same source in the same context, only in a different modality, as assured by the official LegCo website (http://www.legco.gov.hk): 'The records of proceedings of the Council are first presented in the original language as delivered by Members and officials at Council meetings (Floor version). They will then be translated into the English and Chinese versions separately'. (The Chinese version: 立法會會議過程紀錄首先是以議員及官員在立法會會議上發言時所用的語言編製而成 (即場紀錄本)。其後, 即場紀錄本會分別翻譯為中、英文版本。)

5.1 Speech Sources

On the LegCo website, the meeting schedule and records of proceedings are all documented and publicly accessible, as shown in Fig. 1. In the second column from the right, the floor record (會議過程即場紀錄本) of a meeting shows every single turn of speech made by the Chairman, the members, or other government officials, with the content recorded in the original language of the speech. Hence, under each


http://www.legco.gov.hk/general/chinese/counmtg/yr16-20/mt
搜尋...
香港特別行政區立法會 - 會...

● 會議日期,投票結果及會議過程正式紀錄

日　期	會議議程	會議紀要/投票結果	會議過程正式紀錄(中文版)	會議過程即場紀錄本	網上廣播
2017 年 10 月 11 日 (會議安排)	會議議程	會議紀要	會議過程正式紀錄 (中 文 版)	會議過程即場紀錄本	((•))
2017 年 10 月 12 日 (上午10時30分至中午12時) (會議安排)	會議議程 議員累積提問次數	會議紀要	會議過程正式紀錄 (中 文 版)	會議過程即場紀錄本	((•))
2017 年 10 月 18 及 19 日 (會議安排)	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 18.10.2017 19.10.2017	會議過程即場紀錄本 18.10.2017 19.10.2017	((•)) ((•))
2017 年 10 月 25 及 26 日 (會議安排)	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 25.10.2017 26.10.2017	會議過程即場紀錄本 25.10.2017 26.10.2017	((•)) ((•))

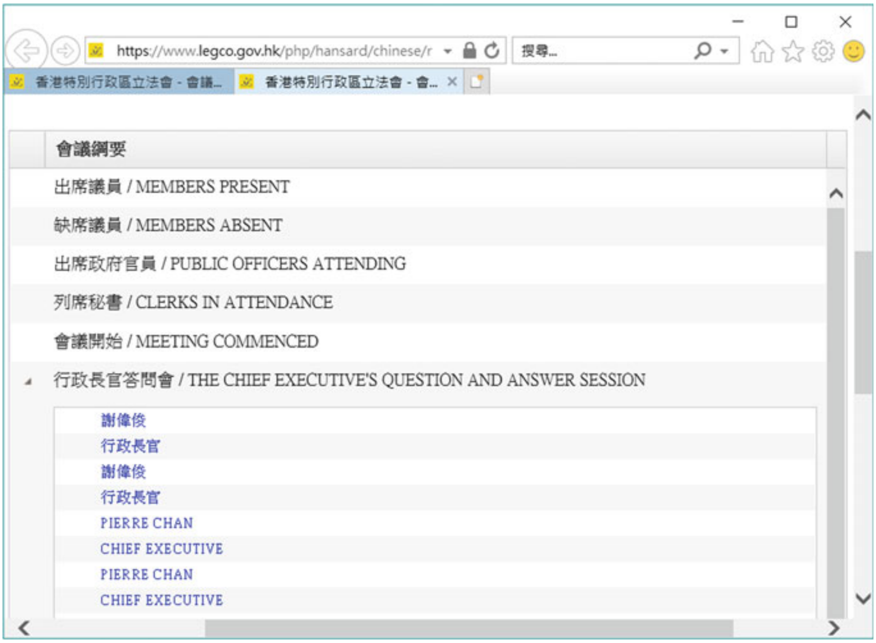
Fig. 1 Meeting schedule and records of proceedings accessible from LegCo website

agenda item, the order of speaking is shown, as in Fig. 2. The speakers' names are listed in Chinese or English, which indicates whether they spoke in Cantonese or English, respectively. Only the English speeches were selected for the current corpus.

Under the rightmost column, webcast (網上廣播), are the links to the official video recording of the meetings. The English speeches were obtained by playing the original version (現場), whereas the interpreted speeches were obtained by choosing to broadcast in Cantonese (粵語), as shown in Fig. 3.

5.2 *Speech Transcription*

The construction of spoken corpora is notoriously expensive particularly for the transcription that has to be done. In this work, the most labour-intensive and time-consuming part is to transcribe the Cantonese SI speeches. We tried to save some effort by starting with the 'Voice typing …' function in Google Docs, setting the language to Chinese (Hong Kong)/中文 (香港). A speech clip was played as voice typing started, to get a draft transcription, as in Fig. 4. The quality of the voice typing varies, depending on the talking speed, clarity of speech, articulation of the speaker, vocabulary used, etc. The draft transcription was then manually checked and corrected with fine listening, which had to be repeated several times. Other information like hesitations (<uh>) and self-repairs (^^) was added.


https://www.legco.gov.hk/php/hansard/chinese/r
搜尋...
香港特別行政區立法會 - 會議...
香港特別行政區立法會 - 會...

會議綱要

出席議員 / MEMBERS PRESENT

缺席議員 / MEMBERS ABSENT

出席政府官員 / PUBLIC OFFICERS ATTENDING

列席秘書 / CLERKS IN ATTENDANCE

會議開始 / MEETING COMMENCED

行政長官答問會 / THE CHIEF EXECUTIVE'S QUESTION AND ANSWER SESSION

謝偉俊
行政長官
謝偉俊
行政長官
PIERRE CHAN
CHIEF EXECUTIVE
PIERRE CHAN
CHIEF EXECUTIVE

Fig. 2 Speakers listed in English or Chinese indicating the language they used

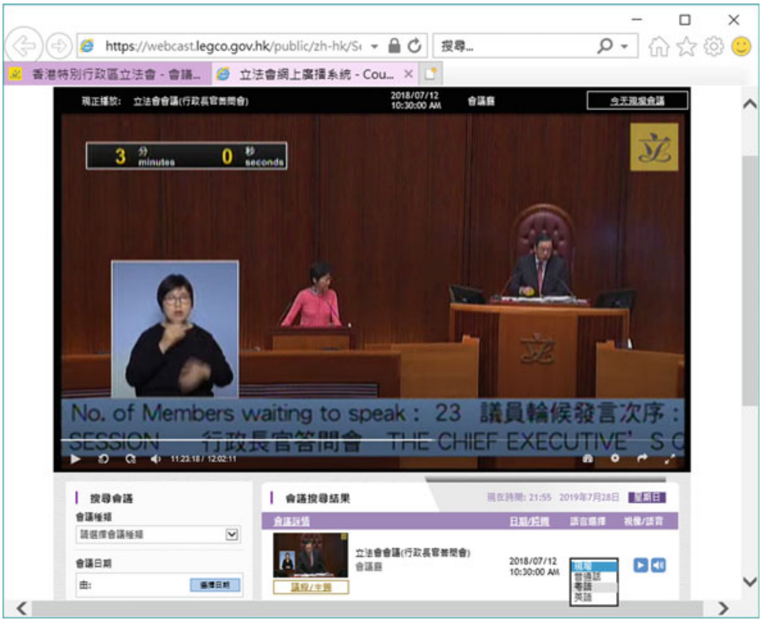
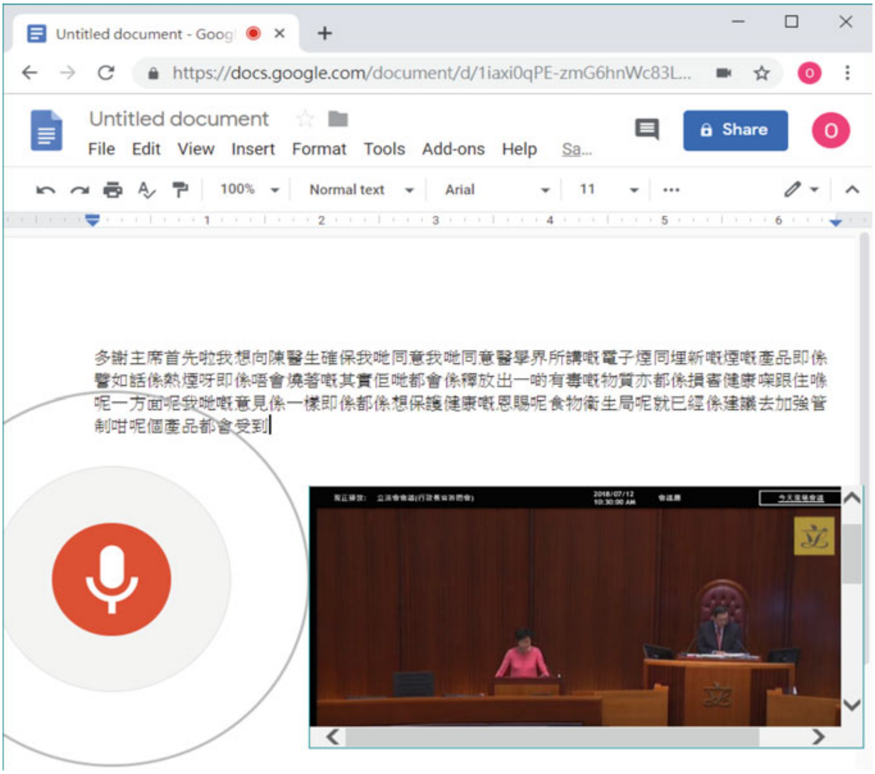

https://webcast.legco.gov.hk/public/zh-hk/S
搜尋...
香港特別行政區立法會 - 會議...
立法會網上廣播系統 - Cou...
現正播放: 立法會會議(行政長官答問會)
2018/07/12
10:30:00 AM
會議廳
今天視像會議
3 分 minutes 0 秒 seconds
No. of Members waiting to speak : 23 議員輪候發言次序 :
SESSION 行政長官答問會 THE CHIEF EXECUTIVE' S
11:23:18 / 12:02:11
搜尋會議
會議種類
請選擇會議種類
會議日期
由:
選擇日期
會議搜尋結果
現在時間: 21:55 2019年7月28日
星期日
會議詳情
日期/時間
語言選擇
視像/語音
立法會會議(行政長官答問會)
會議廳
2018/07/12
10:30:00 AM
現場
普通話
粵語
英語
議程/主題

Fig. 3 Obtaining the speech clips from the official video recording


Untitled document - Goog
https://docs.google.com/document/d/1iaxi0qPE-zmG6hnWc83L...
Untitled document
File Edit View Insert Format Tools Add-ons Help Sa...
Share
100%
Normal text
Arial
11
2018/07/12
10:30:00 AM

多謝主席首先啦我想向陳醫生確保我哋同意我哋同意醫學界所講嘅電子煙同埋新嘅煙嘅產品即係譬如話係熱煙呀即係唔會燒著嘅其實佢哋都會係釋放出一啲有毒嘅物質亦都係損害健康㗎跟住喺呢一方面呢我哋嘅意見係一樣即係都係想保護健康嘅恩賜呢食物衛生局呢就已經係建議去加強管制咁呢個產品都會受到

Fig. 4 Draft transcription with voice typing tool

In addition to the Cantonese interpreted speech, the original English speeches were also transcribed. Together they form our bilingual SI corpus. Transcribing the English speeches was easier, as we could take the written English record available from the LegCo website as a starting point. The speakers mostly followed the written script with occasional improvised changes. With fine listening, the necessary corrections were made and extra information was added accordingly (e.g. indicating hesitations with <uh> and self-repairs with ^^, as shown in the example in Table 1).

5.3 *Parallel Components in the Corpus*

Alongside the floor record of a meeting, there are bilingual written records of proceedings available after each meeting, that is, the Hansard. The portions corresponding to the extracted speeches were collected to form the written corpus in this work. Given that the original speeches were made in English, the Chinese

Table 1 A 2 × 2 alignment of a segment from the T&I corpus

	Written/Tran	Spoken/SI
English	So, we do not have any difference as far as health protection is concerned, and that is why the Food and Health Bureau is proposing to strengthen regulation, such that these products are being regulated, at least on par with the conventional cigarettes for the protection of public health.	So, we <uh> we ^^ we ^^ we don't have any difference as far as <uh> health protection is concerned, and that's why the Food and Health Bureau is proposing to strengthen <uh> regulation, <uh> such that these products are being regulated, <uh> at least on par with the conventional cigarettes for the protection of public health
Chinese/ Cantonese	因此就保障健康而言，兩者並無任何分歧。正因如此，食物及衛生局現正建議加強規管，使這些產品受到至少與傳統香煙看齊的管制，以保障公眾健康。	因此喺呢一方面呢，我哋嘅意見係一樣，即係都係想保護健康嚟，因此呢，食物衛生局呢就已經係建議去加強管制，咁呢一啲嘅產品都會受到規管，最低限度係同傳統嘅煙一樣咁樣受規管呢，去保障香港人嘅健康。

written records are essentially the translation of the corresponding English scripts, hence providing a very important and parallel source for comparing with the interpreting. Although it cannot be ascertained that the interpreters are not also the translators themselves, it is clear that the translation is done after the meetings based on the written verbatim records as source texts, with enough time for more refined lexical choices. Table 1 shows an example with a two-way alignment of a particular segment. In other words, it is a two-way parallel T&I corpus. First, it is parallel with respect to language, between English and Chinese. Second, it is parallel with respect to modality, between written and oral translations. Altogether 60 English speech clips (about 650 min) and their interpreted versions were collected from the council meetings in 2017–2018. The transcription of the interpreted speeches is still in progress. Table 2 shows the amount of parallel data in the various components in the corpus used for the current study, as constrained by the amount of transcribed SI speech. While the corpus is expanding as the transcription goes on, as Shlesinger and Ordan (2012) pointed out, small corpora for interpreting studies could be considered large enough if the phenomenon under review is sufficiently frequent.

Table 2 Size of the T&I corpus for this study

	Written/Tran	Spoken/SI
English (words)	27 K	30 K
Chinese/Cantonese (chars)	43 K	44 K

6 Frequency and Lexical Availability

As mentioned earlier, a parallel T&I corpus will provide much more reliable data for us to compare translation and interpreting and probe their underlying cognitive processes by means of their products. In this section, we will investigate the lexical availability for translators and interpreters, by testing whether the frequency assumption holds in practice. According to the Gravitational Model (Gile 1999), for instance, it has been suggested that interpreters can often only access their active vocabulary under the time pressure, and the more frequently a word is used, the more likely it is gravitated towards an orbit closer to the nucleus. Hence the words more readily retrieved by interpreters are usually more frequent words.

6.1 Sampling

The comparison was based on 50 pairs of samples drawn from the aligned segments, including nouns, verbs and adjectives from the source language. The corresponding lexical choices in the translation and the interpreting were identified to form a sample pair, with their word frequencies retrieved from the Chinese Gigaword 2 Corpus (Traditional), accessed from the Sketch Engine (Kilgarriff et al. 2004).

Although it is assumed that translators and interpreters may differ in their lexical choices as a result of different lexical availabilities, in practice it does not mean that they always end up with a different word in the target language. In fact, most of the time, they may be using exactly the same words, especially for very common words or very specific terminology. At other times they may employ different strategies to handle the translation without any word-for-word equivalence on the surface. So what we really want to check is whether there is a difference in terms of the frequency when they actually use different words. Hence the sampling was done with the following situations excluded:

- When written/spoken or dialectal difference is exhibited between the translated text and the interpreted speech—As the vernacular in Hong Kong, Cantonese is basically a spoken dialect. Although it can be transcribed in written form, some words or expressions are distinguished from the written form in Modern Standard Chinese, which is often the accepted forms in more formal contexts.
- When proper names or technical terms are used in the source text and speech—Since there are often prescribed equivalents for proper names and domain-specific terminology, little difference will be expected between the translator and the interpreter as both are supposed to render the name or the term in the same way.
- When a word in the source language is expressed in a more complicated linguistic unit—The current study aims at word-for-word comparison, and when a source word is rendered in the target language by a linguistic unit above the

word level, the comparison in terms of word frequency will be less straightforward and such cases were thus excluded.

- When a source word does not appear in both the script and the speech, or when it is not rendered in both the translation and interpreting—In such cases, obviously a sample pair cannot be formed. We will discuss our observations on some examples of omissions in Sect. 7.3.2 later.

The various situations are illustrated in Example (1). In all examples hereafter, an aligned segment is presented with the original English written script (EN-W), its Chinese translation (ZH-T), the transcribed English speech (EN-S), and the transcribed simultaneous interpreting in Cantonese (ZH-I). For a given source word in question, any Chinese rendition identified from the example will be accompanied by its pronunciation in Jyutping (a Cantonese Romanisation system proposed by the Linguistic Society of Hong Kong) and a gloss if it is not exactly an equivalent, when it is mentioned for the first time.

(1) EN-W But of course, returning① to the importance② under the international② treaty③, the jurisdictions⑤ should require⑤ that the firm's③ senior③ management④ be responsible⑤ for providing⑤ the necessary⑤ …

ZH-T 但當然，說回在國際條約下的重要性，司法管轄區應要求公司的高層管理人員提供所需的……

EN-S But of course, come ^^ <uh> returning to the <uh> ^^ to the importance <uh> under the international treaty, <uh> the jurisdictions should require that the firm's senior management be responsible for providing the <nece> ^^ necessary …

ZH-I 咁當然啦，返番嚟呢一個國際協議嘅重要性呢，而家呢就話誒銀行嘅高級嘅職員呢……

In Example (1), the potential source words for sampling are underlined and marked as one of the following situations:

① Dialectal difference: 'returning' was translated as 說回 *syut3wui4* in the Hansard but interpreted as 返番嚟 *faan1faan1lai4* in the SI corpus. The latter is obviously spoken Cantonese, so this pair would not be sampled.

② Same lexical choice: 'importance' was rendered as 重要性 *zung6jiu3sing3*, and 'international' as 國際 *gwok3zai3*, in both the translation and the interpreting. They would be ignored.

③ Different lexical choices: 'treaty' was translated as 條約 *tiu4joek3* and interpreted as 協議 *hip3ji5* 'agreement'; 'firm' was translated as 公司 *gung1si1* and interpreted as 銀行 *ngan4hong4* 'bank'; 'senior' was translated as 高層 *gou1ceng4* and interpreted as 高級 *gou1kap1*. These pairs would be sampled.

④ More than a lexical unit: 'management' was translated as 管理人員 *gwun2lei5 jan4jyun4* 'management personnel', which is essentially a compound, and interpreted as a simple word 職員 *zik1jyun4* 'staff member'. These incompatible linguistic units would not be sampled.

⑤ Omission on target side: 'jurisdictions' was translated as 司法管轄區 *si1faat3gwun2hat6keoi1*; 'require' as 要求 *jiu1kau4*; 'providing' as 提供 *tai4gung1*;

‘necessary’ as 所需 *so2seoi1*; and no translation for ‘responsible’. It happens that the speech has been interrupted in the meeting, and all these were missing in the interpreting. These cases would be ignored.

6.2 Results

The hypothesis is straightforward: Where there is a difference in lexical choice between the translator and the interpreter, the latter will tend to use more frequent words than the former. Thus a paired sample *t*-test was done on the word pairs with their word frequencies, with the following null and alternative hypotheses:

- H_0: $\mu_d = 0$
- H_1: $\mu_d > 0$

The difference in the mean frequency of the words chosen by translators and interpreters was found to be statistically significant ($t = 2.66$, $df = 49$, $p < 0.05$). In other words, interpreters do tend to use words of higher frequency when they differ from translators at the lexical level, as shown empirically from our corpus data.

Further analysis was based on our observations of what such a difference in frequency may imply on the actual words being used in the two activities, and several interesting differences were observed.

6.3 Qualitative Characteristics

Hence the paired sample *t*-test has allowed us to verify empirically that when interpreters differ from translators in their lexical choices, they tend to use words that are more frequent. This is in concord with interpreting models describing the lexical access for interpreting, where more frequent words are often more available, and therefore interpreters use them as an immediate response under the time pressure. Nevertheless, what does this difference in frequency really mean? What kinds of qualitative characteristics are behind it? Previous studies have pointed out that lexemes used in interpreting are more informal or unmarked (Shlesinger 2008), as well as less accurate than their translation counterparts (An and Zhang 2014). Is the frequency factor related to formality and accuracy? Specifically, as far as English-Chinese/Cantonese is concerned, what other connections do we find between the frequency effect and the linguistic characteristics? Based on the samples in the statistical test, we observed the following qualitative differences.

6.3.1 Syllabicity or Word Length

The Chinese language is well-known for its four-character idiomatic expressions as a kind of special lexical items. These quadrisyllabic expressions are generally more formal, or of a higher register, than their disyllabic synonyms. It is not difficult to find many examples from our T&I corpus where translators render an English word with a four-character idiom, while interpreters stick to disyllabic words in the same context.

As in Example (2), four-character idioms were used in the translated version, not only for 'stand out', but also for 'unique' that follows in the next utterance. It is interesting that both the translator and the interpreter attempted to make the two sentences more or less parallel. Nevertheless, the subtle differences, if not inaccuracy, in the interpreted speech should be noted. The translated version, with more generous time, allowed for better organisation and followed the order in the source text, rendering 'stand out' as 別樹一幟 *bit6syu6jat1ci3* (literally meaning 'to hoist a flag somewhere else') and 'unique' as 獨一無二 *duk6jat1mou4ji6* (literally meaning 'only one and no two'). On the other hand, the interpreted version has apparently dealt with 'unique' first, rendering it as 獨特 *duk6dak6*, then returned to 'stand out' but did it less accurately with 出色 *ceot1sik1* 'outstanding'.

(2) EN-W Hong Kong does stand out. We are unique.
ZH-T 香港是別樹一幟、獨一無二的。
EN-S Hong Kong does stand out. We are unique.
ZH-I 但係香港呀係獨特, 香港係出色嘅。

In Example (3) below, the adjective 'reasonable' was translated as 合情合理 *hap6cing4hap6lei5* and interpreted as 合理 *hap6lei5*. Obviously, the disyllabic word is much more frequent than the other. Taking the linguistic context into account, it can be seen that the interpreter followed quite straightforwardly the order of adverb and adjective, and added 相當 *soeng1dong1* for 'fairly' before 合理. The translator, on the other hand, has somehow embedded the slight emphasis in the four-character, and higher register, expression.

(3) EN-W Seventeen years ago to be exact, a similar motion was moved, and once only, under fairly reasonable circumstances.
ZH-T 在整整17年前, 政府也曾在合情合理的情況下動議類似的議案, 但只此一次。
EN-S Seventeen years ago, exactly, it was done once, and once only, and under fairly reasonable circumstances.
ZH-I 十七年前做過一次，只係做過一次，而且係^^係相當合理嘅情況下去做㗎。

In fact, apart from being more formal, the longer word lengths may also take up relatively more cognitive resources in speech production, that is, with more syllables to be uttered. As there are a few efforts competing for cognitive capacities at any one time in simultaneous interpreting, it is only natural to use more disyllabic words to accommodate the time pressure.

6.3.2 Polysemy and Generalness

While translators can take their time to be more mindful of their word choices to fit the context as much as possible, interpreters tend to use more general words. Using words that are acceptable in a broader range of context may compromise the precision, but it may also be a safe way to convey the meaning in the source speech, as in Example (4).

(4) EN-W But then Hong Kong's banking system is quite healthy, you would think.
ZH-T 然而, 大家都認為香港的銀行體系十分健全。
EN-S But then Hong Kong's banking <uh> system is quite healthy, you would think.
ZH-I 咁但係香港嘅銀行制度都係誒相當健康啦, 可能你話呀。

The lexicalisation of concepts is seldom uniform across languages, and different sense distinction further complicates the lexical equivalence relations between two languages. According to the Macmillan Dictionary, four senses of 'healthy' are given and defined as follows: (i) physically strong and not ill, (ii) working well and likely to continue to be successful, (iii) a healthy amount of money is a large amount, and (iv) a healthy attitude is good and sensible. Apart from the last two which are defined with respect to specific contexts, the first sense is most general and often relates to physical condition, and the second sense is an extended sense relating more to institutional robustness. The Chinese words 健康 *gin6hong1* 'healthy' and 健全 *gin6cyun4* 'healthy and complete' are near-synonymous, but the former is a more general equivalent to 'healthy' covering almost every sense of it. Their collocation patterns (e.g. the words they often modify) are compared and contrasted with the Sketch Difference function in the Sketch Engine (Kilgarriff et al. 2004), also using the Chinese Gigaword 2 Corpus (Traditional), as shown in Fig. 5.

Words towards the top part are more strongly associated with 健康 and those towards the bottom are more strongly associated with 健全. The former is more used for physical health (e.g. 身體 *san1tai2* 'body', 體魄 *tai2paak3* 'physique'), whereas the latter is more used for abstract systems (e.g. 法制 *faat3zai3* 'legal system', 體系 *tai2hai6* 'system'). In Example (4), the translator has rendered 'system' as 體系 and used 健全 for 'healthy' correspondingly, while the interpreter expressed 'system' as 制度 *zai4dou6* and used the more general word 健康 with it.

A similar case is found in Example (5), for the word 'duties' which was translated as 職責 *zik1zaak3* and interpreted as 責任 *zaak3jam6*. A general sense of

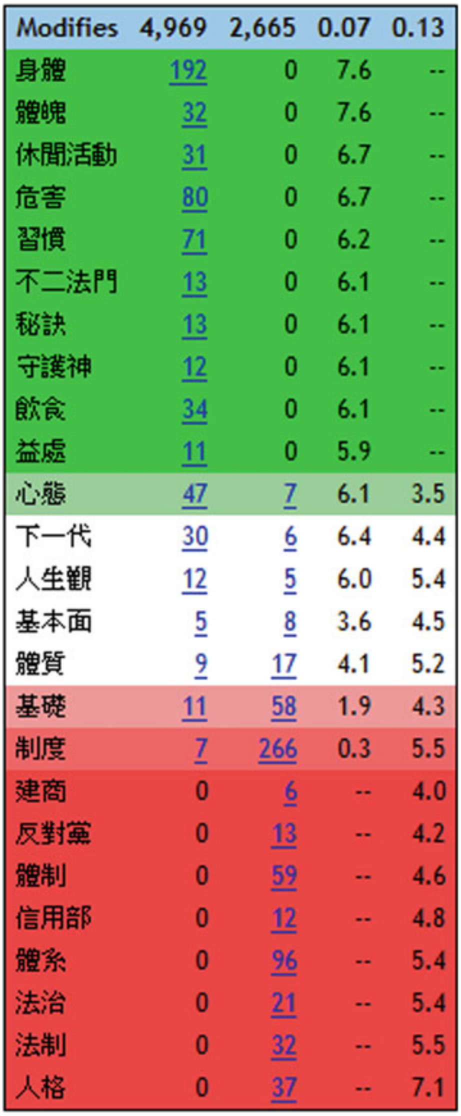


Fig. 5 Sketch difference between 健康 *gin6hong1* and 健全 *gin6cyun4*

Modifies	4,969	2,665	0.07	0.13
身體	192	0	7.6	--
體魄	32	0	7.6	--
休閒活動	31	0	6.7	--
危害	80	0	6.7	--
習慣	71	0	6.2	--
不二法門	13	0	6.1	--
秘訣	13	0	6.1	--
守護神	12	0	6.1	--
飲食	34	0	6.1	--
益處	11	0	5.9	--
心態	47	7	6.1	3.5
下一代	30	6	6.4	4.4
人生觀	12	5	6.0	5.4
基本面	5	8	3.6	4.5
體質	9	17	4.1	5.2
基礎	11	58	1.9	4.3
制度	7	266	0.3	5.5
建商	0	6	--	4.0
反對黨	0	13	--	4.2
體制	0	59	--	4.6
信用部	0	12	--	4.8
體系	0	96	--	5.4
法治	0	21	--	5.4
法制	0	32	--	5.5
人格	0	37	--	7.1

‘duty’ in Macmillan Dictionary is ‘a legal or moral obligation’, with a sub-sense (when used in plural form) as ‘things that you have to do as part of your job’. Thus the translator’s choice is more for the latter, as 職責 more specifically means the duties in a particular job, and 責任 can refer to any kind of responsibility. In other words, the interpreter has taken the more general sense and opted for a more general lexical item.

(5) EN-W However, as legislators, we have to discharge our duties in the best interest of the public, taking a moderate approach instead of going to extremes.
ZH-T 然而, 我們身為議員, 在履行議員職責時, 須以市民的最佳利益為依歸, 採取溫和路線, 而非走向極端。
EN-S However, as legislator, we have to discharge our duties in the best interest of the public, taking a moderate approach instead of going to extremes.
ZH-I 咁但係呢作為立法會議員呢， 我哋就應該係履行我哋嘅責任呢,就係保護香港人嘅公眾利益,我哋應該係有一個比較溫和嘅誒方法, 而唔係話就用啲極端嘅手段。

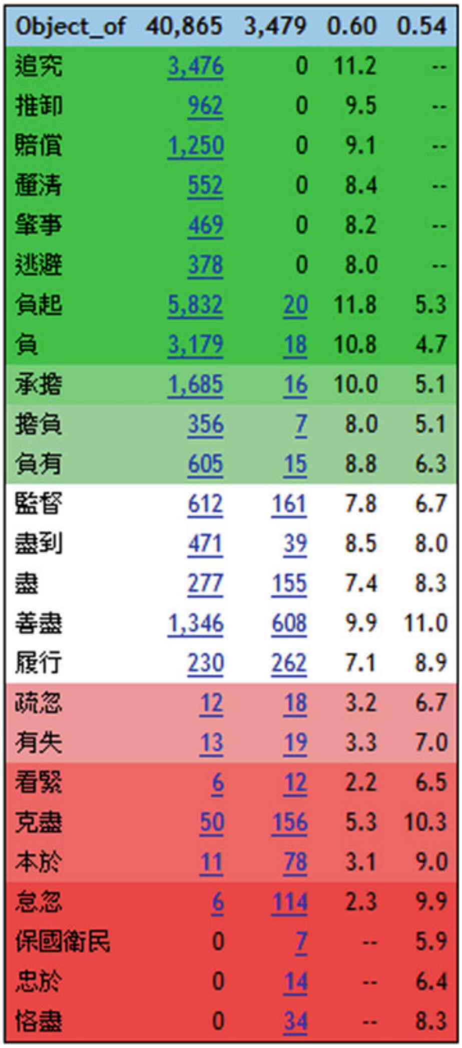
Figure 6 shows the verbs that are strongly associated with 責任 and/or 職責 as their objects. It can be seen that 履行 *lei5hang4* 'discharge' is among those in the middle, indicating that it is used with both nouns. From the frequencies, it may collocate with 職責 more often, and the difference is especially obvious when the collocation frequencies are normalised. Its mutual information score is also higher with 職責 (refer to the last two columns in Fig. 6). On the other hand, 責任 is more general and thus more frequent, and it is more often collocated with other verbs like 負起 *fu6hei2* 'to bear' or 承擔 *sing4daam1* 'to shoulder', which do not take 職責 as object at all.

Example (6) shows the treatment of 'welfare'. In the original utterance, it was conjoined with 'rights', that is, 'rights and welfare'. Similar to Example (2) discussed in Sect. 6.3.1 for 'stand out' and 'unique', some sort of recency effect is also observed here where the interpreter swapped the order and rendered 'welfare' first as 福利 *fuk1lei6*, followed by 權利 *kyun4lei6* for 'rights'. What we want to contrast here is the translator's choice of 福祉 *fuk1zi2* 'well-being' with the interpreter's choice of 福利 for 'welfare'. Obviously the latter is much more frequent, and used more generally. Looking at the Sketch Difference in terms of what they are usually conjoined with, as shown in Fig. 7, apparently 福利 and 權利 tend to co-occur much more often, whereas 福祉 tends to co-occur with more abstract and ideal concepts like 自由 *zi6jau4* 'freedom', 尊嚴 *zyun1jim4* 'dignity', 利益 *lei6jik1* 'benefit', etc.

(6) EN-W I am going to talk about ethnic minority rights and welfare.
ZH-T 我會談論少數族裔的權利及福祉。
EN-S I'm going to talk about <uh> ethnic minority rights and welfare.
ZH-I 我想講講呢就係嗰個少數族裔嘅福利同埋權利。

The examples shown in this section serve to illustrate the cases where there is a choice between some near-synonyms, one with more general and the other with more specific or restrictive sense, the general one is often more readily available to interpreters. Certainly, the more restrictive lexemes are often found in relatively more limited contexts, and are naturally less frequent. They may also be associated with more formal registers, as well as more fine-grained distinction of word senses.

Fig. 6 Sketch difference between 責任 *zaak3jam6* and 職責 *zik1zaak3*

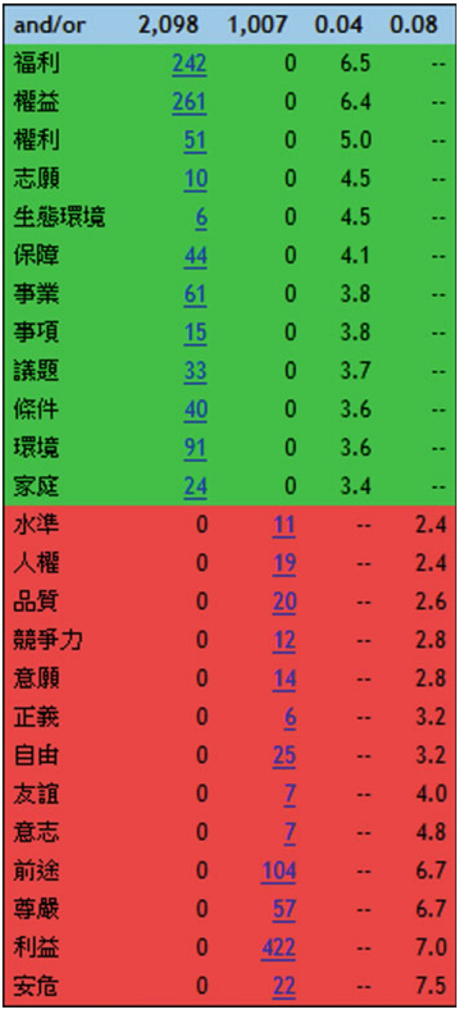


Object_of	40,865	3,479	0.60	0.54
追究	3,476	0	11.2	--
推卸	962	0	9.5	--
賠償	1,250	0	9.1	--
釐清	552	0	8.4	--
肇事	469	0	8.2	--
逃避	378	0	8.0	--
負起	5,832	20	11.8	5.3
負	3,179	18	10.8	4.7
承擔	1,685	16	10.0	5.1
擔負	356	7	8.0	5.1
負有	605	15	8.8	6.3
監督	612	161	7.8	6.7
盡到	471	39	8.5	8.0
盡	277	155	7.4	8.3
善盡	1,346	608	9.9	11.0
履行	230	262	7.1	8.9
疏忽	12	18	3.2	6.7
有失	13	19	3.3	7.0
看緊	6	12	2.2	6.5
克盡	50	156	5.3	10.3
本於	11	78	3.1	9.0
怠忽	6	114	2.3	9.9
保國衛民	0	7	--	5.9
忠於	0	14	--	6.4
恪盡	0	34	--	8.3

6.3.3 Subtle Semantic Differences

Given the word formation mechanisms of Chinese, disyllabic Chinese near-synonyms often contain very subtle semantic differences. When one word in a synonym pair is used in translation and the other is used in interpreting, it does not necessarily mean inaccuracy on any side, but one will be more (im)precise than the other as a result. In this section, we show three examples illustrating a different aspect relevant to the lexical semantic properties found in our sample pairs.

In Example (7), ‘suffered’ is translated as 受累 *sau6leoi6* ‘to get involved (in trouble)’ and interpreted as 受損 *sau6syun2* ‘get damaged’. The meaning of ‘suffer’



and/or	2,098	1,007	0.04	0.08
福利	242	0	6.5	--
權益	261	0	6.4	--
權利	51	0	5.0	--
志願	10	0	4.5	--
生態環境	6	0	4.5	--
保障	44	0	4.1	--
事業	61	0	3.8	--
事項	15	0	3.8	--
議題	33	0	3.7	--
條件	40	0	3.6	--
環境	91	0	3.6	--
家庭	24	0	3.4	--
水準	0	11	--	2.4
人權	0	19	--	2.4
品質	0	20	--	2.6
競爭力	0	12	--	2.8
意願	0	14	--	2.8
正義	0	6	--	3.2
自由	0	25	--	3.2
友誼	0	7	--	4.0
意志	0	7	--	4.8
前途	0	104	--	6.7
尊嚴	0	57	--	6.7
利益	0	422	--	7.0
安危	0	22	--	7.5

Fig. 7 Sketch difference between 福利 *fuk1lei6* and 福祉 *fuk1zi2*

itself is quite vague, as one 'suffers' when one experiences something very unpleasant or painful. Such difficult situations, however, may or may not lead to physical or financial loss. Both Chinese words begin with the morpheme 受 *sau6* 'to endure'. While 受累 simply suggests getting involved in something unpleasant, 受損 embeds the meaning of damage being caused.

(7) EN-W And do not forget that any economic downturn can take place, not just because there is some American economic … sort of negative development taking place over there, and we suffered here.

ZH-T 再者, 別忘記任何經濟衰退均可發生, 不僅是因為美國某項經濟……其他地方出現某些負面發展, 我們這兒便受累。

EN-S And <uh> not to <uh> forget that <uh> our ^^ any economic downturn can take place, not just because, oh, <uh> there's <uh> some <uh> <uh> <uh> American <uh> economic <uh> <uh> <um> sort of negative development taking place over there, and we suffered here.

ZH-I 咁當然啦，我哋唔可以忘記，如果經濟衰退呢..........話因為美國經濟有啲不利發展^^負面發展，咁而令到我哋都受損。

The subtle semantic differences may also relate to the sentiment of individual words. As in Example (8), 'delay' was translated as 延遲 *jin4ci4* and interpreted as 延誤 *jin4ng6* 'delay with loss incurred'. It should be noted that 'delay' is a noun in the source text and speech, and is shifted to a verb in Chinese, in both the translated and interpreted version. The word used in the translation is relatively neutral, with the morphemes 延 *jin4* 'delay' and 遲 *ci4* 'late'. On the contrary, the morpheme 誤 *ng6* 'to incur loss' adds a negative sentiment to the interpreter's word 延誤. As the speaker has said that the delay would not cause any loss, the translator may choose to avoid the negative sense in the translation, and he or she could afford the time to make this refinement.

(8) EN-W The Government will not lose any revenue as a result of the delay in the legislative process for the Stamp Duty (Amendment) Bill 2017 ('the Bill'). The Government has already been collecting all the stamp duties although they have not reached the coffers yet.

ZH-T 《2017年印花稅(修訂)條例草案》("《條例草案》")的立法程序即使延遲，政府也不會因而損失任何稅收。政府一直已在徵收各項印花稅，只是稅款仍未收歸庫房而已。

EN-S And <sec> ^^ and thirdly, they're not losing any money in terms of the <uh> delay, because the government already receiving all the money but it's not in their pocket yet.

ZH-I 第三，佢哋......就算係延誤咗呢條條例草案嘅通過都係唔會有金錢嘅損失，因為政府都係已經係收緊呢啲嘅^^呢啲嘅稅㗎啦。

Example (9) illustrates another issue. There is also a part-of-speech shift in this example, where the adjective 'agreeable' was expressed as the verbs 贊成 *zaan3sing4* 'agree' and 支持 *zi1ci4* 'support' by the translator and the interpreter, respectively. The interesting point here is which of these two near-synonyms is a more precise expression of 'agreeable' in this context. The original English version would suggest merely more willingness to accept the amendment, not necessarily with much enthusiasm. However, 'agree' and 'support' are not exactly synonymous. 'Agree' means to have the same opinion, and 'support' means to agree and give encouragement. In other words, 'support' may entail 'agree', but not the other way round. Hence, in the context of the source text, although 支持 is more frequent and more readily available to the interpreter, apparently 贊成 fits the tone and context more precisely.

(9) EN-WI am actually more agreeable with the amendment that the period should be extended to 12 months, doubling the time.
ZH-T 我其實較贊成修正案的建議, 將限期延長至12個月, 即給予雙倍時間。
EN-S I'm actually <uh> more agreeable with the amendment that <uh> it should be extended to <uh> 12 months, double the time.

ZH-I 我其實呢就想支持嗰個新^^誒延長個期限到十二個月㗎。

It has been observed in past studies that translators tend to use more formal and accurate words than interpreters. From what we have discussed in this section, we find similar support for the formality and accuracy issues. Moreover, we have shown that such differences are often connected to the frequency effect in lexical access as hypothesised in models for simultaneous interpreting. We have found further patterns of the different lexemes used by translators and interpreters with respect to their lexical semantic properties.

7 The Big Picture

Given the complexity of translation, it is nevertheless too simplistic to over-emphasise the significance of the frequency factor to account for the observed differences in lexical choices made by translators and interpreters. First, what we manage to show is when a translator and an interpreter actually opted for a different word to express a certain source word, there is in general a tendency for the interpreter's choice to be a more frequent lexical item. What we have not shown is how often the translators and interpreters actually differ in their lexical choices in practice, and what happens at other times. Second, even when a different word is used, accuracy is not necessarily compromised at the expense of a quick choice. Precision may be less satisfactory but this is often tolerated given the difficulty of simultaneous interpreting. In that case, it raises the question of how we should treat the results of the above comparison. After all, we need to zoom out for the big picture. In this section, we briefly discuss a few other aspects inspired by our corpus examples which could be addressed in future studies.

7.1 Some Words Are Always Available

In practice, the higher availability of more frequent lexical items to interpreters does not exclude other situations where translators and interpreters choose the same words, or at other times when they actually handle a particular word in the source language in different ways in the target language. As described in Sect. 6.1, such cases were outside our sampling criteria. To examine what may happen in those circumstances, we further randomly select 50 samples of nouns, verbs and adjectives each from the source and observe how they are handled by translators and interpreters, respectively.

These samples in fact contain a considerable number of cases where the same target word is used in both the translation and the interpreting. This is especially true for terms that are very frequently used in LegCo debates and relatively common words where the Chinese equivalents are very straightforward. For example, translators and interpreters often unanimously use 解釋 *gaai2sik1* for 'explain', 宣布 *syun1bou3* for 'announce', 改善 *goi2sin6* for 'improve', 關注 *gwaan1zyu3* for 'concerns', 計劃 *gai3waak6* for 'scheme', 受害人 *sau6hoi6jan4* for 'victims', 質素 *zat1sou3* for 'quality', 階段 *gaai1dyun6* for 'stage', 緊逼 *gan2bik1* for 'tight', 成功 *sing4gung1* for 'successful', and even the same four-character word 不偏不倚 *bat1pin1bat1ji2* for 'impartial'. There is basically no controversy in these cases, and apparently the words are quite immediately available. It is also observed that adjectives are obviously weaker in this regard, and we will return to this point in the following discussion.

7.2 Interpretese and Lexical Variety

As measured by previous studies (e.g. Shlesinger 2008), interpretese tends to exhibit even less lexical variety than translationese, while word repetition is more abundant in written and oral translations than in non-translated texts and non-interpreted speeches. Notwithstanding this relative difference, lexical choices made by translators and interpreters are not particularly mechanical. In other words, even for relatively common words with straightforward equivalents, they do not necessarily use the same words every time. For instance, we have shown in Example (4) where the banking system was described as 'healthy', which was rendered as 健康 by the interpreter and 健全 by the translator. Very interestingly, they both had opted for another Chinese word for another occurrence of 'healthy' in the same speech, as in Example (10).

(10) EN-W … the banking system is the core of Hong Kong and we are all responsible for a healthy and modern banking system.

ZH-T … 就是銀行體系是香港的支柱, 而我們所有人均有責任維持穩健而追上時代的銀行體系。

EN-S … <uh> the banking system is the core of Hong Kong and we're all very ^^ we are all responsible for a healthy and modern banking system.

ZH-I … 香港嘅銀行業呢^^銀行體系呢係香港嘅核心嚟嘅，我哋全部人呢都一定係要有責任去維持一個穩定同埋係現代化嘅銀行系統嘅。

Although it was similarly used to describe the banking system, 'healthy' was rendered as 穩健 *wan2gin6* 'stable and healthy' by the translator and 穩定 *wan2ding6* 'stable' by the interpreter. The former is composed with the morphemes 穩 'stable' and 健 'healthy', while the latter may really be more equivalent to 'stable' intuitively, as both morphemes 穩 'stable/steady' and 定 'stable/calm' are near-synonymous. Moreover, the translator's choice 穩健 is less frequent than both 健全 and 健康 in the previous example, whereas the interpreter's choice 穩定 is more frequent than both.

Hence, for the same source word, even when it is used in a similar context, a different target word may be chosen. This is not easy to account for as far as lexical access is concerned. While one could afford more time in translation to evaluate several candidate words, it is not expected to happen in simultaneous interpreting so often, if interpreters are supposed to only select the most available word within an extremely short time. Although the Gravitational Model suggests that lexical availability varies constantly, the lexical variation in response to a similar linguistic context remains to be explained. Given the distinctiveness of lexical repetition as a feature of translationese and interpretese, more in-depth analysis is needed to delve into its dynamics more thoroughly.

7.3 Beyond the Lexical Boundary

There were cases where the translation and interpreting products may not provide directly comparable samples for simple lexical comparison in our analysis above. In this section, we discuss two kinds of scenarios observed from the data.

7.3.1 Formal Equivalence Versus Idiomaticity

Earlier in Sect. 6.3.1, we have discussed one qualitative difference behind the frequency effect, in terms of the use of quadrisyllabic expressions in translation and disyllabic words in interpreting. It was also suggested that those four-character words are often of higher register than their disyllabic near-synonyms. Nevertheless, it should also be noted that the claim may only apply to equivalence limited at the lexical level. On the one hand, four-character expressions are not only available to translators. From time to time, they are also commonly used by interpreters. On the other hand, in those cases, it is often found that interpreters have not confined themselves to individual words in the source speech, especially if it is likely to hinder intelligibility when they do.

As in Example (11), 'agree to disagree' is more or less a fixed phrase, although its meaning can still be figured out from the surface. The four-character word used by the interpreter, 求同存異 *kau4tung4cyun4ji6*, is obviously a concise and accurate expression for it. On the other hand, apparently the translator has chosen to preserve the surface form as much as possible, expressing the phrase in the form of a definition: 認同彼此之間可存有異議 *jing6tung4 bei2ci2 zi1gaan1 ho2 cyun4-jau5 ji6ji5* 'accept that different opinions may exist between each other'. Faithful translation like this, at the expense of idiomaticity, is not rare.

(11) EN-W Very often, we have to agree to disagree.
ZH-T 許多時候, 我們須得認同彼此之間可存有異議。
EN-S Very often, we have to agree to disagree.
ZH-I 好多時候呢, 我哋需要呢係求同存異。

Example (12) demonstrates a similar phenomenon, where the translator tends to follow the surface form of the source text, even when the output may not be very natural in that particular context. For instance, the translation adheres to the original structure 'to [V-inf] for the sake of [V-ing]' and renders it with a word and a synonymous phrase for the two occurrences of 'upset', namely, 觸怒 *zuk1nou6* and 惹它生氣 *je5taa1sang1hei3*, both meaning to make someone/something angry. Instead of such a literal reading, the interpreter's rendition with the four-character word 舉步維艱 *geoi2bou6wai4gaan1* (literally meaning 'very difficult to move a step') seems to make more sense. According to the Macmillan Dictionary, other than 'to make someone feel sad, worried or angry', 'upset' has another sense of 'to make something stop working in the normal way'. The latter interpretation seems to be more suitable than making the government angry.

(12) EN-W … monitoring the Government's work does not mean we have to upset the Government merely for the sake of upsetting it.
ZH-T 監察政府的工作亦不表示我們須單純為觸怒政府而惹它生氣。
EN-S … monitoring the Government's work does not mean to have an ^^ to have to upset the Government for the sake of doing it.

ZH-I 我哋應該監察政府嘅工作，但係呢個唔代表呢我哋係需要係令到呢格誒^^呢一個誒政府呢舉步維艱。

Hence the above examples raise another issue. While lexical availability determines interpreters' word uses to a certain extent, we also need to consider how the dynamics might be altered when problems of intelligibility and idiomaticity emerge in the process.

7.3.2 Context-Dependent Interpretation and Omission

In Sect. 7.1 above, we have mentioned that many a time translators and interpreters actually made the same lexical choices, but this happened less to adjectives than nouns and verbs. A possible reason is that the meaning of individual adjectives could be relatively fluid, depending on the actual context in which they are used, that is, what they are modifying. Although one might find context-free equivalents for them in bilingual dictionaries, it is even more important to render them in a context-sensitive way. Let us refer to two examples below, both for the adjective 'unsatisfactory'.

(13) EN-W Although, as I have said, the authorities may consider this a starting point, it is unsatisfactory.
ZH-T 雖然正如我說，有關當局可視這個指數為一個起點，但這並不是令人滿意的做法。
EN-S Although, as I've said, you may want to do it as a starting point, but it is un- <uh> <s> satisfactory.
ZH-I 雖然我講過啦，可能啦你可以^^可以用佢做一個起點，咁但係就唔^^唔足夠嘅。

In Example (13), 'unsatisfactory' was translated as 不是令人滿意的做法 *bat1si6 ling6 jan4 mun5ji3 dik1 zou6faat3* 'not a satisfactory method' and interpreted as 唔足夠 *m4zuk1gau3* 'not enough'. Note that both ways have already gone beyond the lexical level to render the original English adjective. Similarly, in Example (14), the same word was rendered as 不理想 *bat1lei5soeng2* 'not ideal' and 唔好 *m4hou2* 'not good', respectively.

(14) EN-W Yet the average waiting time is eight weeks, which is highly unsatisfactory.
ZH-T 不過, 輪候時間平均是8個星期, 這也是極不理想的。
EN-S … with an average eight-week waiting time, and this is highly unsatisfactory.

ZH-I 一般嚟講呢個等待^^候嘅時間呢就係八個星期，就非常之唔^^唔好㗎。

Quite obviously, in the latter example, the interpreter may have been experiencing difficulty in accessing a specific word to fit the particular context where the long waiting time was described as ‘unsatisfactory’, and at the end came up with a most general expression equivalent to ‘not good’. In more extreme cases, adjectives might be omitted altogether. As in Example (15), the adjective ‘reputable’ was omitted in the interpreted version.

(15) EN-W The person may either defend himself or get a reputable criminal lawyer if he has financial resources or if he has obtained legal aid.
ZH-T 有關人士可自行抗辯; 又或如果他們有財政資源或已取得法援, 便可聘用聲譽良好的刑事律師。
EN-S Either you do your defence by yourself, or, if you have enough financial resources, you get a very reputable criminal lawyer, or the legal aid.
ZH-I 咁你一係呢就你自己去抗辯, 又或者呢你足夠嘅經濟嘅能力呢, 就係搵一個嘅刑事律師去幫你, 又或者呢去申請法援。

Shlesinger (2008) compared the part-of-speech distributions in written and oral English-Hebrew translations, and found that adjectives ranked second in both, which she attributed to ‘an artifact of the source text, which involved particularly long and frequent strings of modifiers’ (Shlesinger 2008: 247). It was nevertheless noted that the relative frequency of adjectives was actually lower in oral translation (15.3%) than written translation (19.1%). The relatively bigger drop in adjectives compared to nouns, and particularly in contrast to the increase in verbs as well as other classes of function words, seems to indicate the vulnerability of adjectives to omissions in the interpreting process. As we suggest here, it may be related to the context-sensitive interpretation of the meanings of adjectives, which may affect the efficiency and effectiveness of lexical access during interpreting.

7.4 Implications on Lexical Access and Interpreter Training

The human mental lexicon is often modelled as a massive interconnected network of words with the nodes linked by a variety of associative relations of different strengths (Aitchison 2003). While lexical access may be affected by many factors, including word frequency, concreteness, polysemy, etc., in the context of translation and interpreting we must also consider the very important cross-linguistic factors. On the one hand, the complex network of words contains not only paradigmatic and syntagmatic links, but also topical and other subjective associations. On the other hand, when it comes to cross-lingual word association, the situation could be much more complicated, especially for distant language pairs like English and Chinese. First, given the morphological difference, as well as the non-uniform lexicalisation of concepts, complete equivalents are rare. Second, the majority of partial equivalents between two languages thus make a huge amount of many-to-many mappings. Third, the cultural specific concepts or others without direct equivalents will have to be resolved by composing two or more words in the target language following specific linguistic rules.

As what we have discussed in this study so far, there are at least two aspects of the differences in translators' and interpreters' lexical choices which are worth further consideration. First, when translators and interpreters opt for different lexical items, the interpreters' choices are often more frequent words, with which formality and precision may be compromised. Second, when a source word does not have an exact lexical equivalent in the target language or requires to be comprehended with its context closely, translators can have more time to figure out an appropriate expression. Interpreters only have limited time to manoeuvre and may come up with a very general rendition with the broadest sense, or may even omit it at the end. Putting these in the context of lexical access and the availability of linguistic constituents during interpreting, two questions are particularly pertinent:

- Although the more frequent words, which also tend to be less formal and less precise, are often most available, are there any ways to push the second or third available lexical items to the front, especially under specific contexts where they are obviously a more satisfactory lexical choice?
- For cases that require context-sensitive renditions to be natural and intelligible, how can we enhance the flexibility of translators and interpreters and arouse their awareness of the many possibilities in rendering the same source word in connection with a variety of contexts?

Translators may often resort to dictionaries and various computer-aided translation tools. It may be feasible to enhance their lexical access from these lexical resources. These may include additional navigational routes and providing well categorised examples for them to consider the many possibilities available to them (Kwong 2018). For interpreters, however, drilling with appropriate materials may be a more practical way. For example, guiding interpreters to compare specific word choices of translators and interpreters may draw their attention to the real observed differences in practice. Exercises to elicit more formal and precise near-synonyms given a particular source word may also be useful, especially if interpreters are expected to work in a relevant domain at a certain level of formality.

The bilingual parallel T&I corpus presented in this study could be taken as a starting point to provide a sample of stimulation materials for the recommended exercises. Given that the dynamics of lexical availability vary with the task demands, the time constraints, and the cognitive resources, it is advisable to make use of authentic and empirical data for task-oriented approaches to enhance the access of our mental lexicon. The empirical data are also useful for computational modelling of word associations applicable to translation and interpreting, as well as further psycholinguistic investigation of the different lexical access mechanisms of translators and interpreters.

8 Conclusion

Unlike what previous corpus-based translation and interpreting studies have often done to sort out the characteristics or universals of translationese and interpretese from specific quantitative measures like token-type ratio and part-of-speech distributions from corpora, in this study we have tried to compare translation and interpreting more microscopically, using a parallel T&I corpus, and from a new perspective. Based on samples from the corpus, we found empirical and statistical support for the frequency effect on translators' and interpreters' lexical choices, thus the models regarding lexical availability for simultaneous interpreting. In addition, we have been able to find out more qualitative differences in association with the frequency effect, as well as to observe more thoroughly the translators' and interpreters' linguistic behaviour to probe the underlying lexical processing. Although we have only taken lexical access and the frequency effect as the point of departure in this study, there remain many other concurrently operating factors to be studied. To this end, our parallel T&I corpus will have much to offer for further research, especially for us to understand how they collectively affect the processes and shape the products of translation and interpreting.

Acknowledgements The work reported in this chapter was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 14616317).

References

AIIC. 1982. *Practical guide for professional interpreters*. Geneva: AIIC.

Aitchison, Jean. 2003. *Words in the mind: An introduction to the mental Lexicon*. Blackwell Publishers.

An, Xue 安雪, and Ling Zhang 張凌. 2014. 朱鎔基答中外記者問口筆譯對比研究. *Modern Chinese* 《現代語文》2014 (3): 127–130.

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, and Elena Tognini-Bonelli. Amsterdam/Philadelphia, John Benjamins.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7 (2): 223–243.

Bendazzoli, Claudio, and Annalisa Sandrelli. 2005. An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In *Proceedings of MuTra 2005—challenges of multidimensional translation*, 149–160. Saarbrücken.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Chesterman, Andrew. 2004. Paradigm problems? In *Translation research and interpreting research: Traditions, gaps and synergies*, ed. Christina Schäffner, 52–56. Clevedon/Buffalo/Toronto: Multilingual Matters.

Christoffels, Ingrid K., Annette M.B. de Groot, and Judith F. Kroll. 2006. Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language* 54: 324–345.

Gile, Daniel. 1999. Testing the effort models' tightrope hypothesis in simultaneous interpreting—A contribution. *Hermes* 23: 153–172.

Gile, Daniel. 2009. *Basic concepts and models for interpreter and translator training*. Revised. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Guy, Aston. 2018. Acquiring the language of interpreters: A corpus-based approach. In *Making Way in Corpus-based Interpreting Studies*, ed. M. Russo et al. New Frontier in Translation Studies, Springer.

Hu, Kai-bao, and Qing Tao. 2010. The compilation and application of Chinese-English conference interpreting corpus. *Chinese Translators Journal* 2010 (5): 49–56.

Hu, Kai-bao, and Li-xin Xie. 2014. The use of English relative clauses in Chinese-English conference interpreting. *Shandong Foreign Language Teaching Journal* 2014 (4): 8–16.

Ilg, Gérard. 1978. L'apprentissage de l'interprétation simultanée de l'allemand vers le français. *Parallèles n°1, Cahiers de l'E.T.I.*, 69–99. Université de Genève.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London/New York: Longman.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX 2004*, Lorient, France.

Kwong, Oi Yee. 2016. There is something about grammatical category in Chinese visual word recognition. *Journal of Psycholinguistic Research* 45: 1067–1087.

Kwong, Oi Yee. 2018. From fidelity to fluency: Natural language processing for translator training. In *Proceedings of the 5th workshop on natural language processing techniques for educational applications*, 130–134. Melbourne, Australia.

Laviosa, Sara. 2012. Corpora and translation studies. In *Corpus applications in applied linguistics*, ed. Ken Hyland, Chau Meng Huat, and Michael Handford. London: Continuum.

Liu, Minhua, Diane L. Schallert, and Patrick J. Carroll. 2004. Working memory and expertise in simultaneous interpreting. *Interpreting* 6 (1): 19–42.

Pan, Feng. 2014. A corpus-based study of the application of hedges in Chinese-English conference interpreting. *Shandong Foreign Language Teaching Journal* 2014 (4): 24–29.

Santos, Diana. 2004. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems*. Amsterdam: Rodopi.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43 (4): 486–493.

Shlesinger, Miriam. 2008. Towards a definition in Interpretese: An intermodal, corpus-based study. In *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile*, ed. Gyde hansen, Andrew Chesterman, and Heidrun Gerzymisch-Arbogast. John Benjamins Publishing Company.

Shlesinger, Miriam, and Noam Ordan. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24 (1): 43–60.

Stavrakaki, Stavroula, Kalliopi Megari, Mary H. Kosmidis, Maria Apostolidou, and Eleni Takou. 2012. Working memory and verbal fluency in simultaneous interpreters. *Journal of Clinical and Experimental Neuropsychology* 34 (6): 624–633.

Tohyama, Hitomi, and Shigeki Matsubara. 2006. Development of web-based teaching material for simultaneous interpreting learners using Bilingual Speech Corpus. In *Proceedings of world conference on educational multimedia, hypermedia & telecommunications*, 2906–2911. Orlando, Florida.

Wang, Binhua, and Bing Zou. 2018. Exploring language specificity as a variable in Chinese-English interpreting: A corpus-based investigation. In *Making way in corpus-based interpreting studies*, ed. M. Russo et al. New Frontier in Translation Studies, Springer.

Zhang, Wei. 2013. Corpus-related interpreting studies: Principles and approaches. *Technology Enhanced Foreign Language Education* 149: 63–68.

Zou, Bing, and Bin-Hua Wang. 2014. Transcription and annotation of paralinguistic information in interpreting corpora: The status Quo, problems and solutions. *Shandong Foreign Language Teaching Journal* 2014 (4): 17–23.

Oi Yee Kwong is former Associate Professor in the Department of Translation of The Chinese University of Hong Kong. Her research interests span many bilingual and pan-Chinese language processing issues in computational linguistics, corpus linguistics and psycholinguistics, for which she has worked and published on corpus development and annotation, lexical semantics, translation lexicons, name transliteration, mental lexicon, sentiment analysis and discourse processing. She authored the monograph New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation (2012). Her recent research focuses on language resources for computer-aided translation and corpus-based translation studies.

Translationese and Register Variation in English-To-Russian Professional Translation

Maria Kunilovskaya and Gloria Corpas Pastor

Abstract This study explores the impact of register on the properties of translations. We compare sources, translations and non-translated reference texts to describe the linguistic specificity of translations common and unique between four registers. Our approach includes bottom-up identification of translationese effects that can be used to define translations in relation to contrastive properties of each register. The analysis is based on an extended set of features that reflect morphological, syntactic and text-level characteristics of translations. We also experiment with lexis-based features from n-gram language models estimated on large bodies of originally- authored texts from the included registers. Our parallel corpora are built from published English-to-Russian professional translations of general domain mass-media texts, popular-scientific books, fiction and analytical texts on political and economic news. The number of observations and the data sizes for parallel and reference components are comparable within each register and range from 166 (fiction) to 525 (media) text pairs; from 300,000 to 1 million tokens. Methodologically, the research relies on a series of supervised and unsupervised machine learning techniques, including those that facilitate visual data exploration. We learn a number of text classification models and study their performance to assess our hypotheses. Further on, we analyse the usefulness of the features for these classifications to detect the best translationese indicators in each register. The multivariate analysis via text classification is complemented by univariate statistical analysis which helps to explain the observed deviation of translated registers through a number of translationese effects and detect the features that contribute to them. Our results demonstrate that each register generates a unique form of translationese that can be only partially explained by cross-linguistic factors. Translated registers differ in the amount and type of prevalent translationese. The

M. Kunilovskaya (✉) · G. Corpas Pastor
Research Group in Computational Linguistics, University of Wolverhampton,
Wolverhampton, UK
e-mail: maria.kunilovskaya@wlv.ac.uk

G. Corpas Pastor
University of Malaga, Malaga, Spain

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_6

same translationese tendencies in different registers are manifested through different features. In particular, the notorious shining-through effect is more noticeable in general media texts and news commentary and is less prominent in fiction.

Keywords Parallel corpora · Register variation · Translationese trends · Translationese indicators · Machine learning

1 Motivation and Aim

In this chapter we explore and compare translationese effects across several registers in English-to-Russian translation. This research builds on the long-established assumption that the intralinguistic variation between registers can be greater than the cross-linguistic differences between the same registers, famously demonstrated by Biber (1999). We also assume that the cross-linguistic differences are one of the major factors that shape the linguistic make-up of translations. The configuration of differences and similarities between the source language (SL) and the target language (TL) creates a unique language gap in each register and underlies the shining-through effect (Teich 2003) or interference, i.e. the tendency of translated texts to follow the SL patterns rather than conform to the regularities of the TL. Based on these assumptions, we are interested in establishing how the cross-linguistic distance between registers plays out with respect to the properties of translated texts in these registers.

It is especially interesting because the features used in this research to distinguish translations from the originally-authored texts in the target language (also referred to as non-translations or reference texts) are partly inspired by the variational linguistics studies that compare registers (Biber 1988; Katinskaya and Sharoff 2015; Neumann 2013; Nini 2015).

Besides variational studies, our feature selection and engineering process were guided by the previous translationese studies and evidence from the empirical translation studies, especially those that relied on interpretable (rather than surface) linguistic features to describe the typical deviations from TL norm observed in translations. Briefly, we use two feature sets: (i) frequencies of a number of morphosyntactic categories extracted from Universal Dependencies (UD) annotations and (ii) lexical frequency features that reflect the differences in the distribution of n-grams in translated and non-translated language (a detailed description of features is offered in Sect. 3.1; the description of the morphosyntactic features is offered in Appendix).

Typical translationese features for English-to-Russian translation include the overuse of relative clauses, copula verbs, modal predicates, analytical passives, generic nouns and all types of pronouns as shown below. Probably, none of the translation in the examples can be considered ungrammatical in Russian, but there is a Master Yoda-style foreign sound to them. Note that the back translations may come across as perfectly acceptable sentences, because the translations are very literal in the first place. All examples are real-life student translations from Russian Learner Translator corpus (Kutuzov and Kunilovskaya 2014).[1]

(1) Necklaces, at first as pectorals that covered the whole chest, evolved from the prehistoric pendants. Ожерелье—первое нагрудное украшение, **которое** занимало место на всей груди, **которое** стало основой для подвесок [Necklace—first chest decoration**, which** covered the whole chest**, which** became the basis for pendants].

(2) …there are many self-employed people who manage to get money from others by means of falsely pretending to provide them with some benefit or service… Более того, **есть** много **людей**, работающих на себя, **которые** получают деньги обманным путем [Moreover, many **people are**, working for themselves**, who** get the money in a deceitful way].

(3) …differences in self-efficacy may simply mean that some teachers struggle to identify solutions to problems beyond their circle of control. …разница в самооценке **может означать** лишь **то, что** некоторые учителя испытывают сложности в нахождении решений задач **за пределами того, чем** они **могут управлять** […difference in self-evaluation **can** mean only **that** some teachers run into difficulties in finding solutions to tasks beyond the scope of **that what** they **can** control].

(4) It was difficult and exhausting to see. **Это было** тяжело и утомляюще пытаться видеть. [**It was** hard and exhausting to try to see].

These examples demonstrate a number of translation solutions that explain the increase in the frequency of TL items that are less frequent in non-translated TL than their literal counterparts in the SL. In example (2) the generic noun ‘people’ is rendered with a less frequent literal ‘люди’, instead of using a structure with zero subject or other more acceptable ways of expressing unspecified subjects. English and Russian have contrastive ways of expressing subjective modality: modal verbs are a less common choice in non-translated Russian, which prefers parenthetical means of expressing modality. The translation solution in (3) carries over the typical English modal predicate. Example (4) has the notorious literal renderings of the structures with the introductory it, which contributes to the boost of pronouns and copula verbs in translated Russian. Besides, such renditions have a strange word order, which usually interferes with the smooth flow of information in the text. Another source of surplus function words, including pronouns is the tendency to unpack the information from various concise English structures using strings of relative clauses, instead of repackaging the information in a more natural way (see (1) and (3)). Finally, example (2) demonstrates the tendency towards the explicit use of copula verbs in contexts, where a zero copula is typical in Russian.

The overarching goal of this research is to reveal and describe the register-related specificity of English-to-Russian translations in four registers.

[1] https://www.rus-ltc.org/search.

To achieve this goal, we complete several steps and answer the following research questions:

1. How clear are the register distinctions between the translated registers compared to non-translations for the two feature sets tested, provided that the suggested features reliably distinguish registers in originally-authored Russian? If the register distinctions are diluted in translations, the standardisation hypothesis stands.
2. Do registers share translationese indicators, i.e. are there translationese indicators that cut across all registers, provided that we are able to distinguish between translations and non-translations using our features?
3. What are the most important translationese indicators and most prominent translationese trends based on the results of multivariate and univariate analyses in each register?
4. Do the top translationese indicators intersect with the major cross-linguistic differences between the same registers in English and Russian to demonstrate that interference is the most important translationese effect?

These research questions are relevant to the development of the translationese theories and methodologies. The robustness of translationese indicators across registers has to be considered while building translationese detection applications. The register-induced specificity of translations has to be taken into consideration in any translation quality estimation systems based on translationese features.

In what follows, we discuss the theoretical implications of the previous translationese and variational linguistics studies for the current research and define our key concepts (Sect. 2). Section 3describes our research data and the linguistic resources used for language modelling; it also has the description of our methods and experimental setup, starting with the feature sets. The results as per the research questions are presented and commented in Sect. 4, which is followed by their interpretation in Sect. 5. Section 6 summarises the research and outlines future work.

2 Theoretical Background

2.1 Key Concepts and Approaches

The theoretical underpinnings for this research come from translationese studies, a research direction that investigates the peculiarities of translated texts that distinguish them from non-translations. This research field is related to the tasks of testing translationese universals, translationese detection, translation direction detections (including SL identification both for human and machine translation (MT)) as well as more recent studies of translationese variation along a number of dimensions such as translation competence, quality, direction, method, etc. In our necessarily sketchy discussion of the developments in this well-established research area below, we highlight the aspects that are most relevant for the current project.

What is 'translationese'. The foundations of this type of studies were laid by Gellerstam (1986), to whom they attribute the introduction of the term 'translationese'. Gellerstam has demonstrated that there were significant statistical differences in the frequencies of loan words and colloquialisms, among other lexical features, between translated and non-translated Swedish texts. Originally, the term was used to denote statistical deviations of the translated language from the expected target language norm manifested in a reference corpus. Diana Santos (1995) extended the lexical translationese findings to include morphological phenomena such as diverging frequencies of tense and aspect forms in English and Portuguese. Her research was based on a small bidirectional parallel corpus, which provided enough occurrences of the targeted grammatical items for manual analysis. Importantly, her research design gave access to the source text and helped to link the unusual frequencies of grammatical items to the influence of the source text. We will highlight that her understanding of translationese was limited to 'the influence of properties of the source language in a translated text in a target language' (Santos 1995: 61). Her work is relevant for this research because it explicitly mentions the impact of the distance between the languages on the properties of translations. In particular, the author hypothesises that the closer the languages, the more probability of translationese due to the ease of levelling-out the differences between them.

The term translationese is sometimes used metonymically to denote any translated material (see, Nikolaev et al. 2020; Stymne 2017, for example) or to refer to the specificity of translations induced by the SL in opposition to SL/TL-independent properties of translations known as translation universals (see Rabadán et al. 2009; Santos 1995). For the purposes of this project, translationese is defined as a property of being a translation, based on the *statistical differences in frequencies of language items between translations and non-translations in the TL regardless of their hypothesised cause, which mark translations as its own language variety*.

Main translationese effects: Shining-through and independent translationese. Important developments in the descriptive approach to translations are associated with Gideon Toury's *laws of translation* (1995) and Mona Baker's *translation universals hypotheses* (1993). To put it briefly, the former generalised the observations on the properties of translations as two major laws: the law of increasing standardisation, and the law of interference from the source text. Mona Baker's theory suggested that there are universal tendencies in translation that are independent of the source and target languages. Baker's famous definition of the universal features of translation runs as follows: 'features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems' (Baker 1993: 243). Her initial set of hypothesised universals (among the most-tested items) included *explicitation*, i.e. the tendency to spell things out rather than leave them implicit; *simplification*, i.e. the tendency to disambiguate and to avoid any risks of misunderstanding by making texts simpler lexically and structurally; *conventionalisation* (also known as standardisation or levelling-out), i.e. the tendency for translations to exhibit relatively higher level of homogeneity than their sources; *normalisation*, i.e. the tendency to exaggerate features of the TL and to conform to its typical patterns.

The subsequent empiric research into translation universals did not corroborate the initial 'universal' claims for the proposed hypotheses. The results on a variety of translated domains, registers, language pairs and translation varieties were mixed and contradictory. To give some examples, Corpas Pastor et al. (2008) confirmed simplification for some features associated with this trend, but not for the others. Kruger and van Rooy reported limited support for the 'more explicit, more conservative, and simplified language use in the translation corpus' (Kruger and van Rooy 2010: 26).

This is not surprising for three major reasons: (1) the mapping of particular features into descriptive translationese trends can be a matter of debate (as stated in Zanettin 2013: 25); (2) there can be differences in the extraction procedures; (3) translations from different SLs and in different registers produce diverging translationese patterns. To demonstrate some of these factors consider the findings about connectives (also referred to as discourse markers, cohesive markers or conjunctions). Corpas Pastor et al. (2008) expected fewer discourse markers in translations of medical and technical texts from English into Spanish as a sign of simplification, and indeed found that 'non-translated texts use discourse markers significantly more often' in two out of three corpus pairs (Corpas Pastor et al. 2008: 24). At the same time, Koppel and Ordan (2011), while testing on English translations of addresses given in the European Parliament (Europarl) in five other languages, reported that discourse markers were significantly more frequent in translations than in the originally-authored English texts. They were inclined to interpret it as an indication of explicitation. Generally, the increase in the frequencies of discourse markers in translated language and higher cohesiveness of translations is a relatively well-explored translationese phenomenon. However, its interpretation as a manifestation of explicitation, normalisation or SL interference varies across language pairs and text categories or is unclear in some experimental setups (Castagnoli 2009; Kunilovskaya 2017; Olohan 2001). It is especially confusing if connectives are treated individually rather than cumulatively. In Jiang and Tao (2017) the frequencies of individual discourse markers were traced to the corresponding SL items to demonstrate that they contribute to several translation universals. Similarly, Becher insisted that 'every explicitating and implicitating shift has a distinct cause' and needs to be treated on a case-to-case basis (Becher 2011: 215).

In this research we refrain from assigning individual features (indicators) to the trends such as simplification and explicitation a priori. Instead, we follow a bottom-up approach and identify the indicators of some *translationese effects* based on the similarity of their frequency pattern in the source texts (ST), target texts (TT) and reference texts (see Sect. 3.3 for the categorisation of features as contributing to different translationese effects).

The two interpretations of the nature of translations given by Toury and by Baker are complementary and can be seen to represent two major types of translationese. To avoid unnecessary associations with the foreign language acquisition terminology, we would use Elke Teich's term *shining-through* to refer to the cases where the cross-linguistically diverging frequencies of the features are adapted in translations to the SL values, giving rise to significant distinctions between translations and non-translations (Teich 2003). This is the 'interference' type of translationese, which is considered the major factor in shaping the properties of translations (see evidence in Evert and Neumann 2017; Volansky et al. 2015, for example). The features of translations that significantly deviate from both SL and TL, where there are no cross-linguistic differences between non-translations (English source texts and originally-authored Russian texts in our setup), should be considered cases of true *language-pair-independent translationese* in line with Baker's ideas. Some features that spot language contrast can be fully adapted to the TL norm (*adaptation*) or even exaggerate the TL properties (*over-normalisation* or russification in our setup).

Methodological paradigms in translationese studies (features, data and analytical approaches). Over the last few decades, translationese studies as an area of research within translation studies has seen significant developments in the research methods. The earlier investigations were often based on manual extraction of a few features from limited corpus data (sometimes lacking the parallel component) and relied on univariate statistic analysis (Becher 2011; Castagnoli et al. 2011; Nakamura 2007; Puurtinen 2003; Santos 1995). The more recent projects are computationally intensive and involve massive parallel and comparable corpus resources in several language pairs and complex research designs with extensive and elaborate feature sets and methods (see, for example, Dipper, Seiss, and Zinsmeister (2012) who describe the typical corpus resources setup in translationese studies and Evert and Neumann (2017) for the multivariate analysis and feature engineering methodology).

A *machine learning (ML) turn* in the translationese research began with the ground-breaking work by Baroni and Bernardini (2006) who convincingly demonstrated that translations of geopolitical texts into Italian are inherently different from the comparable non-translations by employing a Support Vector Machines (SVM) algorithm to classify them. They experimented with various types of n-grams to represent texts and discovered that bigrams performed best. An important message from their experiments was that a ML algorithm was able to reliably pick the difference between translations and non-translations even when the human subjects (professional translators) were unable to do so as effectively. It brought about a new strand of research known as translationese detection. ML algorithms were used to test the hypothesis about various translationese properties. A good example of this methodology in action is Koppel and Ordan (2011), who reported a series of ML experiments on the Europarl corpus and confirmed that source language plays a crucial role in the make-up of a translated text. They used frequencies of 300 function words as features (which excludes any cultural or topic differences between the corpora). Probably, the most impressive results were reported by Popescu (2011) who reported 99.53% cross-validation accuracy in the task of detecting translations on character string features for an SVM classifier trained on literary translations from French and German into English. However, when they tested a model trained on out-of-French translations on out-of-German translations they received the results at the chance level—an indication that character n-grams capture uninteresting SL-related cues such as proper names. Filtering out those items led to the realistically moderate results of 77.08% in the experiment where they trained on translations from French and books by British authors for reference and testing on translations from German and American fiction for non-translated reference.

In Ilisei et al (2010), a supervised learning approach was employed to identify the most informative features that characterised translations compared to non-translated texts. The learning system was trained on two domains, medical and technical. The novelty of their approach consisted of its language-independent data representation. On the categorisation task, the algorithms achieved an accuracy of 87.16% on a test set and reached up to 97.62% for separate test datasets from the technical domain. The removal of the features, linked by the authors to simplification, from the machine learning process led to decreased accuracy of the classifiers. Therefore, the retrieved results were interpreted as an argument for the existence of the simplification universal.

The book by Gloria Corpas presents the results of several NLP experiments to study translation universals and translationese features. Corpas focuses on three universals: simplification, convergence and transfer (shining-through). Vectors of lexical and syntactic features are used to test various corpora of English and Spanish: (a) a large corpus of Peninsular Spanish (reference corpus of 50 million words), and various comparable corpora: (a) corpus of translation of medical texts by professionals and semi-professionals (from English into Spanish); (b) corpus of non-translated medical texts in Spanish; c) corpus of non-translated medical texts in English, (d) corpus of translation of technical texts by professionals (from English into Spanish); and (d) corpus of non-translated technical texts in Spanish. The main findings support (1) the inexistence of simplification of translated text into Spanish (for most features) (non-translated Spanish texts are even more simple). (2) Convergence (translated texts are more homogeneous among themselves) can be observed only for syntactic features. (3) Transfer can only be observed partially: there is some positive transfer (translated texts show more lexical cognates), but no negative transfer (translated texts show more zero pronouns). Syntactic interference (shining-through) is observed for all translated texts (Corpas Pastor 2008).

After the initial sweeping success of ML approaches to detecting translations on surface and linguistically uninterpretable features, there appeared a research strand that aimed to combine the ML computational power with the corpus-linguistic interest in translationese properties. These efforts can be exemplified by Volansky, Ordan, and Wintner (2015) research, which tested the usefulness of a dozen of linguistically informed features, theoretically attributed to the main translation tendencies (simplification, interference, normalisation and explicitation). In effect, they used ML methodology to perform univariate analysis (they compare the accuracy of a binary translationese classification on each feature) to reveal the features prominence in the identification of translations. Their findings make a strong argument for interference as the major tendency in translation and, concurrently, for language-pair-related nature of translationese in general. The authors also make rigorous claims about the importance of a parallel data, content-independent features and genre-related nature of translationese trends.

The use of automatic text classification as a validation methodology combined with unsupervised and mildly supervised machine learning techniques (namely, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)) was promoted in Evert and Neumann (2017) for revealing the latent distinctions between text types (languages, registers, translations-non-translations) and exploring the sets of features that load on the respective discriminants. Unlike the previous research, the authors advocated the use of the *multivariate techniques* claiming that translationese is a systematic property of a text, not dissimilar to register specificity, and can hardly be conveyed by a single feature, but rather a combination of them (cf. multidimensional approach to register studies introduced by Biber (1988) and similar approach to translation in (Prieels et al. 2015)). An important methodological claim that the authors make is about the resources necessary for translationese studies. They assert that 'it is methodologically impossible to determine differences between translated and non-translated texts without comparing the realisation of a feature in the matching source text' (Evert and Neumann 2017: 49). It is interesting to note that despite their study is based on a balanced corpus involving five registers, the register variation was treated as a confounding factor that shapes translationese; any register-related interpretations were left for future work.

Evidence for language pair specificity of translationese. While developing effective language-independent applications to detect translations can be an interesting engineering task, there is ample evidence that translationese features and effects are indeed language pair and translation direction specific. In fact, the symmetric additions and omissions of items in both translation directions between two languages (demonstrated by Becher (2011), for example) are indicative of the impact of the contrastive properties of the language pair on the translators' choices. Reduced accuracy of the translationese classification, when a model trained on translations from SL1 is tested on translations from SL2, supports the same conclusion (Koppel and Ordan 2011; Popescu 2011). It is common to interpret the linguistic make-up of translations as a complex interplay of the two major forces: the SL shining-through pull and the TL normalisation pull (see, for example, Hansen-Schirra (2011)).

To sum up, the previous translationese research has established that translations are systematically and inherently different from the originally-authored texts due to the specificity of the underlying communicative situation and cognitive processes. It has been shown that the property of 'being a translation' is largely determined by the SL and the register conventions. The intuitive association between some frequency features and translationese universals proved difficult to be confirmed by empirical evidence due to the lack of objective link between the trend and its operationalisation. However, bottom-up exploratory approaches based on ML methods enable to reveal translationese indicators and the unique ways in which they coalesce into patterns in each register of a given translation direction.

In general, the relevance of translationese studies is supported by the renewed interest to the impact the human translated training data exerts on the quality of machine translation (Aharoni et al. 2014; Goutte et al. 2009; Graham et al. 2019; Popovic 2020; Stymne 2017; Zhang and Toral 2019). One of the earlier investigations into this issue by Lembersky, Ordan, and Wintner (2012) demonstrated that the BLEU score can be improved if the language models are trained on the translated texts and not the texts originally written in the TL.

The current project is based on balanced data for four registers, each represented by a combination of (1) a document- and sentence-aligned parallel corpus of professional published translations for English-to-Russian language pair and (2) a comparable corpus of non-translations in the target language. These components are necessary to reliably capture and describe various translationese effects by comparing feature frequencies across three text types in each register: sources, targets and reference texts. Methodologically, we combine multivariate analysis in supervised and unsupervised ML settings and univariate statistical analysis to reveal prominent translationese indicators and describe trends observed within and across the registers. Our features include content-independent morphosyntactic features that allow to abstract from topic and domain information as well as indirect lexical indicators retrieved from language models learnt on separate and much bigger register-comparable resources. Importantly, all features are shared by the two languages involved to enable placing all texts into the same multidimensional feature space.

2.2 Translationese and Register

This research explores the translation properties that are observed in various registers. It is difficult to deny that language is not homogeneous. Language is a combination of subsystems that are employed in specific communicative conditions. One important dimension of language variation, distinct from domain sublanguages, territorial or social dialects, has to do with the dominant communicative function and the generalised type of the situation in which the textual activity takes place. This type of variation is referred to as registers or genres depending on which aspects of the communicative event are focused. David Lee, the author of one of the text categorisation schemes in the British National Corpus (BNC), prefers to think about these competing terms as ‘two different points of view covering the same ground’ (Lee 2001: 46). The term *register* signals that language material is approached from the viewpoint of its internal properties (such as frequencies of linguistic items), which form specific patterns of use predetermined by the communicative conditions (‘the context of the situation’) in which they occur. The major situational factors are typically described following Halliday’s categorisation into field, tenor and mode. Genres are understood as text categories more focused on the text-external and functional parameters; they are text schemata licensed by the culture and superimposed on the register. According to James Martin, ‘no culture combines field, mode and tenor variables freely’ (1992: 562). This approach is in line with Michael Halliday’s interpretation of register (see Register Variation chapter in Halliday and Hasan 1989) and is adapted in a number of corpus and computational linguistics projects, especially based on the BNC (see Lijffijt et al. 2016; Neumann 2013; Santini et al. 2010; Sharoff 2018).

In translationese studies, it seems more typical to refer to the analysed text categories as registers (see Diwersy et al. 2014; Kruger and Rooy 2012; Lapshinova-Koltunski 2017 among other works). However, Delaere (2015) consistently prefers the term ‘genre’ to refer to the text categories of similar names and granularity, because in her research these categories are explicitly annotated using such non-linguistic characteristics (addressor, addressee, channel and communicative purpose), following the methodology in Biber and Conrad (2009).

In the current research, we follow this interpretation of the contextual language variation and refer to the four text categories under comparison (general domain mass-media texts, popular-scientific texts, fiction, political-economic news commentary) as registers.

Register is widely acknowledged as one of the major factors that influences the properties of translations, along with the source language.[2] This is not surprising precisely because of the strong SL pull in translations, given that ‘parallel registers are indeed more similar cross-linguistically than are disparate registers within a single language’ (Biber 1995: 279). In a lot of earlier research, this is corroborated as a by-product of a different research focus and/or as a result of observations from manual analysis of some restricted corpus data. For example, a relatively small-scale study based on half-a-million word corpus by Puurtinen (2003) indicated that genre could be an important factor guiding translation choices. The authors concluded that ‘subgenres of children’s literature … should be investigated separately’ (Puurtinen 2003: 403).

Xiao, He, and Ming (2010) report the construction of a register-balanced corpus of translational Chinese and original Chinese texts after the FLOB sampling frame. In their univariate analysis of several known translationese indicators, they show that the features tested, including lexical density (STTR), mean sentence length, conjunctions and passives frequencies, display ‘genre subtleties’ in translation.

Our research can be compared to Kruger and Rooy (2012), who see the investigation of the relationship between register and the features of translated language as one of their main research goals. They performed univariate analysis for seven features, which represented three translationese universals, to see how the universals would play out within and across their six registers. In their research design, explicitation, normalisation and simplification were operationalised with the (1) frequencies of full forms (as opposed to contractions), that-complementisers,

[2] Earlier studies that suggest that translationese is dependent on register are Steiner (1998), Reiss (1989) and Teich (2003), among others.

linking adverbials; (2) frequencies of coinages, loanwords and common lexical bundles; and (3) values for lexical diversity and mean word length, respectively. Their results provided limited evidence for universal character of translationese, rather each register demonstrated its own pattern of analysed features. In a later research using the same features, the levelling-out of registers, conceptualised as the assumed reduced register variability in favour of a neutral middle register, was not supported either (Redelinghuys 2016).

In recognition of the importance of register in translationese studies, researchers pay special attention to the selection and annotation of the reference corpus of non-translations: Castagnoli (2009) decided to build a new corpus from scratch, Delaere (2015) re-annotated an existing resource, Kunilovskaya and Lapshinova-Koltunski (2020) used a special corpus sampling strategy to extract functionally comparable subsets from larger corpus resources.

The large-scale studies of translated registers that allow reliable application of statistical methods or ML techniques are comparatively rare. There is a case study in Diwersy, Evert, and Neumann (2014), based on a reasonably large register-balanced bidirectional English and German corpus, but its contributions were more of the methodological nature: they reported few findings that characterised individual registers in translation, if any.

Delaere (2015) used the frequencies of linguistic items associated with the general properties of texts such as formal/neutral language and native/borrowed words to profile originally-authored and translated texts and test whether the translators tend to conform to the observed TL norm. Her findings for five genres in several language directions between Dutch, English and French generally confirmed the normalisation trend in translations and the impact of the genre and SL factors, but there was no consistency in the results. The authors attributed this inconsistency to incomplete metadata in the corpus and some unaccounted factors that might govern translators' choices. The sparsity of the indicators and domain disparities could also be confounding factors, given the lexical nature of the operationalisations implemented and the relatively small size of each subcorpus used in the study.

Unlike the previous study, which relied on predefined operationalisations of some properties of translated texts like levels of formality, Lapshinova-Koltunski (2017) employed hierarchical cluster analysis, an unsupervised ML method, and represented English-to-German translations and German non-translations in seven registers as feature vectors using eight lexico-grammatical patterns that were inspired by register studies to see how much the properties of translations were influenced by two factors—the register and the method of translation. Their features are context-independent and characterise texts through ratios of, for example, nominal vs. verbal parts-of-speech or through cumulative frequency values for items expressing modality or evaluation among others. The results of the study showed that the functional text type dimension dominated as a factor for some registers but not others. This research, as well as an earlier research on the same data using SVM classification (Vela and Lapshinova-Koltunski 2015), had its focus on the comparison of human and machine translation across a range of registers.

They found that the two translation varieties were more similar between themselves than any of them were similar to the register-comparable non-translations. In a later work on the same data, they used part-of-speech (PoS) trigrams in a number of binary text classification experiments to reveal and interpret features distinguishing translated registers. They confirmed their earlier finding that 'the genre dimensions in translation variation is much stronger than that of translation method' (Lapshinova-Koltunski and Zampieri 2018: 107). These three studies indicate that human and machine translations are more similar between themselves than any two translated genres, regardless the feature set used and ML approach chosen.

3 Methodology

In translationese research, the results are largely dependent on the features used to represent the texts, including their selection and extraction. Features are usually frequencies or ratios of linguistic items and phenomena, used to operationalise various hypothesised translationese trends or to capture and measure translationese effects in the bottom-up approach.

Another important factor is the type, quality and size of the corpus resources used to produce data tables. As it is shown above, both parallel and comparable components are required to be able to interpret quantitative differences between translation and non-translations.

There can be various ways of looking at the data methodologically, ranging from manual in-depth analysis of a few contrastive linguistic phenomena and/or statistical significance testing to ML experiments, usually cast as text classification problems or various types of factor analysis and computational linguistics methods. While the previous research has reported some tried and tested approaches, they leave a lot of room for development and exploration, especially if new research questions are posed.

Unlike much of the related work, where register effects on translationese properties are used as a backdrop for another primary research questions, the current research employs ML techniques to compare the type and strength of various translationese effects in several registers as well as to reveal the translationese indicators that might cut across all registers. This section has the description of these three major components of our research design: features, data and methods.

3.1 Feature Sets

Similarly to Volansky, Ordan, and Wintner (2015), our features are not selected to get the highest accuracy for the binary classification of originally-authored texts and translations (translationese classification). We seek to investigate the variation in translations along the register dimension in a linguistically interpretable way.

In the literature, the types of features used to capture translationese in the ML setting vary depending on the specific task. Translationese detection and SL identification tasks almost exclusively rely on character, word, lemma, PoS or mixed n-grams of various order[3] and most frequent lemmas (including function words) or PoS.[4] A bold exception is the projects that aim at sentence-level detection of translation direction (Eetemadi and Toutanova 2015; Sominsky and Wintner 2019). They leverage the aligned PoS information from source and target sides of the parallel corpora to achieve the state-of-the-art results. Sominsky and Wintner (2019) reported further improvements of up to 6% accuracy (at the expense of interpretability) for four out of six tested language pairs on distributional 50-dimension pre-trained GloVe word embeddings used to represent words and fed to a neural network of one bidirectional Long Short-Term Memory (BiLSTM) layer.

The more linguistically orientated research, which aims to know more about the linguistic specificity of translations, considers the feature selection the most challenging and creative part of the task. On top of the well-known and most-tested translationese indicators (such as type-to-token ratio, content-to-function words ratio, frequency of connectives/conjunctions and pronouns, ratio of contracted to full forms, average sentence length, mean word rank), the authors suggest more elaborately engineered features. For example, Arase and Zhou (2013) used the frequency of discontinuous structures to capture ‘phrase salad’ in MT. Redelinghuys (2016) calculated readability scores, while Volansky, Ordan, and Wintner (2015) operationalised the normalisation hypothesis with average point-wise mutual information (PMI, one of the association measures used to detect collocations) of all bigrams and ratio of repeated content words along with other features. Lapshinova-Koltunski (2017) suggested a feature set, which included features like frequency of evaluative patterns and degree of nominalisation (ratio of nominal and verbal PoS). Some experimenting was done with the frequency features based on parsed data: Ilisei et al. (2010) calculated ratio of simple sentences and parse tree depth and Kunilovskaya and Kutuzov (2018) extracted and counted syntactic relations tags from UD annotations of their corpora.

In our research the feature selection and engineering process was informed (1) by the findings in the translation and translationese studies, including the practical observations made in English-to-Russian translation textbooks, but never tested empirically and (2) by the practices in the register studies and variational linguistics on the assumption that translations could be viewed as a specific sublanguage, a third code (Duff 1981; Frawley 1984), based on the specificity of distribution of the linguistic features. This is supposed to enable measuring the cross-linguistic distance between the registers as well as between translations and non-translations. This approach effectively means that our feature set is language

[3] See, for instance, Baroni (2006), Kurokawa (2009), Arase (2013), Eetemadi (2015) and Rabinovich (2016).

[4] Some relevant studies are Popescu (2011), Koppel (2011) and Nisioi (2013).

pair specific and would require adaptation to be extended to other language pairs (see such adaptation in Kunilovskaya and Lapshinova-Koltunski 2020). Besides, our research design required that the features (3) should be shared by the languages involved in the experiment. We also focused on (4) content-independent features to reduce the noise from the topic and domain divergence between the parallel and the reference corpora, which excluded the common bag-of-words models from our options. Finally, we avoided (5) less interpretable features and (6) features that defy reliable extraction based on our experience.

Unlike much of the previous research into translationese, overviewed in Sect. 2.1, we do not assign features to the known translationese trends in the top-down manner, but empirically establish their role in producing various translationese effects. The experimental setup in this study can handle irrelevant or collinear features, and we use a reasonably high number of potential translationese indicators to be able to distil the most useful ones through feature selection.

Our feature set is composed of two parts. First, it includes 45 morphosyntactic features that were introduced in Kunilovskaya and Lapshinova-Koltunski (2019) to capture human translation quality. We provide a brief overview of these features below. For the full description of each individual feature, refer to Appendix. The feature codes used in this chapter and the extraction details are given in the Appendix alphabetically. Second, it comprises 11 abstract lexical features to reflect the specificity of the lexical choice in translations.

The morphosyntactic features are extracted from the annotation performed within the Universal Dependencies framework (Straka and Straková 2017), using models pre-trained on 2.5 versions of the EWT and SynTagRus treebanks for English and Russian, respectively.

More than a third of these features (17) are the frequencies of the default UD morphosyntactic tags (such as *ccomp*: clausal complements or *sconj*: subordinating conjunctions) and their combinations (such as *numcls*: number of clauses per sentence counted as the number of relations tagged *csubj, acl:relcl, advcl, acl, xcomp* in one sentence); when extracting PoS tags for various types of pronouns and other closed word classes, we used lists to filter out noise. The other third of the features (16) involved custom rules and extraction patterns, detailed in Appendix. These include *lexical type-to-token ratio, modal predicates, passives, mean dependency distance* (*mdd*, which represents 'comprehension difficulty' defined as 'the distance between words and their parents, measured in terms of intervening words' (Jing and Liu 2015). In developing these features we took into consideration the description in (Evert and Neumann 2017; Nini 2015) for English and in (Katinskaya and Sharoff 2015) for Russian. Further on, the cumulative frequencies for the four semantic types of *connectives, epistemic markers* and *adverbial quantifiers* are extracted using predefined lists compiled from the literature (see more details on the items selection, academic sources, extraction and disambiguation in Appendix).

Generally, our UD-based indicators include morphological forms (e.g. non-finite forms of verbs), syntactic relations (e.g. clausal complements), syntactic functions (e.g. modal predicates), word classes (e.g. pronouns, discourse markers). The extraction quality of these features largely depends on the quality of the UD annotation: for v2.5 mean accuracy on raw text is reported at 93.3/97.8 for universal PoS, 94.2/93.5 for morphological features and 77.0/85.0 for labelled dependency attachment for English/Russian, respectively.[5]

For this project we implemented 11 additional features to approach translationese at the lexical level as well. It is obvious that we cannot rely on frequencies of individual character or word n-grams in our cross-lingual setting. Besides, it is a known fact that sparse vectors of string features do not generalise well across domains (Eetemadi and Toutanova 2015). Instead, we used language model (LM) perplexities and calculated ratios of n-grams from top and bottom frequency quartiles, using the KenLM toolkit (Heafield 2011) and Quest ++ utilities (Specia et al. 2015). These features are used for the analysis of translationese in the research projects, which target translation quality (see Karakanta and Teich 2019 and Quest ++ feature set). We hypothesise that translated texts might have a diverging lexical composition in terms of ratios of n-grams from high- and low-frequency bands and sentence perplexity scores due to unseen sequences induced by the translation process. Our text-level lexical features include:

- mean target sentence perplexity score from the 3-g language models trained on large register-comparable corpora (see 3.2.2 for details);
- standard deviation value for the above sentence perplexities to account for possibly uneven lexical complexity of sentences in the translated texts;
- ratio of uni-, bi-, trigram that were not seen in the n-gram lists from the reference corpora;
- ratio of n-grams from the 1st frequency quartile (low-frequency items)
- ratio of n-grams from the 4th frequency quartile (high-frequency items)

To produce these features, we collected separate language resources for each register making sure they do not intersect with the smaller reference corpora included in our experimental data to exclude unfair bias for these features. Before learning LMs and generating n-gram lists, all corpora had been lemmatised and PoS-tagged with UDPipe (Straka and Straková 2017) to get lempos representation (e.g. as_SCONJ i_PRON look_VERB up_ADP ._PUNCT). This is required because Russian is a morphologically rich language; English is pre-processed for higher consistency and comparability.

As a result of feature extraction, each text in our data was represented as a vector, where individual components corresponded to the value of each feature for this text. The dataset, used in the experiments, can be thought of as a table, which has texts in rows and features in columns. Note that prior to the experiments, the values of each feature were standardised to get the distribution with a mean value 0 and standard deviation of 1. This helps to ensure that all features have the variance of the same order, and each feature makes the same contribution to the differences observed, regardless of large discrepancies in real values between some indicators.

[5] http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_20_models.

3.2 *Research Corpora*

This research relies on several parallel and comparable corpora to explore the linguistic properties of texts translated from English into Russian by professional translators across a variety of registers. We distinguish between the corpora used to conduct experiments (data) and the corpora used to learn language models and produce n-gram frequency lists (linguistic resources).

All corpora were put through the same pre-processing pipeline (spelling unification, text size normalisation, deduplication, noise filtering), annotated with UDPipe and converted to PoS-tagged lemmas (lempos format).

3.2.1 Data

The selection of registers for this project was limited by the availability of the English-Russian parallel and comparable corpora that would store texts of reasonable size and structure. We considered a wide variety of the available parallel corpora, including web corpora (Yandex 1 M-token parallel corpus, Parallel Corpora for European Languages), United Nations corpus, corpora of subtitles and Wiki Titles, TedTalks corpora and mozilla transvision corpus of technical translations. But the units of storage in these corpora were often limited to one sentence or would include a lot of non-textual information and tables. TedTalks transcripts and subtitles have specific translation processes behind them that can unfairly influence the frequencies of our features. It is also more difficult to make assumptions about the translation quality for these corpora and compile non-translated comparable corpora for them.

We focused on the four registers: general domain mass-media texts, popular-scientific texts, fiction and the news commentary texts in the political and economic domain. All translations included in the experiments are published. We only selected the corpora that store texts with respect to their natural text boundaries, which allows the collection of text-level statistics. The parallel subcorpora are document-level and sentence-aligned. The global sources of data in this project can be described as follows.

1. *Mass-media* parallel corpora include data from the three major sources: a quarter comes from the parallel component of the *Russian National Corpus (RNC)*[6] and the rest of the data were manually collected or crawled from *InoSMI.ru* and *BBC.com/russian* (2018–2020).
2. *Popular scientific* parallel corpus is self-compiled from a dozen of full-length English books on a range of subjects including biology, physics, sociology, history, anthropology, robotics, medicine, and their published translations into Russian from 1999 to 2016 period. This corpus is now included into the RNC

[6] https://ruscorpora.ru/.

Table 1 The macro-corpus used for research purposes (k=thousand, m=million)

	Type of data	Words	Sentences	Documents
general media	parallel	731 k	31 k	525
	reference	625 k	33 k	448
popular science	parallel	1 m	42 k	112
	reference	1 m	46 k	101
fiction	parallel	11 m	564 k	149
	reference	12 m	706 k	200
commentary	parallel	301 k	12 k	347
	reference	276 k	13 k	334

parallel resources. While the number of observations is small, the selected unit of storage is a chapter or a part of the book.

3. The parallel data for *fiction* is entirely from the RNC parallel component. It includes 149 source texts of various length and literary genres, but mostly novels representing over a hundred of authors from Dickens to Rowling.
4. Parallel *political and economic articles (commentary)* are extracted from the *WMT News Commentary* corpus (v.15),[7] which contains political and economic commentary crawled from *Project Syndicate* website.

The originally-authored Russian texts to be used as the reference for the former three registers were randomly sampled from the respective register subcorpus of the main 500-million RNC and for the last category—from the 300-million contemporary Russian newspaper corpus, included in the RNC monolingual resources.

Table 1 has the description of the pre-processed and annotated parts of our register-balanced corpus including the parallel and comparable monolingual components. For the parallel data we report the size on the SL side only.

In total we have 3349 documents in two languages, labelled for four registers and three types (sources, targets, reference).

3.2.2 Linguistic Resources

The resources for LM training in all registers, except the English news commentary, come from the British National Corpus (BNC) and the Russian National Corpus (RNC). We relied on the available metadata to ensure maximum comparability with the parallel data in terms of intended audience, text production time and communicative function. The English political and economic commentary reference texts are collected from the WMT News Commentary corpus outside the English-Russian parallel data. Note that these resources exclude the random

[7] http://www.casmacat.eu/corpus/news-commentary.html.

Table 2 Corpora used to train language models and generate n-gram lists

	Language	Words	Sentences	Documents
general media	en	3.9 m	177 k	100
	ru	129 m	6.9 m	226 k
popular science	en	17.7 m	682 k	528
	ru	1.9 m	93 k	378
fiction	en	18.6 m	1.2 m	431
	ru	37.6 m	2.6 m	580
commentary	en	5.9 m	237 k	8.7 k
	ru	5.7 m	252 k	9.5 k

samples used as reference data and described in Table 1. The general shape of the resources after pre-processing and annotations can be found in Table 2.

We will indicate that the mass-media items in the BNC do not observe true document boundaries but are in fact text chunks of varying length. However, it is irrelevant for the purposes of building LMs and n-gram lists.

3.3 Methods

Our methodology combines the data representation and visualisation approaches which were shown to be effective for the study of translations in Evert and Neumann (2017) and the idea that in revealing or measuring translationese effects, the distance between the source and target languages (or, in our case, registers) has to be taken into account. We develop the general approach tested in Kunilovskaya and Lapshinova-Koltunski (2020) on one register for two language pairs.

To represent texts in our data we generate feature vectors, where each component has the value for a particular linguistic parameter. With the exception of the LM perplexity scores, these values are the frequencies or ratios of a targeted linguistic phenomenon, captured through a set of PoS tags or a syntactic pattern. For features based on the search lists, the values are cumulative frequencies of all items on the respective list. For n-gram counts, we used an empirically established frequency threshold of 10, which means that we ignored the n-grams with a frequency lower than 10. This measure helps to avoid zero values for bigram and trigram ratios. Given that our features are the same for all text categories and text types, this representation effectively puts them in a shared feature space. The extraction details are given in Sect. 3.1 and in Appendix.

We resort to PCA, an unsupervised ML technique, for dimensionality reduction to present our observations in scatter plots and visually estimate whether our features reflect the ontological text categories and types. The visual impressions are verified by the results of text classification. In all experiments we rely on the linear SVM algorithm, set to the default scikit-learn parameters (C = 1.0, degree = 3, gamma = auto). The algorithm is fed with the feature vectors that have been centred around the mean and scaled to unit variance and is run in the 'balanced' mode to offset the unequal number of observations in the training classes. We report the results in the tenfold cross-validation setting to reduce the possible biases of any single held-out test set.

In accord with our research questions, given in Sect. 1, the text classifications are designed to capture the following general properties and phenomena:

- translational status: a binary classification for each register;
- register variation: a 4-label classification for non-translations in each language;
- standardisation effect: a 4-label classification for translated texts only.

To determine the position of each translated register with regard to the sources and TL non-translations, we average the real-valued vectors across each of the three text types and calculate the *Euclidean distances* (a square root from the sum of squared differences between the corresponding dimensions of the two vectors) between them. We rely on the Euclidean distance (as opposed to cosine similarity, for example) because in this experiment we use unscaled vectors and the magnitude of the values in each dimension matters. The differences between the three measurements, which can be pictured as triangles, demonstrate the relative proximity (similarity) of the translated texts to the originally-authored registers in the two languages. The idea to measure linguistic (morphosyntactic) distances between languages for the purposes of translationese studies is not new. To this end, Nikolaev et al. (2020) computed the cross-linguistic congruence index as the proportion of matching universal PoS tags and dependency labels for all manually aligned content words in a parallel corpus. They acknowledged that there was no established procedure to achieve it.

The explanatory analysis of the linguistic specificity of translations in each register is based on the best translationese indicators, i.e. the top N features that can be used by the ML learning algorithm to differentiate the classes with the minimum loss in the classifier performance. Our experimental results indicated that the best performance for the top 10 and top 20 features was returned by the *Recursive Feature Elimination (RFE)* feature selection algorithm, which internally used *Support Vector Regressor (SVR)* with the default scikit-learn settings. The same approach was used to reveal register contrast indicators that were necessary to demonstrate the amount of intersection between the translationese and cross-lingual contrast features.

Finally, we perform a succession of the univariate analyses to establish which features contribute to various translationese effects that we distinguish in this study following a procedure described below. In all experiments we used the *two-tailed T-test for samples with unequal variance* and quantified the effect size of the differences with *Cohen's d*. First, we identify the features that have significant differences between translations and non-translations (tgt, ref): these are translationese indicators. Then, we establish whether there are differences between the two cross-linguistic registers (src, ref) with respect to a given feature (the language gap). Finally, we compare the average frequency for the feature in translations with those in the source and target languages to determine how it relates to these values (greater or smaller).

Combinations of these tests outcomes yield the feature sets for the following translationese effects:

1. shining-through effect: translationese features in the language gap, i.e. we observe significant differences between translations and non-translations and between English and Russian non-translations; and the frequencies of features from translations are smaller than in English but significantly greater than in non-translated Russian (src > tgt > ref) or greater than in English but smaller than in Russian (src < tgt < ref);
2. anglicisation: translationese features demonstrating frequencies outside the English extent of the significant language gap;
3. SL/TL-independent translationese: translationese features with significant differences from both languages and no language gap;
4. over-normalisation: translationese features demonstrating frequencies outside the Russian extent of the significant language gap;
5. adaptation: features that have significant differences for the two languages, but not translationese features, i.e. their frequencies are adapted to the TL norm.

This procedure is also supposed to reveal features that are useless for our purposes: the feature that has the same frequencies in translations and non-translations, and also do not distinguish the languages.

4 Results

In this section, we first report the results of the two classification experiments that test the ability of our feature sets (1) to distinguish translations and non-translations in each register, (2) to capture the register variation in the originally-authored texts in each language. We also look at the performance of the register classification on the translated registers to check whether the register distinctions are diluted by the translation process. If the translated registers are more difficult to classify, we can confirm the levelling-out hypothesis. The second paragraph demonstrates how the translated registers are positioned against comparable non-translations in both languages (src, ref) based on the Euclidean distances in our setup. We complement the spacial representation of translated and non-translated registers with histograms for values on the strongest PCA dimension, which appears to mostly capture register variation in our data. Finally, we describe the subsets of features that are revealed through feature selection and comparative frequency analysis and represent several translationese effects. Feature analysis is performed to explain the observed specificity of each translated register with regard to their sources and reference non-translations.

4.1 Translationese and Register Distinctions

For a preliminary investigation of the data, given our features, we visualised the distinctions between all text types on the full feature set and on its morphosyntactic and lexical parts. For example, Fig. 1 has a scatter plot, where each document is represented by the values on the first two PCA dimensions, i.e. the result of the dimensionality reduction of the 45-dimensional morphosyntactic vector. Unlike lexical features (not shown for the consideration of space), the morphosyntactic features manage good separation of the registers and the two languages. It seems that the register variation is found on Dimension 1, which explains the most variance in the data, while Dimension 2 (shown on the vertical axis in Fig. 1) captures the language contrast. The lexical features are not able to achieve this representation of data on the most prominent known properties of the texts: they squeeze all variance into the first dimension. It means that in terms of ratios of high-frequency and low-frequency n-grams the similarity between registers from different languages is stronger than the differences between languages. This observation is confirmed by the language contrast classification (English vs. Russian original texts) results: for morphosyntax 100% accuracy can be achieved

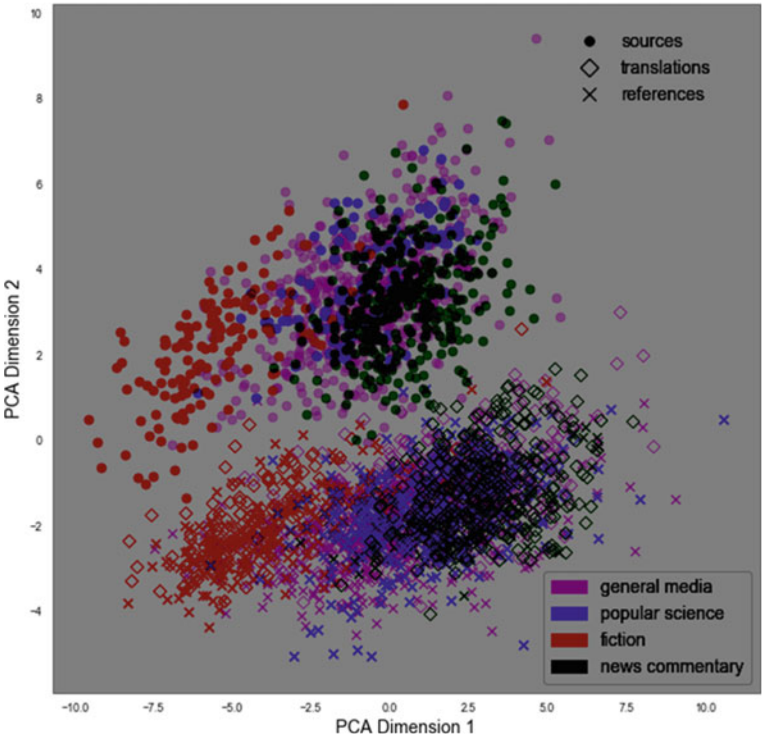


Fig. 1 Values on the first two PCA dimensions derived from the morphosyntactic features

on just 3 features (*aux, aux:pass, parataxis*), while the 11 lexical features returned only 85%.

The concatenation of the two feature sets captures the register distinctions on Dimension 1 and language distinctions on Dimension 2 more clearly (see Fig. 3).

However, the distinctions of translations and non-translations, required by the first step in our methodology, are clouded. To bring them to the fore for closer exploration, we tried to cast the full feature vectors of size 56 for translations and non-translated Russian texts to a bidimensional space by PCA and produced a scatter plot of the resulting data. The independent subplots in Fig. 2 position the texts in each register according to the values received on the first two principle components.

It can be seen that translations are shifted away from the non-translations, especially in general mass media and news commentary. It means that our features do register some divergence of translated Russian from the expected TL norm in these registers represented by non-translations. Admittedly, the visual impressions are more subtle in the other two registers. Note that PCA is unsupervised: it is unaware of any text types that are colour-coded in the plots. Besides, PCA reduces the 56 dimensions to just two, necessary to plot the data, which inevitably leads to the loss of information and distortions. That is why we verify the visual impressions

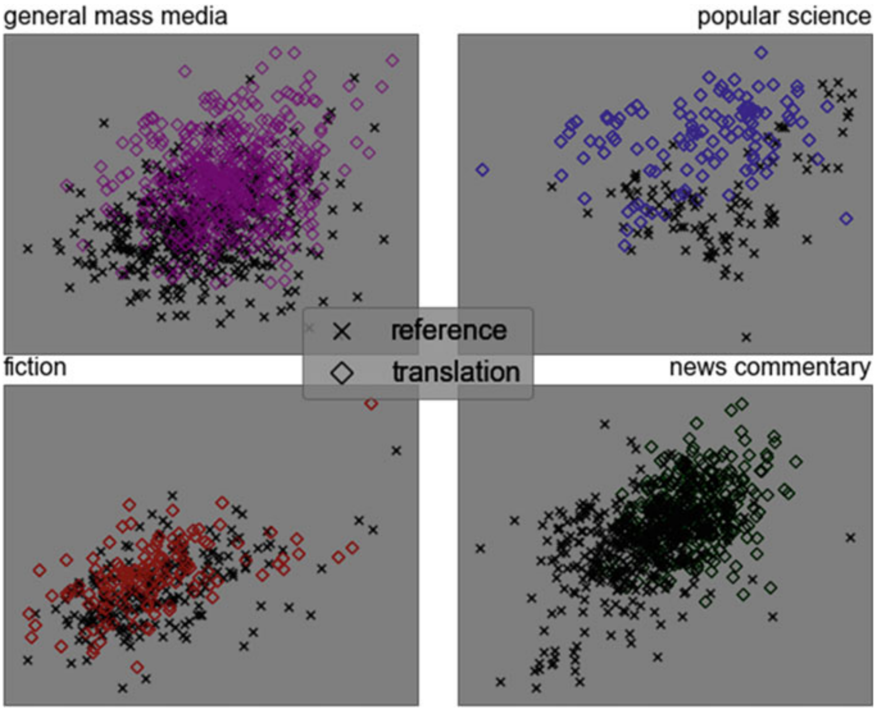


Fig. 2 Differences between translations and non-translations by register

Table 3 SVM performance on the translationese classification in each register

	N texts	56 features		45 morphosyntax		11 lexis	
general media	973	87%	0.872	**87%**	**0.869**	75%	0.750
popular science	213	98%	0.977	**98%**	**0.981**	76%	0.754
fiction	349	95%	0.953	94%	0.936	77%	0.759
commentary	681	95%	0.947	93%	0.934	**88%**	**0.879**

with a series of binary translationese classifications using SVM. The classification results confirm that PCA visualisations can be, indeed, misleading, because the registers with seemingly different visual distinctions (fiction and news commentary) achieve the same high classification accuracy, while the accuracy for general mass media is lower, in contrast with what is observed in Fig. 2.

The cross-validation results are presented in Table 3, which shows SVM performance on the translationese classification, taking into account accuracies and macro F1 scores. On the full feature set in three registers, SVM achieves the accuracy of over 95%, while for mass-media texts it is 87%, which is still reasonable high. We have fairly balanced classes in all registers, so the chance level never exceeds 50%.

The classification experiments on morphosyntactic and lexical feature sets separately indicate that the result in the 56 features column (see Table 3) is mostly produced by the morphosyntactic features. If lexical features are eliminated the classifier performance does not degrade much in any registers: the loss amounts to 1% and 2% in accuracy for fiction and commentary at most. However, switching to just lexical features results in the drops in performance ranging from minimum 7% (*news commentary*) to maximum 17% (*popular science*). It means that for the translationese classification (1) *news commentary* relies on the lexical features most, i.e. they demonstrate the highest divergence from non-translations; (2) for *popular science* structure is most important, i.e. translations differ from non-translations in morphosyntax; (3) in *general media* both feature sets perform the worst, possibly because of the higher variation in the respective subcorpora observed in Fig. 3.

Secondly, we are interested in finding out whether our features model the register diversity in both non-translated languages well. In Fig. 3 we plotted the originally-authored texts in the two languages, represented by their values on the first two PCA dimensions generated by the PCA transform of the full feature vector of size 56. Most variance is explained by Dimension 1, which captures register variation. Texts from different registers seem to occupy specific areas along the horizontal axis, especially in Russian. The second dimension has the clear separation of the two languages. The plot in Fig. 3 also indicates that some eponymous registers are closer together across languages than others. For example, *fiction* and *news commentary* seem to be more similar along the vertical 'language contrast' dimension than *general mass-media texts* and *popular science*.

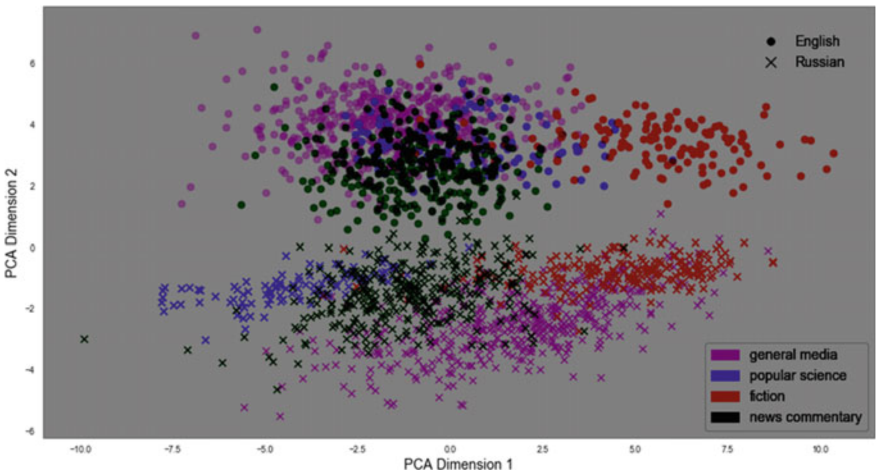


Fig. 3 PCA representation of registers in non-translations in English and Russian (56 features)

Popular science has the most expressed register differences in the cross-linguistic perspective of the four registers (notice the horizontal mismatch of the respective blue areas in the plot). *Mass-media texts* display a lot of in-category variation along the horizontal 'register' axis, especially in Russian. Judging by the upward and downward shifts of the respective clouds, this register passes some register distinctions on to Dimension 2, which ideally would capture only the language contrast. PCA on our features also struggles with distinguishing popular science and news commentary in English.

The classification results confirm that our features separate the four registers fairly well. For all 56 features, the SVM classifier, which predicted the four classes, returned 97% accuracy for each languages (F1-score 0.966 and 0.974 for English and Russian respectively). The chance level is 30% for English and 34% for Russian, with correction for imbalances between the four classes. In line with the visual impressions, most classification errors were between *mass media, commentary* and *popular science* in English and between *media* and *fiction* in Russian.

As expected in this experiment, the lexical features performed better: the 11 features were only 1% worse than 56 for English, while for Russian the decrease in performance amounted to 4%. The morphosyntactic features (45) alone were able to achieve only 78% and 81% accuracy for English and Russian, respectively. We can tentatively conclude that in our setting the register distinctions in English are conveyed through lexis to a greater extent than in Russian, where registers have more morphosyntactic specificity.

Finally, we tested whether the register distinctions in the SL are flattened out by the translation process—an assumption made by the levelling-out hypothesis (the tendency of translations to gravitate towards unmarked features in contrast to non-translated texts (Baker 1996)). The plot in Fig. 4 shows the difference in the

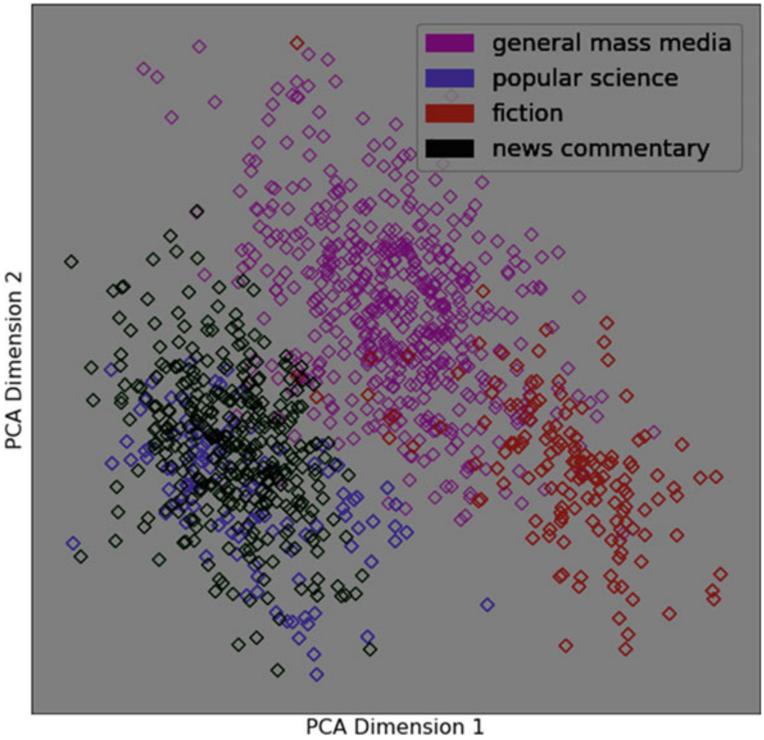


Fig. 4 Translated registers in Russian: PCA transformation of 56-dimensional feature vectors

localisation of the registers, some of which are even better separated than in the non-translated Russian (compare to the bottom part of the plot in Fig. 3). The translation process seems to import some confusion between *popular-scientific texts* and *news commentary*, on the one hand, and reinforce the separation between these two and *mass media* and *fiction*, on the other.

In this experiment, the SVM achieved the average tenfold cross-validation accuracy of 99% with a macro F1-score of 0.982 on the full feature set. Interestingly, the errors in the contingency table were between other classes than in non-translated registers: they were predictably between *news commentary* and *popular-scientific texts* (same as in the classification for English originals), rather than between *mass media* and *fiction* (as was the case in the classification for Russian originals).

Another intriguing observation is that the importance of lexical features for predicting translated registers increased compared to the texts originally written in Russian. The accuracy of register classification on the lexical feature set went up from 93 to 99% and was better than on all the 56 features. At the same time, the morphosyntax of translations introduced some noise: the classification on the 45 features from UD annotations for translation was 1% worse than for the texts

Table 4 Register distinctions in the original texts and translations for different feature sets (accuracies and macro F1 scores)

	N texts	56 features		45 morphosyntax		11 lexis	
English sources	1133	97%	0.966	78%	0.789	96%	0.955
Russian reference	1083	97%	0.974	81%	0.831	93%	0.934
Russian targets	1133	99%	0.982	80%	0.806	99%	0.983

originally written in Russian (80 vs. 81% accuracy). It indicates that the translation process does interfere with the target language register system on the structural level, but in terms of lexis translators tend to conform to the conventional distributions seen in the respective register. Table 4 systematises the results of the 4-class register classifications run on the three feature sets for each type of text in this project.

4.2 Euclidean Distances Between Translations and Non-Translations

To measure the apparent change of register properties in the translated language, we calculated the Euclidean distances between the register vectors for each text type (sources, targets, references). They were produced by averaging the text vectors across each category. The resulting distances are shown in Fig. 5 as a scale of the real values indicated in the diagrams. While lexical features did not contribute much to defining the specificity of translations, they were not used in measuring these distances. Besides, due to the drastic differences in the magnitudes between

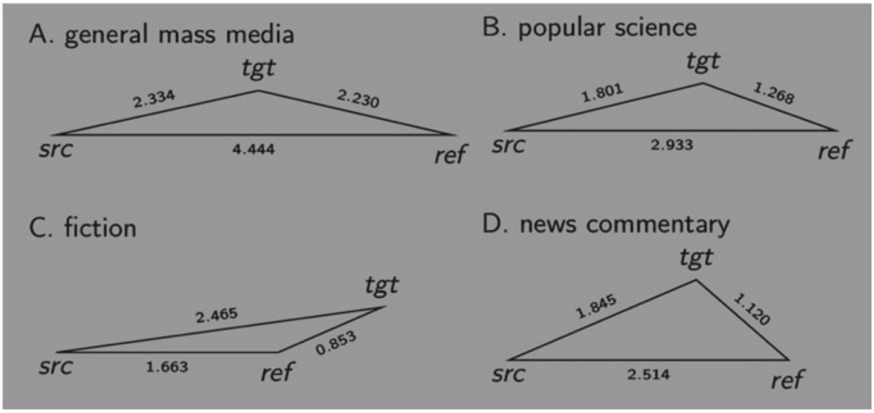


Fig. 5 Euclidean distances between the text types in each register

morphosyntactic and lexical features the latter overshadowed the former in this distance measure.

The translations in each register demonstrate some differences in how they are related to their sources and the expected target language norm. The *mass media* and *popular science* texts seem to have the most similar translationese properties, though the scale of differences is greater in the former. This generalised representation of translations from the *news commentary* subcorpora makes translations appear to be shifted more towards the TL than in the previous two registers, but at the same time the translations are more distinct from either of languages (this is indicated by the greater elevation of the *tgt* apex over the *src-tgt* plain and can be a sign of the greater amount of SL/TL-independent translationese in this register). Finally, *fiction* stands out as demonstrating an uncommon translationese shape: the diagram indicates the prevalence of adaptation or over-normalisation over shining-through effects. Note that the distances between originally-authored texts (*src* and *ref* in Fig. 5) replicate the visual results from Fig. 3.

As an additional sanity check, we computed the same measure for the random halves of the reference corpora: the average distances over 10 iterations range from 0.169 (media) to 0.712 (fiction). This confirms that translations in Russian are systematically different from the texts in the same register originally written in Russian.

The peculiarities of translationese flavours in various registers are best captured on the PCA ‘register’ dimension (Dimension 1) obtained from the full feature set for all texts in this project (see Fig. 6). The register properties of translations (solid coloured lines) do not necessarily replicate one language or the other, and the similarities between translations and non-translations can be seen under various register contrast conditions. The greatest mismatch of the cross-linguistic registers is seen in *general media* and *popular science*, but in the former translations tend to be in the language gap, and in the latter they appear to reproduce the TL norms. In *fiction* and *news commentary* register conventions seem to be most similar in English and Russian, and yet translations either faithfully coincide with these conventions or deviate from both.

The representations in these plots should not be taken literary, however. They do not account for the distinctions captured on the other PCA dimension and are based on the crude 2-dimensional transformation of the full feature vector. Contrary to the visual impression, translations are easily distinguishable from non-translations in all registers (Table 3).

To test Biber’s claim that registers can be more distant intra-linguistically than cross-linguistically (Biber 1995: 279), we used the same approach to measure pairwise distances between registers in non-translated English and Russian. The results in Table 5, considered together with distances between *src* and *ref* for each register in Fig. 5, support this claim. In both languages fiction is more isolated from other registers structurally, especially in English, while cross-linguistically it returns the smallest distance of 1.663.

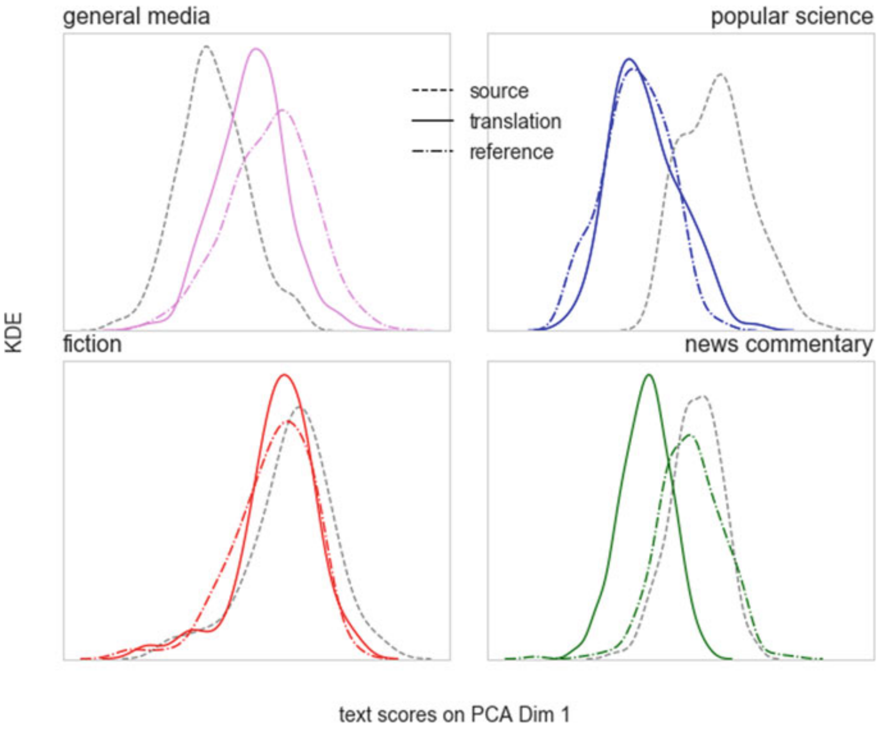


Fig. 6 Kernel Density Estimation (KDE) for the values on the PCA Dimension 1 (56 features)

Table 5 Euclidean distances between intralinguistic registers based on structural properties (values for English are under the diagonal; values for Russian are above the diagonal)

	general media	popular science	fiction	commentary
general media		1.522	3.070	2.274
popular science	0.250		4.558	0.913
fiction	5.791	5.640		5.273
commentary	0.587	0.693	6.137	

4.3 Translationese Effects and Features

In this paragraph we explore the specificity of translationese in each register through feature analysis. The results of the procedure based on the univariate analyses for *tgt-ref* (translationese), *src-ref* (language gap) and *src-tgt* (proximity to sources) are presented in Table 6. It aims to associate our features with the translationese effects described in paragraph 3.3. For the consideration of space, the table lists the 20 best translationese indicators in each register. In brackets we indicate the

Table 6 Features associated with translationese effects (based on univariate analysis of 56 features)

	shining-through	anglicisation	SL/TL independent	over-normalisation	adaptation	useless
General media	but, nnargs, relativ, interrog, whconj, neg, **parataxis**, **comp**, uoov, **trioov**, advers (23)	xcomp, **acl**, sconj (4)	epist, passives (4)	- (9)	**pverbals**, **lexdens**, **bifreq**, bioov (16)	- (0)
Popular science	**nnargs**, relativ, **deverbals**, **copula**, whconj, **aux**, **parataxis**, nsubj:pass, infs, mpred (29)	**xcomp** (2)	passives, **sconj** (5)	tempseq, simple, acl (8)	caus, advers, **mdd** (10)	**epist** (2)
Fiction	nnargs, **tempseq**, **pverbals**, whconj, **neg**, **ppron**, parataxis, **ccomp**, interrog (18)	- (0)	epist, **sconj** (3)	**xcomp**, simple, **mdd** (10)	bioov, bifreq, finites, uoov, trifreq, **lexdens** (23)	deverbals (2)
Commentary	**but**, relativ, **aux**, **parataxis**, **comp**, sup, **infs**, **advers** (20)	epist, passives, sconj (7)	ccomp (4)	**finites**, bifreq, **bioov**, **lexdens**, uoov, trioov (20)	**lexTTR**, ppron (4)	- (1)
SHARED	parataxis (6)	- (0)	- (0)	- (2)	- (0)	- (0)

Note that the three translationese features (*sconj, epist, parataxis*) shared between the lists of most important translationese indicators for our four registers, given in Table 6, are outside of the top 10 translationese indicators. Two of them (*epist, sconj*) are associated with different translationese trends in different registers total number of features (out of 56) that fall with the respective translationese effect according to the frequency analysis. The bold font indicates the features that are among the 20 most important register contrast indicators in the respective cross-linguistic register classifications. In all four cross-linguistic register classifications (*media_src vs media_ref, fiction_src vs fiction_ref*, etc.), the accuracy on the selected features is 100%.

To identify the best translationese and the best register contrast indicators mentioned above, we relied on the Recursive Feature Elimination (RFE) algorithm in scikit-learn, a Python library. In effect, this algorithm performs an ablation study on a given feature set by recursively pruning the least important features in the multivariate setting, based an external estimator (SVR in our case). The univariate approach to feature selection based on ANOVA (SelectKBest algorithm in scikit-learn) returned a higher loss in classification performance for all experiments: on average the classification on the 20 best ANOVA features performed 2.9% worse than on the full feature set. For RFE-SVR this loss in the same experiments was only 0.9%. However, the two feature selection algorithms demonstrate contrasting performance on *popular-scientific texts*, where ANOVA is better, and on *fiction,* where the RFE 20 features do well, while ANOVA features demonstrate 5.8% decrease in performance on the F1 score. It indicates that in the first case the multivariate analysis approach fails to reveal meaningful correlations between the features frequencies, while for *fiction* the discovered patterns explain the difference between translations and non-translations better than mere univariate comparison of features. Nonetheless, the intersection between the 20 best indicators, returned by RFE and ANOVA, ranges from 9 to 13 features for different experiments.

We should reiterate here from Sect. 3.3 that ‘adaptation’ and ‘useless’ sets include features that are not translationese indicators per se, because there are no statistically significant differences for their frequencies in translations and non-translations. Nonetheless, they are not irrelevant for characterising translations. As we will see below they are also important for the machine classification.

It can be seen from Table 6 that *fiction* has the minimum number of shining-through features (18) and the maximum number of over-normalised (10) and totally adapted features (23) together, which explains the shape of the triangle for fiction shown in Fig. 5 and the matching lines in Fig. 6.

News commentary is peculiar for having the maximum number of anglicised (7) and over-normalised features (20). It makes the translated texts in this register stand out as being more distinct from both SL and TL, indicated in Fig. 5 as a greater elevation of the translations apex over the *src-ref* plain and in Fig. 6 by the location of the translations outside the area shared by sources and reference.

Another immediate observation is that the registers tend to have no shared features for the suggested translationese effects, except shining-through and over-normalisation. However, even these effects seem to be achieved through widely different sets of features: only 6 features are shared among the average of 23 features for shining-through (*nnargs, relativ, whconj, parataxis, interrog, mpred*) and there are two shared over-normalisation indicators (*possdet, correl*).

It is also clear from Table 6 that, in terms of the number of features, shining-through is by far the most important type of deviation from the expected norm in translation.

We failed to detect any pattern in the relation of the features prominent in cross-linguistic register classifications (in bold) and the features important for the translationese classification (named in Table 6). Some of the contrastive register features are adapted to the TL norms and some are carried over from the SL.

The lists in Table 6 should be taken with caution, though. One limitation is that some features have negligibly small values and calculations for them are less reliable. For others, the differences in frequencies can be significant but the effect size is small. Besides, the impact of some feature sets associated with a given translationese effect can be comparatively small in the classification task, despite their size.

To verify the observations from the univariate analysis, we extracted the absolute weights of the features associated with each effect for each register from the SVM translationese classifier, and calculated the mean and standard deviation (SD) for these weights. Feature weights from a linear SVM classifier can be used to identify the features that contributed most to the classifier decision. This approach is known to be reliable in feature ranking (Chang and Lin 2008) . Additionally, we looked at the effect size (measured as Cohen's d) for the features with significant differences in frequencies between translations and non-translations (at $p < 0.05$). We report the findings for the most prominent trends by register in Table 7.

It can be seen from Table 7 that the effect size in the last column did not correlate with the classifier weights. Some features with the observed greater magnitude of

Table 7 The most prominent translationese effects in each register (in the order of importance based on the classifier weights)

	effect	N features	Mean weights	SD	Cohen's d
General media	anglicisation	4	0.645	0.210	0.851
	shining-through	23	0.348	0.395	0.232
	adaptation	16	0.325	0.316	–
Popular science	SL/ TL-independent	5	0.243	0.137	0.063
	adaptation	16	0.223	0.118	–
	shining-through	29	0.183	0.161	0.079
Fiction	shining-through	18	0.483	0.371	0.274
	over-normalisation	10	0.451	0.387	0.099
	adaptation	23	0.398	0.242	–
News commentary	adaptation	4	0.662	0.280	–
	anglicisation	7	0.583	0.380	0.601
	shining-through	20	0.553	0.302	0.361
	over-normalisation	20	0.449	0.292	0.200

differences were not selected by the algorithm as important. The comparison of the performance of the two feature selection algorithms, given above, shows that from a machine point of view finding patterns in the data is more effective than relying on separate features in most cases. It is not clear, however, which translationese effects are more visible (if any) to a human user.

5 Register-Based Translationese Varieties

We have seen that professional translations deviate from non-translations in the TL in all registers, which is particularly noticeable on the structural level. These deviations accommodate a number of trends, including shining-through, over-normalisation and adaptation.

The size and the combination of the translationese effects is register-specific, especially if we consider the associated sets of features. Our registers have just one intersecting translationese indicator in the top 20 most important translationese features (*parataxis*). It captures one strong and universal trend across our registers in translations—to spot more introductory and parenthetical elements and non-linear syntax. In general, the lexical features perform much worse than the structural (morphosyntactic) ones, with the difference in accuracies of the translationese classifications ranging from 22% (*popular science*) to 5% (*news commentary*).

As for the translationese effects, shining-through is the strongest trend in all registers, judging by the number of features identified as such and by their weights in the classifier. It is complemented by tendencies with less features, but sometimes higher prominence, to create a unique linguistic make-up for each category, described below.

1. In *general media* the strong pull towards the SL is emphasised by anglicised features and is to an extent counter-balanced by the fully adapted features. The prevailing trend is still to exploit the SL patterns where possible. On the one hand, it is understandably hard for translators to assimilate the considerable cross-linguistic distance in this register. On the other hand, the expected TL norm is less defined in Russian mass-media corpus than in the other registers (note the broad spread of the media texts in Russian in Fig. 3).
2. *Popular scientific* translations have the record number of shining-through indicators, but a third of them are lexical features that do not contribute much to the translationese classification according to the classifier weights and the analysis above, particularly in this register. The prevailing trend is towards adaptation, which is reasonable, if we bear in mind a clearer delineation of this register in the TL. This is the only register where the SL/TL-independent translationese features are important for the classifier. Notably, this register has a significantly lower frequency of passives and significantly higher frequency of

subordinate conjunctions than in either original English or Russian, without a cross-linguistic contrast for this feature.

3. *Fiction* has the least shining-through indicators, and yet, according to the classifier, these features rank high in importance. The second strongest tendency is over-normalisation (or russification). The pull towards the TL norm is reinforced by the considerable input from the record number (23) of fully adapted features. This register appears to be the most Russian-like in translation.
4. In *news commentary* the few fully adapted features are assigned the biggest weights. We will highlight that this register has the largest list of over-normalised features (20) with relatively high weights. The other two effects with comparably high average feature weights are anglicisation and shining-through. It looks like this register is sharply torn between the two languages.

The suggested feature sets are also fairly reliable for defining the contrastive properties of the registers. They can be used to distinguish the four text categories with 97% accuracy. However, the importance of morphosyntactic and lexical features is reverse compared to the translationese classification. The lexical features outperformed morphosyntax in register classification. Besides, we were able to capture less morphosyntactic variation across English registers than across their Russian counterparts. The translated registers exhibit clearer register distinctions than the comparable TL non-translations, especially on the lexical level. However, using morphosyntactic features only, it is more difficult to predict registers in translations than in non-translations. It means that on the structural level the translated registers are a bit less well-defined than non-translations in the TL (see Table 4). It indicates that the translation process does not level out the distinctions between the registers. Additionally, one can claim that the register conversions are exaggerated and amplified, which leads to (1) higher similarity of translated texts from one register and/or to (2) greater distances between the registers.

We put these two hypotheses to a quick test by (1) comparing the averaged distance from centroid (corpus average vector) to each text vector for translated and non-translated registers in Russian ('degree of homogeneity' measure) and by (2) measuring the Euclidean distances between the translated registers (and use the distances in Table 5 for reference).

These experiments show that (1) translations are less diverse than their non-translated counterparts in all registers; (2) the second hypothesis holds only for translated fiction, which is even stronger isolated from the other registers than in non-translations (see Fig. 4), but not for the other registers, where the relatively clear distinctions in the original Russian are blurred in translation in terms of morphosyntax.

Now, the question is whether the amount and type of translationese can be explained by the degree of the cross-linguistic similarity between the registers or they have to be attributed to the extralinguistic factors such as translational norms operating in the contemporary professional community and the other translation process variables such as the input of editors and working conditions. Or in other words, is translationese a function of the linguistic distance between registers? From our observations in Fig. 6 this not likely to be the case.

The previous research on human translations reports different results in this respect based on translationese properties induced by different SLs. Diana Santos observes that languages closeness as a factor in translations has a paradoxical effect: ‘the closer the languages the larger the quantity of false friends and cognates, both in lexicon and in grammar’, because it is easier to carry over the SL properties (Santos 1995: 64). Sominsky and Wintner concluded that ‘translationese is more pronounced, and interference is more powerful, when the two languages are more distant’ based on their classification result in the SL detection task (Sominsky and Wintner 2019: 1138).

An apparent reconciliation for these competing observations is found in (Nikolaev et al. 2020). They explore the predictability of translations and find differences between translations from structurally similar and structurally dissimilar source languages. In the former case translations tend to employ an intersection of syntactic patterns found in both languages, which makes them less rich, more repetitive, in the latter case ‘translators find it hard to fully rework the original morphosyntactic patterns and produce unpredictable/entropic non-idiomatic translations’ (Nikolaev et al. 2020).

In our setting this should be observed as the difference for the degrees of homogeneity of the respective translated corpora: the more cross-linguistically similar registers (fiction and news commentary) should demonstrate higher degree of homogeneity in translation. This was indeed observed in our data where the averaged vector distance to centroid was 3.050 and 2.488 for *fiction* and *news commentary*. For more distant registers—*media* and *popular science*—this measure returned 3.354 and 3.281. Note that for distances the smaller numbers mean more similar texts.

6 Conclusion

In this chapter we investigated the impact of register on the properties of translations in the English-Russian language pair. We used parallel corpora of professional translations and comparable reference corpora from the national corpora in four registers (general media, popular science, fiction, news commentary) to explore the relations of the original texts in the two languages and the translated registers. Our approach exploits linguistically interpretable features and is contingent on their selection and effectiveness for capturing differences between registers, on the one hand, and translationally relevant text types (sources, targets, and TL reference), on the other. For both tasks we tested and described the behaviour of 45 morphosyntactic and 11 lexical features. The former represent the text structure in terms of general text properties, frequencies of PoS and syntactic phenomena, the latter provide text characteristics from the point of view of lexical predictability scores and the ratios of high-frequency and low-frequency n-grams.

The results demonstrate that our experimental setup, including the suggested features, is reliable for distinguishing registers in translated and non-translated language as well as for predicting translations in each register, and, therefore, can be used for revealing the register-related specificity of translations in the given language pair. Admittedly, the features used are language pair specific, and out findings apply for English-to-Russian translation. We leave testing the suggested methodology on other language pairs for future work.

Our findings contribute to the understanding of the linguistic properties of Russian translations from English in general and to the investigation of their specificity across registers. We suggested a distance-based method to estimate the general shapes of translationese in a register-balanced corpus for comparative analysis, taking into account the cross-linguistic properties of each register. A novel bottom-up approach was used to associate the linguistic features with a number of translationese effects and to disentangle the opposite translational tendencies.

We demonstrated that (1) professional translations in all registers are easily distinguishable from non-translations and these distinctions mostly involve morphosyntactic, rather than lexical, properties; (2) more than a third of all translationese indicators have their frequencies shifted towards the values observed in the SL (shining-through features), but their actual impact on the classification results varies and can be overshadowed by strong features representing other trends; (3) each register generates a unique form of translationese, with the various translationese effects contributing to a different extent and being realised through widely diverging sets of features; (4) translated registers have more regularity in feature frequencies and higher intra-category homogeneity than their non-translated English and Russian counterparts. The more cross-linguistically similar registers seem to generate the more homogeneous translations.

One important message from this research is that human translations vary depending on the register. Some of this variation can be explained linguistically. However, some of the translation strategies are likely to be dictated by the established practice and professional norms operating in each register, including the tolerance to translationese.

The scope of this work did not allow us to perform in-depth analysis of the individual features that were identified as having translationally interesting behaviours. The machine learning results can be convincing mathematically, but they remain a noumenon unless they are related to human perception.

Although this research takes into account the specificity of the given language pair, it would certainly be interesting to extend it to other target languages or language pairs. The more immediate development would be to consider other registers in the explored language pair, if the necessary corpus resources are available. We hope that this research will promote the idea that register is one of the central factors in translationese studies, even if its impact on the translation properties is not defined by purely linguistic matters.

Acknowledgements The research presented in this paper has been partially carried out in the framework of projects in the framework of the projects VIP (FFI2016-75831-P), TRIAGE (UMA18-FEDERJA-067) and MI4ALL (CEI-RIS3). The authors would like to thank two anonymous reviewers for their valuable comments.

Appendix

The UD-based and list-based features in alphabetical order.

Preliminary Notes

1. Normalisation measures

We use several norms to make features comparable across different-size corpora, depending on the nature of the feature. Most of the features, including all types of discourse markers, negative particles, passives, types of verb forms, relative clauses, correlative constructions, adverbial clauses introduced by pronominal adverbs coordinating and subordinating conjunctions, simple sentences, number of clauses per sentence, are normalised to the number of sentences (30 features). Such features as personal, possessive and other noun substitutes, nouns, adverbial quantifiers, determiners are normalised to the running words (6 features). Counts for syntactic relations are represented as probabilities, normalised to the number of sentences (7 features). Some features have their own normalisation basis: comparative and superlative degrees are normalised to the total number of adjectives and adverbs, nouns in the functions of subject, object or indirect object are normalised to the total number of these roles in the text.

2. Groups of discourse markers

The classification of connectives (discourse markers) follows the descriptions in Halliday and Hasan (1976) and in Biber et al. (1999). Table A has the number of items in each group and most frequent examples. The lists were initially produced independently from grammar reference books, dictionaries of function words and relevant research papers (for English we used Biber et al. (1999), Fraser (2006), Liu (2008); for Russian—Novikova (2008), Priyatkina (2015), Russian Grammar (Shvedova 1980) to name just a few sources for each language). After the initial selection, the lists were verified for comparability. Following Fraser (2006), discourse markers are treated functionally and include items of various morphological and structural types (conjunctions, adverbs, particles, parenthetical phrases). Though most items on the lists are set phrases, we allowed for possible lexical and structural variability at the extraction time. We also used orthography and punctuation to disambiguate our items. The output of the extraction procedure was manually checked to exclude greedy matching.

Table 8 Number of listed connectives and discourse markers by category for each of the project languages and top five most frequent items

	English	Russian
Additive	52	52
	Also, such as, for example, not only, for instance, in particular, moreover, in other words, namely	Также, при этом, например, кроме того, в частности, к тому же, на самом деле, а именно, иными словами, точнее, причем, вдобавок
Adversative	46	34
	Still, however, rather than, instead, though, on the other hand, in fact, despite	Однако, хотя, впрочем, правда, несмотря на, в отличие от, вместе с тем, всё-таки, но на самом деле, наоборот, напротив, зато
Causative	42	49
	Because, so, due to, so that, therefore, as a result, after all, for this reason, consequently	Потому, поэтому, поскольку, ведь, так,, в результате, ради того, чтобы, затем, что, получается, в этом случае, в связи с тем, дабы, тем более что
Temporal and sequential	110	48
	While, since, soon, and then, eventually, further, anyway, thus, at the same time, ultimately, meanwhile	Пока, наконец, затем, в целом, в то время, как, в заключение, в конце концов, во-первых, в то же время
Epistemic markers	64	86
	Really, at least, perhaps, of course, probably, in any case, for sure, in reality, no doubt, arguably, clearly, indeed, I/we think, I/we am/are (un) convinced/sure	Конечно, возможно, может быть, действительно, говорят, на мой взгляд, якобы, полагаю, по сути, в любом случае, кажется, бесспорно, пожалуй

3. The alphabetic list of 45 morphosyntactic features

acl

finite and non-finite clausal modifier of noun (adjectival clause), including relative clauses as a subtype (used only in EN and RU); extraction is based on UD default annotation (e.g. *the person showing (acl) her around; help people do something to overcome (acl) it; людей, следящих (acl) за политикой)*

addit

additive connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

advers

adversative (contrastive) connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

attrib
adjectives and participles functioning as attributes; all words tagged as ADJ or VerbForm = Part with the *amod* dependency to their head (e.g. *the rising sun; the coloured face; fried green tomatoes)*

aux
auxiliary verbs; extraction is based on UD default annotation

aux:pass
auxiliary verbs in passive forms; extraction is based on UD default annotation

but
contrastive coordinating conjunction *but* (*но*), if not followed but *also/и*, *также* and not in the absolute sentence end

caus
causative connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

ccomp
clausal complement as annotated in UD (e.g. *help people to do (ccomp) smth; не ожидали, что придет (ccomp))*

cconj
coordinating conjunctions: lemmas in *and, or, both, yet, either, &,* nor*, plus, neither, ether / и, а, или, ни, да, причем, либо, зато, иначе, только, ан, и/или, иль* tagged CCONJ. Lists are used to filter out noise.

comp
comparative degree of comparison for adjectives and adverbs; synthetic forms are extracted based on the tag Degree = Comp, while analytical forms are counted as adjectives and adverbs with a dependent *more/более* (*больший)*

copula
copula verbs; lemmas of *be, быть, это* that have a *cop* relation to their head, excluding constructions with *there* as head for English

correl
correlative constructions of all types, where a PRON/DET (*those, such*) is syntactically or semantically connected to subsequent CONJ. In English they make a subset of relative clauses; in Russian they can also be a subtype of a clausal complement (e.g. *of those who voted for him, raising the living standards of those that are poor)*

demdets
pronominal determiners; lemmas in the function *det* from the lists *this, some, these, that, any, all, every, another, each, those, either, such / этот, весь, тот, такой, какой, каждый, любой, некоторый, какой-то, один, сей, это, всякий, некий, какой-либо, какой-нибудь, кое-какой*

deverbals
deverbal nouns, names of processes, actions, states. The extraction for English accounts for affixation (with most productive *-ment, -tion/ -ung, -tion*) and conversion as types of derivation. In the first case the output is filtered with an empirically driven stop list. Converted nouns are counted from a list of true procedural nouns that were not fully substantivised. To produce this list we looked through the nounal occurrences of lemmas that also appear as verbs and filtered out items that prevail in their fully substantivised lexico-semantic variants in our data (such as *design, set, measure, mark, press, stick, cross, trap, handle*). For Russian we extracted nouns in *-тие, -ение, -ание, -ство, -ция, -ота* and employed a 150-items long stop list to exclude fully substantivised words such as *собрание, месторождение, министерство, телевидение, творчество, решение*.

epist
epistemic stance discourse markers; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

finites
verbs in finite form; extraction is based on UD default annotation VerbForm = Fin

indef
noun substitutes, i.e. pronouns par excellence, of indefinite, total and negative semantic subtypes; extraction is based on PRON tag with a filter list: *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, somebody, someone, something, elsewhere, nowhere, everywhere, somewhere, anywhere / когда, где, куда, откуда, отчего, почему, зачем* and words with *-то*|*-нибудь*|*-либо*, except starting with *какой*; and items from *кто-кто, кого-кого, кому-кому, кем-кем, ком-ком, что-что, чего-чего, чему-чему, чем-чем, куда-куда, где-где*

infs
infinitives: all cases of a verb form tagged VerbForm = Inf with a dependent *to* particle and cases of true bare infinitive, excluding after modal verbs and *have to*, *going to* and modal adjectival predicates, but including cases after *help, make, bid, let, see, hear, watch, dare, feel*. For Russian all occurrences of verb forms with the feature VerbForm = Inf except after modal predicates and with the dependent *быть* to exclude future forms (e.g. *отношения будут ухудшаться*).

interrog
interrogative sentences: all sentences ending in ?

lexdens

lexical density: ratio of PoS disambiguated content words types (look_VERB vs look_NOUN) to all tokens

lexTTR
lexical type-to-token ratio: ratio of PoS disambiguated content words types (look_VERB vs look_NOUN) to their tokens. Content words include lemmas in ADJ, ADV, VERB, NOUN part-of-speech categories.

mdd
mean dependency distance (MDD, aka comprehension difficulty) as ‘the distance between words and their parents, measured in terms of intervening words’ (Jing and Liu 2015: 162)

mhd
mean hierarchical distance (MHD, aka production (speaker’s difficulty) as the average value of all path lengths travelling from the root to all nodes along the dependency edges (Jing and Liu 2015: 164)

mpred
modal predicates; for English all verbs tagged as MD in XPOS, except *will/shall*, constructions with have-to-Inf and all adjectival modal predicates (given a list of 17 predicatives such as *impossible, likely, sure* with a dependent AUX). For Russian: lemma *мочь*, lemma *следовать* with a dependent infinitive, three modal adverbs (*можно, нельзя, надо*) and 11 adjectives from the modal predicative list in the short form Variant = Short (e.g. *должен, способный, возможный*)

mquantif
adverbial quantifiers; listed lemmas tagged ADV. The support lists include 37 English items (e.g. *barely, completely, intensely, almost*), 80 Russian items (*абсолютно, полностью, сплошь, необыкновенно, достаточно, совершенно, невыносимо, примерно*). For Russian we additionally provide for functionally similar non-adverbial quantifiers such as *еле, очень, вшестеро, невыразимо, излишне, еле-еле, чуть-чуть, едва-едва, только, капельку, чуточку, едва*.

neg
negative particles or main sentence negation: counts of lemmas in *no, not, neither* / нет, не

nnargs
core verbal arguments represented by nouns or proper names; ratio of nouns and proper names in the functions of *nsubj, obj, iobj* to the count of these functions

nsubj:pass
subjects of verbs in the passive voice; extraction is based on UD default *nsubj:pass* annotation

numcls
number of clauses per sentence; number of relations from the list *csubj, acl:relcl, advcl, acl, xcomp, parataxis* annotated in one sentence

passives
passive constructions with expressed agentive role; all verbs tagged Voice = Pass and a dependent aux:pass (for English). For Russian we account for two morphological forms (*война велась, политика была направлена*) and for semantic passive (*стадион возводят на новом месте, во Владикавказе ему готовят радушную встречу*)

parataxis
asyndatically connected coordinated clauses (often direct speech or clauses joined ':' or a ';' as well as parenthetical clauses); extraction is based on UD default annotation

pasttense
verbs in the past tense: all occurrences of the feature Tense = Past

pied
correlative constructions with displaced (pied-piped) preposition (e.g. *technology for which Sony could take credit; speech in which he made this argument; о таком, о каком вы не слыхали; скандал, в котором; трагедии, с которыми, в той конструкции, в какой она*)

possdet
possessive pronouns; for English lemma in *my, your, his, her, its, our, their* tagged DET, PRON and Poss = Yes. For Russian lemma in *мой, твой, ваш, его, ее, её, наш, их, ихний, свой* tagged DET

ppron
personal pronouns; tokens tagged PRON, with any value of attribute Person = that do not have Poss = Yes feature and are on the list: *i, you, he, she, it, we, they, me, him, her, us, them / я, ты, вы, он, она, оно, мы, они, меня, тебя, его, её, ее, нас, вас, их, неё, нее, него, них, мне, тебе, ей, ему, нам, вам, им, ней, нему, ним, меня, тебя, него, мной, мною, тобой, тобою, Вами, им, ей, ею, нами, вами, ими, ним, нем, нём, ней, нею*

pverbals
participles: for English all occurrences of VerbForm = Part or VerbForm = Ger not in attributive function *amod* or part of an analytical form. For Russian VerbForm = Part not in the short form and not in the attributive function, without a dependent auxiliary, and VerbForm = Conv without dependent auxiliary (e.g. a*fter years of translating emails, webinars and other materials*)

relativ
all relative clauses, including correlative constructions and pied-piping construction. Extraction is based on affirmative sentences only. For English: *which, that, whose, whom, what, who* tagged as PRON, excluding cases when relative PRON has a dependent preposition and follows its head (e.g. *But we will return to that (PRON) later*). For Russian: *который, что, кто, какой* and a comma in the left window of 3

sconj
subordinating conjunctions: lemma in *that, if, as, of, while, because, by, for, to, than, whether, in, about, before, after, on, with, from, like, although, though, since, once, so, at, without, until, into, despite, unless, whereas, over, upon, whilst, beyond, towards, toward, but, except, cause, together / что, как, если, чтобы, то, когда, чем, хотя, поскольку, пока, тем, ведь, нежели, ибо, пусть, будто, словно, дабы,раз, насколько, тот, коли, коль, хоть, разве, сколь,ежели, покуда, постольку* tagged SCONJ. Lists are used to filter out noise.

sentlength
number of words per sentence averaged over all sentences in the text. The extraction accounts for typical sentence tokenisation errors such as sentences ending in:,;, Mr., Dr.

simple
simple sentence; a sentence where no words have relations: *csubj, acl:relcl, advcl, acl, xcomp, parataxis*

sup
superlative degree of comparison for adjective and adverbs; synthetic forms are extracted based on the tag Degree = Sup, while analytical forms are counted as adjectives and adverbs with a dependent most/наиболее/самый and for Russian words starting with *наи-* with the exception of a few homonymous adverbs (*наискосок*)

tempseq
temporal and sequential connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

whconj
adverbial clause introduced by a pronominal ADV *when, where, why / когда, где, куда, откуда, отчего, почему, зачем*

xcomp
a predicative or clausal complement without its own subject, annotated after phrasal verbs (e.g. *started to sing*), in case of infinitive constructions (e.g. *asked me to leave*), etc.; extraction is based on UD default annotation

References

Aharoni, R., M. Koppel, and Y. Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL 2014)*, Vol. 1: Long Papers, ed. K. Toutanova, and H. Wu, 289–295. Association for Computational Linguistics https://doi.org/10.3115/v1/p14-2048.

Arase, Y., and M. Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st annual meeting of the association for computational linguistics*, Vol. 1: Long Papers, ed. H. Schütze, F. Pascale, and M. Poesio, 1597–1607. Association for Computational Linguistics.

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and technology: In honour of John Sinclair*, ed. M. Baker, G. Francis, and E. Tognini-Bonelli, 232–250. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.64.15bak.

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and translation: Studies in language engineering, in honour of Juan C. Sager*, ed. H. Somers, 175–186. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/btl.18.17bak.

Baroni, M., and S. Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21 (3): 259–274. https://doi.org/10.1093/llc/fqi039.

Becher, V. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts* [Doctoral dissertation, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky]. https://ediss.sub.uni-hamburg.de/bitstream/ediss/4186/1/Dissertation.pdf.

Biber, D. 1988. *Variation across speech and writing*, 2nd ed. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871.

Biber, D., and S. Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and R. Quirk. 1999. *Longman grammar of spoken and written English*, vol. 2. Cambridge, MA: The MIT Press.

Castagnoli, S. 2009. *Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation* [Doctoral dissertation, University of Pisa, Italy]. ETD System, electronic theses and dissertations repository. https://etd.adm.unipi.it/t/etd-04252009-135411/.

Castagnoli, S., D. Ciobanu, K. Kunz, N. Kübler, and A. Volanschi. 2011. Designing a learner translator corpus for training purposes. In *Corpora, language, teaching, and resources: From theory to practice*, Vol. 12, ed. N. Kubler, 221–248. Frankfurt: Peter Lang.

Chang, Y., and C. Lin. 2008. Feature ranking using linear SVM. In *Proceedings of the workshop on the causation and prediction challenge at WCCI 2008*, ed. I. Guyon, C. Aliferis, and G. Cooper, 53–64. Proceedings of Machine Learning Research.

Corpas Pastor, G. 2008. *Investigar con corpus en traducción: Los retos de un nuevo paradigma*. Frankfurt: Peter Lang. https://doi.org/10.4000/bulletinhispanique.1301.

Corpas Pastor, G., R. Mitkov, N. Afzal, and V. Pekar. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th conference of the association for machine translation in the Americas (AMTA'08)*, 21–25.

Delaere, I. 2015. *Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity*. [Doctoral dissertation, Ghent University]. Academic Bibliography. https://biblio.ugent.be/publication/5888594.

Dipper, S., M. Seiss, and H. Zinsmeister. 2012. The use of parallel and comparable data for analysis of abstract anaphora in German and English. In *Proceedings of the 8th international* *conference on language resources and evaluation (LREC 2012)*, ed. N. Calzolari, Kh. Choukri, Th. Declerck, M. Uğur Doğan, et al., 138–145. European Language Resources Association.

Diwersy, S., S. Evert, and S. Neumann. 2014. A semi-supervised multivariate approach to the study of language variation. In *Linguistic variation in text and speech, within and across languages*, ed. B. Szmrecsanyi, and B. Wälchli, 174–204. Berlin: De Gruyter Mouton.

Duff, A. 1981. *The third language: Recurrent problems of translation into English*. Oxford: Pergamon.

Eetemadi, S., and K. Toutanova. 2015. Detecting translation direction: A cross-domain study. In *Proceedings of NAACL-HLT 2015 student research workshop (SRW)*, ed. D. Inkpen, S. Muresan, Sh. Lahiri, K. Mazidi, and A. Zhila, 103–109. https://doi.org/10.3115/v1/N15-2014.

Evert, S., and S. Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In *Empirical translation studies: New methodological and theoretical traditions*, vol. 300, ed. G. De Sutter, M. Lefer, and I. Delaere, 47–80. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110459586-003.

Fraser, B. 2006. Towards a theory of discourse markers. In *Approaches to discourse particles*, ed. K. Fischer, 189–204. London: Elsevier.

Frawley, W. 1984. Prolegomenon to a theory of translation. In *Translation: Literary, linguistic & philosophical perspectives*, ed. W. Frawley, 159–175. Newark: University of Delaware Press.

Gellerstam, M. 1986. Translationese in Swedish novels translated from English. In *Translation studies in Scandinavia*, ed. L. Wollin and H. Lindquist, 88–95. Lund: CWK Gleerup.

Goutte, C., D. Kurokawa, and P. Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of the 12th machine translation summit (MT Summit XII)*, 81–88.

Graham, Y., B. Haddow, and P. Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 72–81). Association for Computational Linguistics.

Halliday, M.A.K., and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Halliday, M., and R. Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective* (2nd ed.). Oxford University Press.

Hansen-Schirra, S. 2011. Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language. In *Multilingual discourse production: Diachronic and synchronic perspectives*, ed. S. Kranich, 133–162. Amsterdam: John Benjamins.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 sixth workshop on statistical machine translation*, ed. Ch. Callison-Burch, Ph. Koehn, Ch. Monz, and O. Zaidan, 187–197. Association for Computational Linguistics.

Ilisei, I., D. Inkpen, G. Corpas Pastor, and R. Mitkov. 2010. Identification of translationese: A machine learning approach. *International conference on intelligent text processing and computational linguistics*, 503–511.

Jiang, Z., and Y. Tao. 2017. Translation universals of discourse markers in Russian-to-Chinese academic texts: A corpus-based approach. *Zeitschrift Fur Slawistik* 62 (4): 583–605. https://doi.org/10.1515/slaw-2017-0037.

Jing, Y., and H. Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, ed. J. Nivre and E. Hajicova, 161–170. Uppsala University.

Karakanta, A., and E. Teich. 2019. Detecting and analysing translationese with probabilistic language models translationese. In *Translation in Transition* 4: 38–39.

Katinskaya, A., and S. Sharoff. 2015. Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proceedings of the 5th workshop on Balto-Slavic natural language processing*, ed. J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev, and R. Yangarber, 65–74. INCOMA Ltd. http://www.aclweb.org/anthology/W15-5311.

Koppel, M., and N. Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, Vol.

1, ed. D. Lin, Yu. Matsumoto, and R. Mihalcea, 1318–1326. Association for Computational Linguistics.

Kruger, H., and B. Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13 (1): 33–65. https://doi.org/10.1556/Acr.13.2012.1.3.

Kruger, H., and B. van Rooy. 2010. The features of non-literary translated language: A pilot study. In *Proceedings of using corpora in contrastive and translation studies (UCCTS 2010)*, ed. R. Xiao, 59–79.

Kunilovskaya, M. 2017. Linguistic tendencies in English to Russian translation: The case of connectives. In *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2017"*, Vol. 2, ed. V.P. Selegey, A.V. Baytin, V.I. Belikov, I.M. Boguslavsky, B.V. Dobrov, et al., 221–233. Computational Linguistics and Intellectual Technologies.

Kunilovskaya, M., and A. Kutuzov. 2018. Universal dependencies-based syntactic features in detecting human translation varieties. In *Proceedings of the 16th international workshop on treebanks and linguistic theories (TLT16),* ed. J. Hajič, 27–36. Association for Computational Linguistics.

Kunilovskaya, M., and E. Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)*, ed. N. Calzolari, F. Bechet, Ph. Blache, Kh. Choukri, et al., 4102–4112. The European Language Resources Association (ELRA).

Kunilovskaya, M., and E. Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the 2nd workshop on human-informed translation and interpreting technology (HiT-IT 2019)*, ed. I. Temnikova, C. Orasan, G. Corpas Pastor, and R. Mitkov, 47–56. INCOMA Ltd. https://doi.org/10.26615/issn.2683-0078.2019_006.

Kutuzov, A., and M. Kunilovskaya. 2014. Russian learner translator corpus: Design, research potential and applications. In *Proceedings of the 17th international conference text, speech and dialogue*, vol. 8655, ed. P. Sojka, A. Horák, I. Kopeček, and K. Pala, 315–323. Springer.

Lapshinova-Koltunski, E. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In *Empirical translation studies. New theoretical and methodological traditions*, vol. 300, ed. G. De Sutter, M. Lefer, and I. Delaere, 207–234. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110459586-008.

Lapshinova-Koltunski, E., and M. Zampieri. 2018. Linguistic features of genre and method variation in translation: A computational perspective. *The grammar of genres and styles: From discrete to non-discrete units*, (TiLSM, *320*), 92–112. Berlin: De Gruyter Mouton.

Lee, D.Y.W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5 (3): 37–72. https://doi.org/10.1016/S1364-6613(00)01594-1.

Lembersky, G., N. Ordan, and S. Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38 (4): 799–825. https://doi.org/10.1162/COLI_a_00111.

Lijffijt, J., T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki, and H. Mannila. 2016. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31 (2): 374–397. https://doi.org/10.1093/llc/fqu064.

Liu, D. 2008. Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics* 13 (4): 491–518. https://doi.org/10.1075/ijcl.13.4.05liu.

Martin, J.R. 1992. *English text: System and structure*. Amsterdam: John Benjamins.

Nakamura, S. 2007. Comparison of features of texts translated by professional and learner translators. In *Proceedings of the 4th corpus linguistics conference.* University of Birmingham.

Neumann, S. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: De Gruyter Mouton.

Nikolaev, D., T. Karidi, N. Kenneth, V. Mitnik, L. Saeboe, and O. Abend. 2020. Morphosyntactic predictability of translationese. *Linguistics Vanguard*, 6 (1).

Nini, A. 2019. The multi-dimensional analysis tagger. In *Multi-dimensional analysis: research methods and current issues*, ed. T. Berber Sardinha, and M. Veirano Pinto, 67–94. London; New York: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0012.

Nisioi, S., and L.P. Dinu. 2013. A clustering approach for translationese identification. In *Proceedings of the international conference recent advances in natural language processing (RANLP 2013),* ed. R. Mitkov, G. Angelova, and K. Bontcheva, 532–538. INCOMA Ltd. http://www.aclweb.org/anthology/R13-1070.

Novikova, N.I. 2008. Connectives as cohesive devices in an asyndetic composite sentence [Konnektory kak svjazujushhie sredstva v bessojuznom slozhnom predlozhenii]. In Herald of the Voronezh state Architecture University, advanced linguistic and pedagogical research series [Ser.: Sovremennye lingvisticheskie i metodiko-didakticheskie issledovanija], 92–100.

Olohan, M. 2001. Spelling out the optionals in translation: A corpus study. *UCREL Technical Papers* 13: 423–432.

Popescu, M. 2011. Studying translationese at the character level. In *Proceedings of the international conference recent advances in natural language processing (RANLP 2011)*, 634–639. http://aclweb.org/anthology/R11-1091.

Popovic, M. 2020. On the differences between human translations. In *Proceedings of the 22nd annual conference of the European association for machine translation*, ed. A. Martins, H. Moniz, S. Fumega, M. Martins, F. Batista, L. Coheur, C. Parra, … M. Forcada, 365–374. European Association for Machine Translation.

Prieels, L., I. Delaere, K. Plevoets, and G. De Sutter. 2015. A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation. *Across Languages and Cultures* 16 (2): 209–231. https://doi.org/10.1556/084.2015.16.2.4.

Priyatkina, A.F., E.A. Starodumova, G.N. Sergeeva, et al. (eds.). 2001. *A Russian dictionary of functional words [Slovar' sluzhebnyh slov russkogo jazyka]*. Vladivostok: Far-East State University Press.

Puurtinen, T. 2003. Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing* 18 (4): 389–406. https://doi.org/10.1093/llc/18.4.389.

Rabadán, R., B. Labrador, and N. Ramón. 2009. Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment. *Babel* 55 (4): 303–328. https://doi.org/10.1075/babel.55.4.01rab.

Rabinovich, E., and S. Wintner. 2013. Unsupervised identification of tr association for computational linguistics anslationese. *Transactions of the Association for Computational Linguistics* 3: 419–432. https://doi.org/10.1162/tacl_a_00148.

Redelinghuys, K. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: A corpus-based study. *Stellenbosch Papers in Linguistics* 45: 189–220. https://doi.org/10.5774/45-0-198.

Santini, M., A. Mehler, and S. Sharoff. 2010. Riding the rough waves of genre on the web concepts and research questions. In *Genres on the web: Computational models and empirical studies*, vol. 42, ed. A. Mehler, S. Sharoff, and M. Santini, 3–30. Springer Science & Business Media.

Santos, D. 1995. On grammatical translationese. In *Proceedings of the 10th Nordic conference of computational linguistics (NODALIDA 1995),* ed. K. Koskenniemi, 59–66. University of Helsinki.

Sharoff, S. 2018. Functional text dimensions for annotation of web corpora. *Corpora* 13 (1): 65–95. https://doi.org/10.3366/cor.2018.0136.

Shvedova, N. (ed.). 1980. *Russian grammar*. Moscow, Science [Nauka].

Sominsky, I., and S. Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2019)*, ed. R. Mitkov and G. Angelova, 1131–1140. INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_130.

Specia, L., G.H. Paetzold, and C. Scarton. 2015. Multi-level translation quality prediction with QUEST++. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, ed. H. Chen, and K. Markert, 115–120. Association for Computational Linguistics. https://doi.org/10.3115/v1/p15-4020.

Straka, M., and Straková, J. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies,* ed. D. Zeman, J. Hajic, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, 88–99. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-300.

Stymne, S. 2017. The effect of translationese on tuning for statistical machine translation. In *Proceedings of the 21st Nordic conference of computational linguistics,* ed. J. Tiederman, 241–246. Linköping University Electronic Press.

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts.* (TTCP, 5). Berlin: De Gruyter Mouton.

Toury, G. 1995. *Descriptive trantslation studies-and beyond*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4.

Vela, M., and E. Lapshinova-Koltunski. 2015. Register-based machine translation evaluation with text classification techniques. In *Proceedings of the 15th machine translation summit (Vol. 1: MT Researchers' Track),* ed. Y. Al-Onaizan, and W. Lewis, 215–228. Association for Machine Translation in the Americas.

Volansky, V., N. Ordan, and S. Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30 (1): 98–118. https://doi.org/10.1093/llc/fqt031.

Xiao, R., L. He, and Y. Ming. 2010. In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In *Using corpora in contrastive and translation studies*, ed. R. Xiao, 182–214. New Castle: Cambridge Scholars Publishing.

Zanettin, F. 2013. Corpus methods for descriptive translation studies. *Procedia-Social and Behavioral Sciences* 95: 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618.

Zhang, M., and A. Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the fourth conference on machine translation (Volume 1: Research Papers),* ed. O. Bojar, R. Chatterjee, Ch. Federmann, M. Fishel, Y. Graham, ... K. Verspoor, 73–81. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5208.

Maria Kunilovskaya PhD in contrastive linguistics, has an extensive experience in translator education and corpus linguistics. Her research is recently focused on human translation quality estimation, translationese studies and varieties, building and exploiting comparable and parallel corpora. Maria's other interests include computational and empirical approaches to comparing languages and understanding translation processes.

Gloria Corpas Pastor PhD in English philology, Professor in translation, interpreting and translation technology and an active member of several international and national editorial and scientific committees. She is the head of the LEXYTRAD research group. Her research interests include specialised translation and interpreting technologies, corpus-based translation studies, phraseology, lexicography and terminology.

Normalization, Motivation, and Reception: A Corpus-Based Lexical Study of the Four English Translations of Louis Cha’s Martial Arts Fiction

Kan Wu and Dechao Li

Abstract Combining both qualitative and quantitative methods in the research design, the present study examines the normalizing tendencies in the translations of Louis Cha’s martial arts fiction in such categories as lexical richness, normalized POS distributions, high-frequency words, and the naturalizing percentages of special martial-arts terminology. The result shows that among the four existing translations of the fiction, Minford’s one is marked by the highest degree of lexical normalization and enjoys the best reception in overseas markets. Closely following Minford’s tendency of lexical choices, Holmwood’s translation has gained instant popularity since its debut in 2018. Meanwhile, Mok’s rendition, which has a medium level of lexical normalization, has received the most negative feedback from foreign readers. On the other hand, Earnshaw’s translation, which has the lowest level of lexical normalization, has the most divided and varied reception among its readers. It is further revealed that the translators’ decisions for their lexical choices are highly governed by their translating motivation, which in turn affects the reception of their translations.

Keywords Normalization · Reception · Translation of martial arts fiction · Louis Cha

This is an updated and expanded version of the authors’ Chinese paper 基於語料庫的英譯金庸小說詞彙範化研究及啟示 “Lexical Normalization in English Translations of Jin Yong’s Martial Arts Fiction: A Corpus-based Study” which was published in 《亞太跨學科翻譯研究》(第七輯)“Asia Pacific Interdisciplinary Translation Studies” (issue no. 7), Beijing: Tsinghua University Press, 2018/2, pp. 93–106.

A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y202044693).

K. Wu (✉)
School of Foreign Languages, Zhejiang University of Finance and Economics Dongfang College, No.2, Yangshan Rd., Haining, Zhejiang, China
e-mail: wukanq@163.com

D. Li
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, No. 11 Yuk Choi Rd., Hung Hom, Kowloon, Hong Kong, China
e-mail: dechao.li@polyu.edu.hk

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_7

1 Introduction

Normalization in translation refers to the tendency of a translated language following the norms of the target language (Baker 1993). Since Mona Baker started the research of applying corpora to the study of the linguistic patterns of translation (such as explicitation, simplification and convergence, among others) in the early 1990s, scholars from Europe (May 1997; Kenny 2001; Williams 2005; Ippolito 2014; Moreno 2016; Frankenberg-Garcia 2017, etc.) have examined normalizing tendencies at either lexical or sentential levels with bilingual parallel corpora and/or monolingual comparable corpora, highlighting the ways that culture diversities between source language (SL) and target language (TL) have affected such tendencies in translated languages. In China, scholars (Hu 2007; Hu 2011; Xia 2014; Wang 2016; Wang et al. 2018, etc.) investigated the normalizing features in Chinese translations of English literature with self-compiled English-Chinese parallel corpora and/or Chinese comparable corpora, exploring diachronic shifts of their features and the causes of them. Indeed, these studies have deepened our understanding of normalization in translated languages across different language pairs from multiple perspectives. But since most of the findings are based on general balanced corpora that are not meant for research into a specific text type, they cannot be used to explain the linguistic patterns for a particular genre that is not included in the corpora, such as martial arts fiction.

Louis Cha, also known as Jin Yong, is a Hong Kong martial arts fiction writer. For decades, his stories are among the most popular literary readings for ordinary Chinese readers both home and abroad, making him one of the best-known writers among Chinese readers around the globe. In recent years, the translation of martial arts fiction has increasingly attracted research attention in literary translation studies (Luo 2011; Xiao 2012; Lu 2014; Hong 2014, etc.) under the aegis of the Chinese national project *"Chinese Literature Going Overseas."* Although these studies have greatly enhanced our understanding about the translation of this unique literature genre, they more often than not abound with analyses of text samples that are randomly chosen, which leaves the readers with a feeling of impressionism and subjectivity on the part of the researchers. There is an obvious lack of study that examines lexical normalization from both qualitative and quantitative (i.e., corpus-based) approaches.

To fill this gap, the present paper attempts to explore normalization in the English translations of Louis Cha's martial arts fiction through a specialized literary corpus, the "*Chinese-English Parallel Corpus of Louis Cha's Martial Arts Fiction*" (hereafter, the Louis Cha Corpus), which is a self-built specialized corpus for the research on the English translation of Louis Cha's fiction. Formal features exhibited in the corpus such as lexical richness, normalized POS (part of speech) distributions, and overlapping rates of high-frequency words against those of their contemporaneous non-translated English fiction will be calculated so as to give an exact picture about the different normalization tendencies in the translations. In addition, the naturalizing rates of martial-arts-specific terminology in the target texts

(TTs) and the relationship between normalization and its possible connections with translator's motivation and reader's reception will be discussed in detail. A survey about the online receptions of these versions will also be conducted to gain better insight into the popularity of these translations in overseas book markets.

2 Texts and Corpora

2.1 The Four Complete English Translations of Louis Cha's Martial Arts Fiction

Currently, there are four complete English translations of Louis Cha's novels: *Fox Volant of the Snowy Mountain* (《雪山飛狐》) in 1993 by the Hong Kong scholar Olivia Mok, *The Deer and the Cauldron* (《鹿鼎記》) in 1997 by the British Sinologist John Minford, *The Book and The Sword* (《書劍恩仇錄》) in 2004 by the British publisher Graham Earnshaw, and *A Hero Born: Legends of the Condor Heroes* (*Vol.1)* (《射雕英雄傳》) in 2018 by the British translator Anna Holmwood. These four English translations, together with their corresponding source texts (STs), constitute the specialized parallel corpus in this research.

2.2 Corpus Design and Compilation

In addition to the specialized corpus mentioned above, a comparable corpus that contains the four English translations and the fiction subset of BNC Baby Edition[1] is used in the research as a reference corpus, against which normalizing tendencies at lexical level in the four English translations will be measured and compared.

To compile the specialized corpus, all English and Chinese texts have been first converted into "txt" files before a noise-cleaning process is conducted to ensure the integrity and accuracy of these textual data. The Chinese files (encoded as UTF-8) are then segmented automatically by Jieba, a Python package developed for Chinese natural language processing, to ensure their consistency and compatibility with Wordsmith Tools 6.0. What follows next is the POS-tagging of the English files online via the CLAWS tagger (C5 tagset) to ensure their comparability with the BNC files. Finally, Transmate Aligner, a freeware for alignment, is used to align the STs with their corresponding TTs at sentence level for comparative analysis of ST-TT normalizing shifts.

[1] This subset, which includes the mainstream non-translated British fiction in the 90s and beyond, is directly downloadable from the website of University of Oxford Text Archive (http://ota.ox.ac.uk/catalogue/index.html) for free.

3 Findings

3.1 Lexical Richness

The type-token ratio (TTR), which is obtained by dividing the total number of different words in a text by the total number of words in it, is one of the important indicators to measure the lexical richness of a text. However, the TTR values retrieved from different texts cannot be directly compared, as the varied sizes of texts can dramatically influence their accuracy in the measurement of lexical richness. To solve this problem, different linguists used different ways to measure the lexical richness of a text based on type-token information of a text. Some scholars, such as Scott (1998), used the Standardized TTR—a parameter that computes TTRs of different texts based on every 1,000 words—for this purpose; others such as Daller et al. (2003), applied the Guirand's Index ($G = \text{types}/\sqrt{tokens}$) to calculate TTR. Despite the difference in their methods of calculation, they all endeavored to minimize the influence of varied text sizes on the analysis results. For this research, type-token information regarding the four English translation and BNC Baby (fiction) is calculated by Word Smith 6.0 and listed in Table 1:

The present research selects STTR value as the parameter to measure the differences of lexical richness between each translation and BNC Baby (fiction). Based on the STTR value of the reference corpus (45.41), One Sample t Test on the STTR value of each translation is run separately. The results show that $t = 0.079$, $p = 0.925$ ($p > 0.05$), meaning no statistical significance between these English translations and BNC Baby (fiction) with regard to STTR values. This result further implies that the normalizing tendencies of these English translations are rather strong in terms of lexical richness. As shown in Table 1, the STTR value of Minford's *The Deer and the Cauldron* is of the smallest difference to that of BNC Baby (fiction) (i.e., the difference is 0.1); whereas Earnshaw's shows the greatest difference (i.e., the difference is 3.25). The STTR values of Holmwood's translation is of moderate difference (2.33 of difference) to the reference corpus, while Mok's

Table 1 Type–Token Information of the Four English translations and BNC Baby (fiction)

English translations	Translators	Type	Token	TTR	STTR
Fox Volant of the Snowy Mountain	Olivia Mok	8,565	117,750	7	44.77
The Book and Sword	Graham Earnshaw	19,506	171,857	11	48.65
The Deer and the Cauldron	John Minford	38,484	592,441	7	45.51
A Hero Born: Legends of the Condor Heroes vol.1	Anna Holmwood	7,735	126,829	6.1	43.08
BNC Baby (fiction)		36,209	1,023,554	4	45.41

rendition is of mild difference (0.64 of difference) to the same corpus. The comparison of STTR values of different versions reveals that Minford's translation shows the highest degree of normalization, Earnshaw's the lowest, Mok's moderate, and Holmwood's mild levels.

3.2 Normalized POS Distribution

To a certain extent, normalized POS (part of speech) distribution (per million) reflects the typological features of a language (Wang 2010). It is generally assumed that within the same genre of a language, the closer the POS distribution in a translation is to its counterpart in a non-translated text of the same language, the higher the normalizing tendency. In the present research, nine categories of POS (i.e., verb, noun, adjective, adverb, pronoun, conjunction, preposition, auxiliary, and article) were investigated, respectively, in the four English translations and BNC Baby (fiction). Detailed results are illustrated in Table 2.

For the sake of convenience, different text was assigned with a letter (i.e., "a," "x," "y," "z," and "w") for comparing the absolute value between them. In this research, "a" stands for the reference corpus, while "x," "y," "z," and "w" represent, respectively, for Mok's, Earnshaw's, Minford's, and Holmwood's translations. Theoretically, the difference between each of the four translations and the reference corpus across these nine POS categories is reflected in the absolute value that can be obtained by subtracting "a," respectively, from "x," "y," "z," and "w." For instance, if we are to measure such difference in terms of the use of adjective, we can subtract, respectively, the value of "a" from those of "x," "y," "z," and "w" under the column of "adj." to get the absolute value that signifies the difference.

Table 2 shows that Minford's translation bears the smallest difference to the reference corpus in terms of the frequencies of adverb, noun, and auxiliary; a mild difference regarding adjective, preposition, and verb; a moderate difference regarding pronoun and article; but the greatest difference in terms of conjunction. Mok's translation shares the greatest similarity with the reference corpus regarding the frequencies of adjective and preposition; a moderate similarity in terms of auxiliary; a mild similarity in terms of noun and conjunction; but the least similarity in terms of adverb, pronoun, verb, and article. Earnshaw's translation maintains a mild difference to the reference corpus in the frequencies of adverb, noun, pronoun, article, and conjunction; a moderate difference regarding verb; but the greatest difference in terms of adjective, preposition, and auxiliary. Holmwood's translation has the smallest difference to the reference corpus regarding the frequencies of pronoun, verb, conjunction, and article; a moderate difference regarding adjective, adverb, preposition, and auxiliary; and the greatest difference in terms of noun.

If we use "4" to indicate the largest value, "3" upper-middle value, "2" lower-middle value, and "1" the smallest value, we can calculate the absolute values of "x-a," "y-a," "z-a," and "w-a" in each of the nine POS categories in Table 2. Then, we can further sum up all these ranks across the nine categories to

Table 2 Normalized POS distribution (per million) in the English translations and reference corpus

	adj	adv	noun	pron	prep	verb	conj	aux	art	Total (Rank)
BNC Fiction (a)	87,751	84,520	208,057	135,362	109,795	216,449	55,003	15,452	80,308	
Fox Volant of the Snowy Mountain (X)	64,393	60,637	224,479	60,340	84,149	162,097	47,796	12,811	97,684	
The Book and the Sword (y)	48,216	62,219	219,029	73,923	62,841	203,460	51,603	10,860	79,598	
The Deer and the Cauldron (z)	64,067	71,474	198,362	73,304	72,175	208,330	42,166	17,238	82,638	
A Hero Born: Legends of the Condor Heroes (W)	62,134	61,157	234,517	74,568	65,748	213,873	53,679	11,564	80,437	
x-a	−23,358	−23,883	16,422	−75,022	−25,646	−54,352	−7,207	−2,641	17,376	
y−a	−39,535	−22,301	10,972	−61,439	−46,954	−12,989	−3,400	−4,592	−710	
z-a	−23,684	−13,046	−9,695	−62,058	−37,620	−8,119	−12,837	1,786	2,330	
w-a	−25,617	−23,363	26,460	−60,794	−44,047	−2,576	−1,324	−3,888	129	
Rank |x-a|Mok	1	4	3	4	1	4	3	2	4	26
Rank |y-a|Earnshaw	4	2	2	2	4	3	2	4	2	25
Rank |z-a| Minford	2	1	1	3	2	2	4	1	3	19
rank |w-a|Holmwood	3	3	4	1	3	1	1	3	1	20

get the final result: i.e., "|x-a|" equals 26, "|y-a|" 25, "|z-a|" 19, and "|w-a|" 20. This shows that in terms of total normalized POS frequencies of all these four versions, Minford's translation maintains the smallest difference to the reference corpus, Holmwood's a mild one, Earnshaw's a moderate one, and Mok's translation the greatest. As a result, the normalization tendency in the four English translations in terms of normalized POS distribution can be summed up as follows: Minford's translation has the highest degree of normalization, Holmwood's translation a moderate level, Earnshaw's a mild level, and Mok's translation the lowest level.

3.3 Overlapping Rates of High-Frequency Words

High-frequency words in the present study refer to words with highly frequent occurrences in a text. It is usually assumed that the proportion of high-frequency words of a text can reflect its linguistic features. Findings from Laviosa (1998) and McEnery et al. (2010) suggest that a translated language generally contains a higher proportion of high-frequency words than its non-translated counterpart. To operationalize the calculation process, the high-frequency words of a translation are described within certain ranges, with their results being compared with those in the reference corpus. The resultant overlapping rates are then used to examine the degree of normalization for the translation. Generally, the higher the overlapping[2] rate of high-frequency words between a translation and its reference corpus, the smaller the difference between the translation and the reference corpus in terms of the use of words and expressions. And hence, a stronger tendency of lexical normalization. The present research confines the high-frequency words in the four translations and the BNC Baby (fiction) to the following five groups, namely, Top 50, Top 100, Top 200, Top 300, and Top 400, to calculate the overlapping rates in each group. The analysis results are presented in Fig. 1.

Overall, the overlapping rates of high-frequency words between each of the four translations and BNC Baby (fiction) drop as we expand our investigation groups (i.e., from Top 50 to Top 400). A closer look at Fig. 1 reveals that Minford's translation has the highest overlapping rates among the four translations across all the five groups (i.e., from Top 400 to Top 50) with a maximum overlapping rate at 86% and a minimum rate at 73.7%. On the other hand, Holmwood's translation bears a moderate overlapping rates among the four translations across the same groups: all its rates are less than 77%, covering a range between 60.32 and 76.48%. Meanwhile, Earnshaw's translation and Mok's translation stand, respectively, in the lowest and mild range: the former falls within a range between 55.8 and 72% and the latter between 62.7 and 78%. All these statistics point to the fact that Minford's use of high-frequency words in his translation is of the smallest difference to that of

[2] Overlapping in this study refers to the situation where the same word occurs in both reference corpus and a translation at a high frequency.

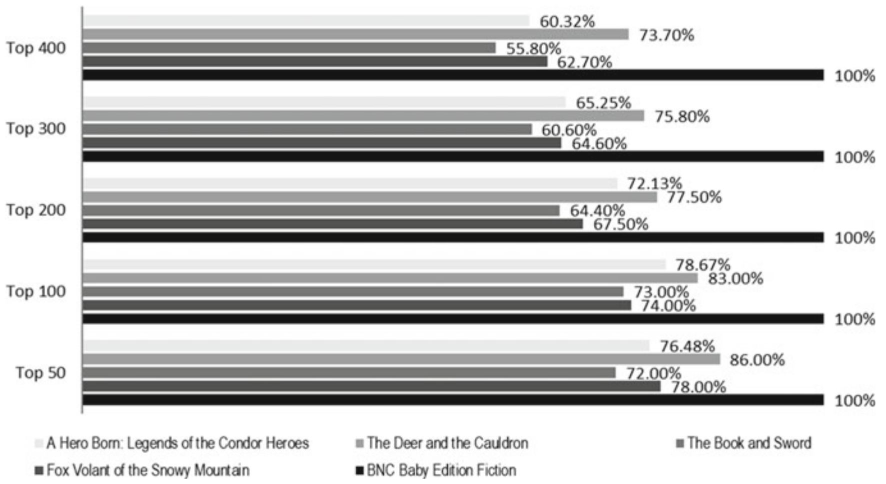


Fig.1 Overlapping rates of high-frequency words in the English translations and reference corpus

BNC Baby (fiction), indicating the highest degree of lexical normalization in his translation. Earnshaw's employment of high-frequency words in his translation is of the greatest difference to that of BNC Baby (fiction), signifying the degree of normalization of the same type in his production is the lowest; Holmwood's translation and Mok's translation are of medium differences to BNC Baby (fiction) in terms of high-frequency words use, revealing their respective translations are mild or moderate in their tendencies of lexical normalization.

3.4 Translation Shifts of Martial-Arts-Specific Terminology

When translating fictional works that are heavily loaded with cultural elements (e.g., the martial arts fiction) into English, some translators may choose to "submit by locating the same in a cultural other, pursuing a cultural narcissism that is imperialistic abroad and conservative, even reactionary, in maintaining canons at home" (Venuti 1995: 308). To operationalize as actual translating strategies, such submission could be partially reflected in the strategies of naturalizing (i.e., replace ST elements that may sound linguistically and/or culturally unfamiliar to TT readers with those that are more acceptable to them) and omitting (i.e., omit ST elements that may sound linguistically and/or culturally unfamiliar to TT readers), which are the means to dilute ST cultural elements through TT recreation. The present study assumes that the frequencies of naturalizing and omitting methods adopted by translators are closely associated with the degrees of normalization in TT, because the overall result of employing these two strategies throughout the translation would naturally lead to a more linguistically and culturally accessible version for the target reader. The direct opposite to the above two normalizing translation strategies is the foreignizing method, which retains the original cultural elements in a literal manner to "resist by locating the alien in a cultural other, pursuing cultural diversity, foregrounding the linguistic and cultural differences of the source-language text and transforming the hierarchy of cultural values in the target language" (Venuti 1995: 308).

Based on these definitions, we then categorized the four translators' translation strategies of the two types of martial-arts-specific terms (i.e., Kungfu names and character names) into three types (i.e., naturalizing, foreignizing, and omitting) and calculated the total normalizing rates (i.e., naturalizing rates plus omitting rates) by using the Concord function of Wordsmith 6.0. In addition to naturalizing rates, omitting rates are also included to calculate the total normalizing rates, because we regard omitting in translating activities as a special and extreme form of naturalizing. Details of these frequencies are illustrated in Table 3.

As shown in Table 3, Minford mainly resorts to the strategy of naturalizing when translating these terms. For the 26 Kungfu names in the original, Minford naturalized 19 out of 26 in his translation, foreignized 7, and omitted none, posting a total normalizing rate of 73.1%. He also naturalized 96 out of 183 character names, which translates into a total normalizing rate of 52.4%. By contrast, Mok shows a tendency of favoring foreignizing when she translated these terms in *Fox Volant of the Snowy Mountain*. Of all the 20 Kungfu names in the ST, she only naturalized four terms in her translation (20.0% total normalizing rate), but foreignized the other 16. Similarly, she only naturalized 35 out of 317 character names (11.0% total normalizing rate), but foreignized all the rest terms (i.e., 282 of them). Like Minford, Holmwood adopted a similar preference for translating the similar terms in *A Hero Born: Legends of the Condor Heroes*, as her total normalizing rate is almost as high as Minford's. Although her total normalizing rate for all the 56 Kungfu names is 48.2% (with 27 terms being naturalized), which is a rather modest number, her total naturalizing rate for the 62 character names is much higher: altogether she normalized 47 of the terms, giving it a 75.8% of total normalizing rate. In fact, almost all the major characters in the novel have all been translated into idiomatic English, such as "梅超風" as "Cyclone Mei," "歐陽克"as "Gallant Ouyang," or "韓寶駒" as "Ryder Han." Different from Minford and Holmwood's practices, Earnshaw adopted translation strategies that are more or less similar to those of Mok when dealing with these terms, clearly favoring the foreignizing strategy. For example, he only naturalized 4 out of 31 Kungfu names (total normalizing rate of 19.3%) in his translation, but foreignized 25 terms. Furthermore, he foreignized 159 out of the 170 character names, but only naturalized 11 of them (total normalizing rate of 6.4%).

Overall speaking, both Minford and Holmwood tend to naturalize or semi-naturalize these culturally loaded terms by using idiomatic English expressions (plus transliteration), which give both of their versions the highest degree of normalization. It is also noticed that Minford's overall normalizing rate is slightly higher than that of Holmwood's (i.e., 73.1% plus 52.4 vs. 48.2% plus 75.8%), ranking his translation the highest among the four in terms of the degree of

Table 3 Frequencies of translation strategies in the four translations and their total naturalizing rates

	Fox Volant of the Snowy Mountain		The Book and Sword		The Deer and the Cauldron		A Hero Born: Legends of the Condor Heroes	
	Kungfu	Character	Kungfu	Character	Kungfu	Character	Kungfu	Character
Naturalizing	4	35	4	11	19	96	27	47
Foreignizing	16	282	25	159	7	87	29	15
Omitting	0	0	2	0	0	0	0	0
Total	20	317	31	170	26	183	56	62
Total normalizing rates	20.00%	11.00%	19.30%	6.40%	73.10%	52.40%	48.20%	75.80%

normalization. Holmwood's ranks the second by following closely in total normalizing rates. By contrast, both Mok and Earnshaw tend to exoticize or omit when dealing with these culturally loaded terms. This means they tend to keep the Chinese martial arts elements of the ST. Hence, the degrees of normalization in the same regard are comparatively lower in their translations. Mok's total normalizing rates (20% plus 11%) are a bit higher than those of Earnshaw's (19.3% plus 6.4%), which indicates a higher degree normalization than the latter.

4 Discussion

From the above analysis of the four lexical aspects in the four translations, we can map out the normalization tendencies in these versions. Overall speaking, Minford's translation has the highest degree of normalization among the four translations, followed by Holmwood's translation that shows a moderate level of such tendency. Whereas Mok's translation can only be said to have a mild level of normalization features in her use of vocabulary, Earnshaw's version has the lowest level in this aspect. To discover the underlying reasons for these translators' decisions on different degrees of lexical normalization, the present research will discuss their motivation for undertaking the translation task and explore whether there is a relationship between the receptions of these translations and the different normalization tendencies.

4.1 Translators' Motivations and Strategies

This section explores possible connections between these translators' motivations and the translating strategies they have used in their TTs. It is found that the four translators have different motivations when translating Cha's martial arts fiction, which in turn leads to varied reception of these translations among the target readership.

4.1.1 Promoting Chinese Literature Overseas

Being a translation scholar, Olivia Mok clearly stated her major motivation of translating *Fox Volant of the Snowy Mountain* as promoting Chinese martial arts cultures overseas as well as introducing the writer Louis Cha—the master of Chinese martial arts fiction—to the Western literary academia (Mok 2001a, b). In the preface to her translation, she also stresses that a translator of Chinese martial arts fiction needs to keep not only the plots and narrative flow of the ST, but also all the martial arts elements (Mok et al. 1993: 24). Nevertheless, the task of promoting Chinese literature is never a straightforward issue. Today, the dominance of English literature in international book markets gives rise to a relatively marginalized position of literature in languages (e.g., Chinese literature) other than English (e.g. Luo 2011: 54; Hu 2018: 19). To solve this problem, Mok adopted a balanced strategy: on the one hand, she tried to promote the Chinese marital arts cultures by retaining the martial arts essence in Louis Cha's fiction; on the other hand, she fine-tuned her language to the taste of English-speaking readers by replacing certain well-known Chinese cultural images with their dynamic equivalent elements in the Western culture. For example, she translated "大俠"[3] as "the paragon of all chivalric deeds"; "貂蟬"[4] and "張飛,"[5] respectively, as "kingdom-quelling beauty" and "ferocious warrior." In this way, she replaced these novel characters who are well-known to Chinese readers with images of "chivalry," "warrior," or "beauty" which are more familiar to English readers. In addition, a scrutiny of her translation reveals that she has omitted some contents in the ST and rearranged its paragraphs to make some fictional details more accessible to native English-speaking readers. This rearrangement enhances the readability of her translation to average English-speaking readers as hoped by the translator (Mok 1993: i-ii). Consequently, such translating motivation and strategies have more or less given her English translation of *Fox Volant of the Snowy Mountain* a rather mild level of lexical normalization.

4.1.2 Winning Readership in English-Speaking Worlds

John Minford is a renowned Sinologist, who had successfully translated many Chinese classics into English. In the preface to his English translation, he wrote that he wanted to bring English readers the same pleasure of millions of Chinese readers who read the Louis Cha's martial arts fiction, even though it is a big challenge (Minford 1997: 9). It seems safe to say that this challenge partially comes from rendering the fiction that is fully loaded with Chinese martial arts cultures into English properly. Responding to our question for his motivation in doing the translation,[6] Minford noted that "*I was also personally motivated as a translator by the challenge of putting a full-length Martial Arts novel into an equivalent style of English.*"[7] This may partly explain why he spent great efforts in reconstructing the Kungfu fighting scenes so as to bring his English readers the same pleasure as ordinary Chinese readers would usually have when reading the original descriptions. In his rendition, verbs or verb phrases are frequently used (second only to Holmwood's translation in terms of the normalized frequencies as indicated in Table 2) to expand the original Kungfu fighting scenes in a detailed manner. For

[3] Chinese heroes who are physically and morally superior to others.

[4] A well-known ancient Chinese beauty based on Chinese folk legend.

[5] A well-known ancient Chinese warrior in *Romance of the Three Kingdoms.*

[6] Quoted from our personal interview email with John Minford on 23/10/2018.

[7] Quoted from our personal interview email with John Minford on 23/10/2018.

example, a common fighting description “拔刃夾擊” (to attack with swords) in the ST was described more vividly in the TT by using two verb phrases, i.e., “drew their swords” and “joined the fray.” Other similar examples include the translation of “不顧義氣” (ingratitude) as “escaping,” “上陣交鋒” (to join the fight) as “led his cavalry into battle,” and “摔将下来” as “threw their riders to the ground.” All these added verbal descriptions echo Minford’s preference for verb/verbal phrase as suggested in Table 2, lending his version an additional vividness for the fighting scenes.

4.1.3 Introducing the Chinese Martial Arts Culture to the West

As a literary agent and dedicated translator, Holmwood saw the potential of the Chinese martial arts culture among those devoted English-speaking fans of Fantasy and Chivalry fiction in the West (Peng 2018). She is keen on introducing this special type of classic Chinese literature to the West, as she sees the culture of Chinese martial arts resonate with the spirit of Western chivalry in a fundamental way. In an interview (Peng 2018) with the Chinese media “thepepar.cn,” Holmwood believed that the core of Chinese martial arts culture is not alien to English-speaking readers of western chivalry and fantasy fiction, since all of them valued heroism. This belief gives confidence to Holmwood to introduce the Chinese martial arts culture to the West, where readers’ passion with chivalric spirits and fantasy could intermingle with their curiosity about ancient Chinese heroism. Hence, to arouse this particular group of readers’ interest for the Chinese martial arts culture, it is necessary for Holmwood to make the language of translated martial arts fiction accessible to these potential English-speaking readers as much as possible. This motivation might to some extent explain why her English translation shows a moderate degree of lexical normalization. For example, she rendered the Chinese character “鵰” (literally, “an eagle”) in the fiction title as “Condor,” because this term would evoke a better and more impressive picture about this ferocious bird among English-speaking readers as they try to understand the important cultural significance for this unique symbol in the Chinese martial arts culture (Holmwood 2018: 238). Likewise, she translated the Kungfu name “九陰白骨爪” as “Nine Yin Skeleton Claw,” and “摧心掌” as “Heartbreaker Palm.” Compared with the usual Pinyin equivalents (e.g., “Cuixin Palm” and “Jiuyinbaigu Claw”) adopted by the previous translators, her more idiomatic and culturally accessible version should be able to reach her target readers better, thus laying a solid foundation for introducing the Chinese martial arts culture to the West.

4.1.4 Learning the Chinese Language and Cultures

Earnshaw is both a journalist and Sinologist who has a great passion for the Chinese language and culture. When working as a journalist for the *South China Morning Post*, he was already an avid fan of Chinese literature and culture, believing that reading and translating Chinese literary works were one of the best ways to learn the Chinese language. In an annual gathering of the Hong Kong Translation Society, he claimed that his chief motivation for translating *The Book and Sword* into English stemmed from his personal interest in learning Chinese (Earnshaw 2005). This motivation was reiterated in his response to our email interview, which he clearly indicated that "*the purpose of translating this novel is to improve my Chinese and learn more about Chinese culture.*"[8] In other words, his primary concern in the translation is whether his translation can represent the Chinese language and culture accurately and effectively. In addition, his philosophy of literary translation also affects his translating strategies: he believes that literal translation, as a translating strategy, shall prevail in translating a martial arts fiction like *The Book and Sword* into English (Earnshaw 2005). Accordingly, he deliberately retained most of the martial arts elements in his TT in a very faithful manner. For instance, ancient Chinese weapons used in Kungfu fighting scenes such as "軟鞭," "懷杖," and "鬼頭刀" are, respectively, rendered as "a whip," "a staff," and "Devil's Head Knife." Such graphic and direct translations of these images of ancient Chinese weapons present non-Chinese readers the special charm of ancient Chinese martial arts culture. His other method of keeping the original cultural elements transparent is to translate almost all Chinese idioms in a literal manner, retaining both the form and the content as much as possible. For instance, he translated the idiom "龍有頭, 人有主" literally as "just as a dragon has a head, men have masters"[9] to retain its original rhetorical and cultural features. To sum up, it is due to Earnshaw's motivation for retaining the original linguistic and cultural elements more closely in his rendition that gives his translation the lowest degree of lexical normalization (as reflected in the overall findings of Sect. 3) when compared with the other three translations.

4.2 Readers' Reception of the Translations in the English-Speaking World

Nowadays, a great majority of literary publications are available from online book vendors. These online vendors, while selling books to readers, also provide readers' feedback to their potential customers. In addition, online book forums, where readers can share their comments on a book, are also available for the purchasers of the publications to express their views. These feedback and comments are not only important indicators of a book's popularity, but also illustrations of the reader's reading preferences. To investigate the reception of these four translations, the online reviews and comments from the key book-selling websites, such as from

[8] Quoted from our personal interview email with Graham Earnshaw on 22/10/2018.

[9] A more semantic translation in this case could be "there must be a man of charge to take responsibilities."

Table 4 Receptions of the four English translations (up to 02/2019)

	Amazon	Goodreads	Novelupdates
Fox Volant of the Snowy Mountain	3.20 of 5	3.77 of 5	3.00 of 5
The Deer and the Cauldron	4.30 of 5	4.21 of 5	4.40 of 5
The Book and The Sword	4.70 of 5	3.82 of 5	3.20 of 5
A Hero Born: Legends of the Condor Heroes Vol.1	4.40 of 5	4.19 of 5	4.30 of 5

"Amazon.com," "Goodread.com," and "Novelupdates.com" were retrieved and analyzed to catch a glimpse of the general reading preferences of their readers. These websites rate the popularity of a publication based on a 5-star ranking system by their readers. While1-star means that a publication is the least popular among their readers, 5-star signifies the highest popularity among the readers. We believe that the popularity rating as reflected by this ranking system may illustrate the reception of the publication from readers to a certain extent, not least because these three websites have a strong influence and enjoy a good reputation among authors and readers from the English-speaking worlds. Ratings on the four translations are presented in Table 4.

Table 4 summarizes readers' rating of the four English translations from the three websites up to Feb. 2019. As indicated in Table 4, Minford's translation of *The Deer and the Cauldron* enjoys the best reception among the four translations, with Holmwood's *A Hero Born: Legends of the Condor Heroes* closely follows in its steps. While Mok's translation of *Fox Volant of the Snowy Mountain* has received the most negative comments, Earnshaw's translation of *The Book and Sword* has got a moderate reception.

For Minford's translation, readers from "Novelupdates.com" give it a 4.4 out of 5 stars (32 ratings, 2 reviews), readers from "Amazon.com" 4.3 out of 5 stars (17 reviews), and readers from "Goodread.com" 4.21 out of 5 stars (198 ratings, 17 reviews). Some comments on his translation include: "I hope more people see the beauty and importance of Cha's work and John Milford's brilliant translation enough to support new editions in physical and e-book formats" and "Thank you to the person who took the time to translate it into English." Such positive comments are clear evidence that the readers enjoy reading the translation as Minford has hoped.

In a similar fashion, instant popularity of Holmwood's rendition is clearly shown in readers' ratings: 4.40 out of 5 stars on "Amazon.com" (20 reviews), 4.19 out of 5 stars on "Goodread.com" (482 ratings, 79 reviews), and 4.30 out 5 stars on "Novelupdates.com" (15 ratings, no review). These readers give high praise to Holmwood's effort in introducing martial arts cultures to the West through her translation. Typical readers' comments are: "When I first heard of this book, I was very excited. But then, when I first picked it up in my hands, I became concerned," "*A Hero Born* reads like a Chinese version of the classic *Lord of the Rings,*" and "A great novel about the martial arts culture, and umpteen kinds of kungfu." As expected, most readers from the three sites are gripped by Holmwood's vibrant translation of the fairytale-like Chinese martial arts world depicted by Cha.

By contrast, Mok's translation receives only 3.0 out of 5 stars from readers at "Novelupdates.com" (6 ratings, no review), 3.2 out of 5 stars from readers at "Amazon.com" (6 reviews), and 3.77 out of 5 stars from "Goodread.com" (412 ratings, 24 reviews). Some readers from these sites commented that "How could anyone use a dictionary to translate martial arts terminology" and "I really don't understand the translation of 'chi' as 'pneuma'. I guess that more English speakers know what 'chi' is than 'pneuma.'" Apparently, these readers are not satisfied with Mok's English rendition which closely mirrors the original diction.

As for Earnshaw's translation, the situation is more complicated. Readers from different sites have divided comments on his work: readers from "Amazon.com" rate it 4.7 out of 5 stars (9 reviews), but readers from "Novelupdates.com" give it 3.2 out of 5 stars (11 ratings, 2 reviews) and readers from "Goodreads.com" give it 3.65 out of 5 stars (930 ratings, 189 reviews). One reader from "Amazon.com" commented that "Translating something so culturally specific is always difficult, but I feel that the translator did a great job"; but another reader from "Novelupdates.com" said that "After flipping through a few pages, I became frustrated as I couldn't recognize the names of the characters due to the 'Pinyin.'"

From the ratings of and reviews on the four English translations of Cha's martial arts fiction by these online readers, we see some potential connections between readers' reception and lexical normalization in the English translations of Chinese martial arts fiction. In Minford's translation and Holmwood's translation, relatively high degrees of lexical normalization win their translations favorable receptions from English-speaking readers. Such degrees of lexical normalization are partly related to the two translators' translating motivations and strategies (cf. 4.1.2 and 4.1.3). By contrast, a comparatively less favorable reception for Mok's translation could be associated with its mild degree of lexical normalization, which partly stems from Mok's translation motivation and strategies (cf. 4.1.1). As for Earnshaw's translation, the divided receptions of his rendition are not only possibly connected with the low degree of lexical normalization of his translation (which partly stems from his translating motivations and strategies as discussed in Sect. 4.1.4), but also the varied types and preferences of his readers.

On-line profiles of those registered readers from Goodreads.com and Novelupadate.com reveal that readers who are interested in the English translation of Chinese martial arts fiction are primarily fans of Historical Novel, Fantasy Novel, and Chivalry Novel. When examining these readers' profiles in depth, we notice that these fans' interests in Chinese martial arts fiction partly derive from their fascination for the fantasized Chinese history and culture as depicted by this unique Chinese literary genre. In addition, by combining "martial arts" with "chivalrous spirits," the martial arts fiction also manages to arrest the attention of Chivalry Novel readers. In other words, the similar fantasy and historical elements between martial arts genre and the three preferred types of novel (i.e., historical, fantasy, chivalry) as indicated by the fans attract them to this special type of Chinese fiction. But since differences among these four literary genres (i.e., martial arts, historical, fantasy, chivalry) do exist, the readers will appreciate and interpret the Chinese martial arts fiction through their idiosyncratic eyes, and consequently give martial arts fiction varied ratings. This might also help explain the divided receptions of Earnshaw's rendition.

5 Conclusion

The present research examined lexical normalization in the four complete English translations of Louis Cha's martial arts fiction through the self-built Louis Cha Corpus and BNC Baby (fiction). Our findings show that among the four translations, Minford's one is colored by the highest degree of lexical normalization and enjoys the best reception in overseas markets. Holmwood's translation, following that of Minford in the degree of lexical normalization, gains instant popularity soon after its publication. Meanwhile, Mok's rendition, which has a mild level of lexical normalization, has received the most negative feedback from foreign readers. Earnshaw's translation, on the other hand, has the lowest level of lexical normalization and its reception among the readers is the most divided and varied. Further investigation in this research shows that translator's motivation governs his/her choice of translating strategies, which in turn influences the degree of lexical normalization in translations. This degree of lexical normalization could be one of the many factors (e.g., content of the original, target readership, time of publication) that affects reader's receptions.

Finally, it is worth pointing out that the interplay among normalization, translator's motivation, and reader's reception explored by the present study may methodologically contribute to the existing corpus-based translation research on normalization by introducing a qualitative perspective in the corpus text analysis on top of the frequently used quantitative methods. Such a combination of macro and micro analyses offered by these two perspectives will help us to get a more in-depth understanding on the nature of normalization in translation.

Acknowledgements The authors would like to express sincere thanks to both Prof. John Minford and Mr. Graham Earnshaw for their reply to our survey questions.

References

Baker, M. 1993. Corpus linguistics and translation studies-implications and applications. *American Journal of Physiology* 274 (1): 321–327.

Daller, H., R.V. Hout, and J. Treffers-Daller. 2003. Lexical richness in the spontaneous speech of Bilinguals. *Applied Linguistics* 24 (2): 197–222.

Earnshaw, G., and L. Cha. 2004. *The book and the sword: A martial arts novel.* Oxford: Oxford University Press.

Earnshaw, G. 2005. *The Book and the sword*. http://www.earnshaw.com/other_writings/content.php?id=391. Accessed 24 Dec. 2018.

Holmwood, A. 2018. *A hero born: Legends of the condor Heroes*. London: MacLehose Press.

Hong, J. 2014. Wuxia xiaoshuo yijie yanjiu sanshinian [Studies on Translation of Martial Arts Fiction in the Past Three Decades]. *Foreign Languages and Their Teaching* 1: 73–79.

Hu, A.J. 2018. Gaige kaifang sishinian zhongguo wenxue zouchuqu de chengjiu yu fansi [Four Decades of Promoting Chinese Literature Overseas: Achievements and Reflections]. *Chinese Translators Journal* 6: 18–20.

Hu, K.B. 2011. *Yuliaoku fanyixue gailun [Introducing Corpus-based Translation Studies]*. Shanghai: Shanghai Jiaotong University Press.

Hu, X.Y. 2007. Jiyu yuliaoku de hanyu fanyi xiaoshuo ciyu tezheng yanjiu [A Corpus-based Study on the Translational Norms of Contemporary Chinese Translated Fiction]. *Foreign Language Teaching and Research* 39 (3): 214–220.

Ippolito, M. 2014. *Simplification, explicitation and normalization: Corpus-based research into English to Italian translations of children's classics*. Newcastle: Cambridge Scholars Publishing.

Jin, Y. 1994. *Xueshan feihu [Fox Volant of the Snowy Mountain]*. Beijing: SDX Joint Publishing Company.

Jin, Y. 2002. *Ludingji [The Deer and the Cauldron]*. Guangzhou: Guangzhou Publishing Company.

Jin, Y. 1999. *Shujian enchoulu [The Book and The Sword]*. Beijing: SDX Joint Publishing Company.

Jin, Y. 2008. *Shediao yingxiongzhuan (Vol.1) [A Hero Born: Legends of the Condor Heroes Vol.1]*. Taipei: Zhihu Publishing House.

Kenny, D. 2001. *Lexis and creativity in translation: A corpus-based approach*. Manchester: St. Jerome.

Laviosa, S. 1996. *The English Comparable Corpus (ECC): A resource and a methodology for the empirical study of translation*. University of Manchester.

Laviosa, S. 1998. The corpus-based approach: A new paradigm in translation studies: A new paradigm in translation studies. *Meta* 43 (4): 474.

Lu, J.J. 2014. Xueshanfeihu yingyi mingxihua celv jianlun wuxia xiaoshuo fanyi zhidao [Explicitation Strategy of the English Translation of *Fox Volant of the Snowy Mountain*—On Translation of Martial Arts Fiction]. *Journal of xi'an International Studies University* 22 (1): 126–129.

Luo, Y.Z. 2011. Jinyong xiaoshuo yingyi yanjiu jianlun zhongguo wenxue zouchuqu [Studies of Translating Jin Yong's Martial Arts Fiction and Promoting Chinese Literature Overseas]. *Chinese Translators Journal* 3: 51–55.

May, R. 1997. Sensible elocution: How translation works in and upon punctuation. *The Translator* 3 (1): 1–20.

McEnery, T., and R. Xiao. 2010. *Corpus-based contrastive studies of English and Chinese*. London: Routledge.

Mok, O., and J. Yong. 1993. *Fox Volant of the snowy mountain*. Hong Kong: The Chinese University Press.

Mok, O. 2001. Strategies of translating martial arts fiction. *Babel* 47 (1): 1–9.

Mok, O. 2001b. Translational migration of martial arts fiction east and west. *Target: International Journal of Translation Studies* 13 (1): 81–102.

Minford, J., and L. Cha. 1997. *The deer and the cauldron*. Oxford: Oxford University Press.

Moreno, G.A. 2016. Delivered translation versus published translation: Reflections on normalization in translation for publishing industry from a case study. *Meta* 61 (2): 396–420.

Peng, S.S. 2018. *Zhuanfang shediao yingwen yizhe* [*Interview with Anna Holmwood*]. https://www.thepaper.cn/newsDetail_forward_1840736. Accessed 31 Jan. 2019.

Scott, M. 1998. *Wordsmith tools manual*. Oxford: Oxford University Press.

Venuti, L. 1995. *The translator's invisibility: A history of translation*. London: Routledge.

Wang, Q. 2010. *Jiyu Yuliaoku de Ulysses hanyiben yizhe fengge yanjiu [A Corpus-based Study on Translator's Style in the Chinese Translation of Ulysses]*. Jinan: Shandong University.

Wang, X.L., and X.Q. Li. 2016. Jiyu yuliaoku de Shakespeare xiju hanyiben fanhua tezheng yanjiu [A corpus-based study of normalization in Chinese translations of Shakespeare's plays]. *Journal of Foreign Languages* 3: 103–112.

Williams, I.A. 2005. Thematic items referring to research and researchers in the discussion section of Spanish biomedical articles and English-Spanish translations. *Babel* 51 (2): 124–160.

Xia, Y. 2014. *Normalization in translation: Corpus-based Diachronic research into twentieth-century English-Chinese fictional translation*. Newcastle: Cambridge Scholars Publishing.

Xiao, C.W. 2012. Wuxia xiaoshuo yingyi huodongzhongde yishixingtai he shixue yi Olivia Mok xueshanfeihu yiben weili [Ideology and Poetics in Translating Martial Arts Fiction: Take Olivia Mok's *Fox Volant of the Snowy Mountain* as an Example]. *Journal of Yunnan Agricultural University (social Science)* 6 (4): 107–110.

Kan Wu Lecturer of translation studies, School of Foreign Languages, Zhejiang University of Finance and Economics, Dongfang College. His research interests include corpus-based translation studies and digital humanity.

Dechao Li Associate Professor of translation studies, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University. His research interests include corpus-based translation studies and translation theories.

Cognition and Translation Equivalents

Translating Principles of Translation: Cross-Cultural and Multi-Brain Perspectives

Chu-Ren Huang and Xiaowen Wang

Abstract Yan Fu's 譯事三難 has rarely been directly challenged but is frequently compared with Tytler's Principles of Translation. These two sets of principles match both in number and in the exact order of three parallel concepts. Given the canonical status of Tytler's principles since its publication in 1790, it is hard to imagine that Yan was not influenced by Tytler in forming his principles. Following this assumption, we explore possible accounts of why and how Tytler's three Principles of Translation could be "translated" into Yan's "信達雅" "*Xìn-Dá-Yǎ.*" We note that Tytler's was ranked in descending order of importance: the First General Law, which is the most important, requires a complete transcript of the original ideas (i.e., *Xìn*), whereas expressing the style and manner of the original writing (i.e., *Dá*) and achieving the ease of the original composition (i.e., *Yǎ*) are supplementary as the Second and Third General Laws. Yet Yan's principles tend to be understood in reverse order of importance. We explore this mismatch from cross-cultural and multi-brain perspectives based on a comparable corpora approach. Through comparing BNC with the Gigaword Chinese Corpus, it is revealed that cultural differences in the meanings of the ordinals *first* and *third* lead to the overlook of the foundational concept in English that "the first principle" is the most important, and the mistaking of the third law (*Yǎ*) as the highest one in China. This reconfiguration of cultural meanings underlines the nature of translation as a multi-brain activity situated in cultural contexts.

Keywords Translation principles · Ordinals · Comparable corpora · Multi-brain perspective · Cross-cultural translation

C.-R. Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University,
Hong Kong SAR, China
e-mail: churen.huang@polyu.edu.hk

X. Wang (✉)
School of English Education, Guangdong University of Foreign Studies, Guangzhou, China
e-mail: wangxiaowen_annie@gdufs.edu.cn

X. Wang
Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong SAR, China

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_8

1 Introduction

Tytler (1747–1813) precedes 嚴復 Yán Fù (1854–1921) as two important figures in the history of translation theories prior to the twentieth century. In English, Tytler's (1790 [1907]; 1813[1978])[1] *Essay on the Principles of Translation*, first released in 1790, is viewed as "the first comprehensive and systematic study of translation" (Munday 2016: 45). In Chinese, Yan's translation principles of "信達雅" "*Xìn-Dá-Yǎ,*" proposed in his Translator's Preface to 天演論 *Tiān-yǎn-lùn* (a translation of Thomas H. Huxley's *Evolution and Ethics*) in 1898,[2] have been regarded as a translation theory "unequaled by any other theoretical work" produced in the past century in China (Chan 2004: 65). Yet, it has long been suspected that Yan's theory on translation is influenced by Tytler, as the proposed principles are said to be highly identical (Fan 2008; Shen 沈蘇儒 1998; Wang 王宏印 2003; Xu and Xu 徐守平 and 徐守勤 1994; Zhao and Shi 趙巍 and 石春讓 2005). It is important to note that Tytler (1790 [1907]; 1813[1978]) clearly ranked his first, second, and third principles in descending order of importance, but Yan (1898) did not offer a clear ranking for the importance of his three principles, leading to numerous debates on "*Xìn-Dá-Yǎ"* in China ever since. Previous scholars (e.g., Liang梁啟超 1922; Luo 羅新璋 1983; Shen 沈蘇儒 1998; Wang 王宏印 2003; Wang 王宏志 1999) mostly argued about how *Xìn*, *Dá, and Yǎ* should be interpreted based on Yan's original statements; however, no one has systematically explored the deep-rooted reasons behind the long-standing disputes over the hierarchy of "*Xìn-Dá-Yǎ"* for more than a century. In this study, we aim to fill in this gap by investigating why and how Tytler's translation principles could possibly be "translated" into Yan's "*Xìn-Dá-Yǎ,*" and what mis-transformation regarding the internal relations of principles could have occurred in such a translation into the Chinese context. After a careful comparison of the translation principles proposed by Tytler and Yan Fu, we take a comparable corpora approach to investigate mechanisms behind their different accounts of the importance of *Xìn*, *Dá,* and *Yǎ* from cross-cultural and multi-brain perspectives.

2 Literature Review

2.1 Tytler's Three General Laws of Translation

In his *Essay on the Principles of Translation*, Tytler (1790[1907]; 1813[1978]: 16) raised three general laws, or rules of translation:

I. THAT the Translation should give a complete transcript of the ideas of the original work;
II. THAT the style and manner of writing should be of the same character with that of the original;
III. THAT the Translation should have all the ease of original composition.

The first law requires "a faithful transfusion of the sense and meaning of an author" (Tytler 1813[1978]: 109). This law is only about content, while his **second law** focuses on form (Rener 1989), calling for "an assimilation of the style and manner of writing in the translation to that of the original" (Tytler 1813[1978]: 109). Regarding the second law, Tytler (1813[1978]: 110) indicated that there are classes of styles for the author: "the grave, the elevated, the easy, the lively, the florid and ornamented, or the simple and unaffected," and the translator should not only deliver the right sense, but should also express in an accurate class of style that the original writer belongs to. According to Rener's (1989: 193) interpretation, the word "style" in this law is based on the subject matter, referring to "the style of the writer as the member of a group," whereas the "manner of speaking" means the writer's "personal style." If the translator does not obey this law, the author's idea would be expressed in "a distorting medium or a garb that is unsuitable to his character" (Tytler 1813[1978]: 110).

The third law relates to "the attainment of ease of style" (Tytler 1813[1978]: 213). This is the most abstract level among the three laws, and Tytler (1813[1978]: 211) had to use some metaphors to describe it. A good translator is like a walker who "exhibit an air of *grace* and freedom while walking" (Tytler 1813 [1978]: 211, emphasis added). He is also like a painter imitating a copy of a picture: "if the original is easy and *graceful*, the copy will have the same qualities" (Tytler 1813 [1978]: 211, emphasis added). This law requires a translator to "reflect the ease and spirit of the original," that is, he "must adopt the very soul of his author, which must speak through his own organs" (Tytler 1813[1978]: 212). However, Tytler did not clearly reveal what this "spirit" or "soul" exactly refers to (Munday 2016: 46).

To us, since the expression of "ease" is repeatedly paralleled with "grace" or "graceful" in Tytler's metaphorical descriptions of this law, the "spirit" or "soul" is most likely related to the esthetic and artistic value of the work. The fulfillment of this law would naturally involve a fluent expression of the sentences, but the esthetic pursuit actually goes much beyond fluency.

Tytler's three laws are ranked in order of importance, with the first law being the most important and the third law being the least important. He also noted that the three laws might conflict with each other, and if that happens, the first law should be the last to sacrifice. He (Tytler 1813 [1978]]: 224, emphasis added) elucidated such an order very explicitly:

> IF the order in which I have classed the three general laws of translation be their just and natural arrangement, which I think will hardly be denied, it will follow, that in all cases where a sacrifice is necessary to be made of one of those laws to another, a due regard ought to be paid to their rank and comparative importance. The different genius of the languages of the original and translation, will sometimes make it necessary to depart from, the *manner* of the original, in order to convey a faithful picture of the *sense*; but it would be highly preposterous to depart, in any case, from the *sense*, for the sake of imitating the *manner*. Equally improper would it be, to sacrifice either the *sense* or *manner* of the original, (if these can be preserved consistently with purity of expression), to a fancied *ease* or superior *gracefulness* of composition.

2.2 Yan Fu's Translation Principles

Yan Fu studied at the Royal Naval College in Greenwich from 1877–79, nearly 90 years after the publication of Tytler (1790) and at the time when Tytler's three laws had been widely cited and followed. In his 譯例言 *Yì-lì-yán* "Translator's preface" to 天演論 *Tiān-yǎn-lùn*, Yan (嚴復 1898) proposed 譯事三難 *Yì-shì-sān-nán* "Three Challenges to Translation," namely, 信 *Xìn* "faithfulness; fidelity; trueness; trustworthiness; loyalty," 達 *Dá* "expressiveness; fluency; readability; intelligibility; comprehensibility, smoothness," and 雅 *Yǎ* "elegance; refinement; gracefulness." Since then, the three aspects have been upheld as criteria for judging the quality of translation in China, often commented as the central concepts for Chinese translation theory and practice in the past century (Munday 2016).

In the preface, Yan (嚴復 1898) firstly described the relations between *Xìn* and *Dá*:

> "譯事三難: 信、達、雅。求其信, 已大難矣! 顧信矣不達, 雖譯猶不譯也, 則達尚焉。"
>
> *Yì shì sān nán: Xìn, dá, yǎ. Qiú qí xìn, yǐ dà nán yǐ! Gù xìn yǐ bù dá, suī yì yóu bù yì yě, zé dá shàng yān.*
>
> "Translation involves three requirements difficult to fulfill: faithfulness (*Xìn*), expressive (*Dá*), and elegance (*Yǎ*). It is plenty difficult to seek to reach *Xìn* (faithful). Yet if the translation is faithful but not as expressive as the original/failed to reach *Dá*, then the translation is futile and one may as well not translate. Hence *Dá* tops *Xìn*." (Our translation)

Yan (嚴復 1898) proposed the concept of *Dá* out of concerns for the differences between English and Chinese, such as disparate patterns in syntax, cohesion, and coherence. It is, therefore, necessary for a translator to understand thoroughly and digest the whole text, and then rewrite the original work in the best manner possible (Chan 2004), which he refers to as *Dá* "expressiveness." He (Yan 嚴復 1898) argued:

> 至原文詞理本深, 難於共喻, 則當前後引襯, 以顯其意。凡此經營, 皆以為達, 為達即所以為信也。
>
> *Zhì yuánwén cí lǐ běn shēn, nányú gòng yù, zé dāng qián hòu yǐn chèn, yǐ xiǎn qí yì. Fán cǐ jīngyíng, jiē yǐ wèi dá, wéi dá jí suǒyǐ wéi xìn yě.*
>
> "Since the original is profound in thought and involved in style, which are difficult to convey together, the translator should correlate what precedes and what follows to bring out the theme. All his effort is to achieve expressiveness, for only when a piece of translation is expressive can it be regarded as faithful". (Adapted from Hsu's translation of 譯例言 collected in Chan 2004: 69)

Further, Yan (嚴復 1898) added *Yǎ* in addition to *Xìn* and *Dá*, and elaborated on them based on ideas from ancient Chinese works:

> 《易》曰: 修辭立誠。子曰: 辭達而已。又曰: 言之無文, 行之不遠。三者乃文章正軌, 亦即為譯事楷模。故信、達而外, 求其爾雅。此不僅期以行遠已耳, 實則精理微言, 用漢以前字法、句法, 則為達易; 用近世利俗文字, 則求達難。
>
> *Yì yuē: Xiūcí lì chéng. Zǐ yuē: Cí dá éryǐ. Yòu yuē: Yán zhī wú wén, xíng zhī bù yuǎn. Sān zhě nǎi wénzhāng zhèngguǐ, yì jí wéi yì shì kǎimó. Gù xìn, dá ér wài, qiú qí ěr yǎ. Cǐ bùjǐn*

> *qī yǐ xíng yuǎn yǐ ěr, shízé jīng lǐ wēi yán, yòng hàn yǐqián zìfǎ, jùfǎ, zé wèi dá yì; yòng jìnshì lì sú wénzì, zé qiú dá nán.*
>
> "The *Book of Changes* says: 'Trustworthiness is the basis of writing.' Confucius commented on this and said: 'The purpose of words is to communicate.' He also commented that, 'language with no elaboration won't go far.' These three dicta set the right course for writing and should also be the standards for translation. Thus, in addition to *Xìn* and *Dá*, we must seek to reach its status of elegance. This is not only to make sure that the translation can go a far way (i.e., reaching *Dá*), but also to really master the rationale and philosophy and understand the unspoken. For instance, using the (classical) grammar of pre-Han, it is easy to achieve the goal of being fully expressive. Using the vulgar modern vernacular, then it is difficult to expect reaching expressiveness." (Our translation)

Since Yan did not illustrate the three concepts in detail and did not explicitly rank the three in terms of importance, his statement has triggered endless debates in translation studies in China for over a century, with scholars holding various interpretations on the meaning of and the relations among the three principles.

2.3 *The Possible Influence of Tytler's Theory on Yan Fu's Translation Principles*

Yan did not explicitly refer to Tytler's Three Laws of Translation at all in his work, yet scholars (Fan 2008; Shen 沈蘇儒 1998; Wang 王宏印 2003; Xu and Xu 徐守平 and 徐守勤 1994; Zhao and Shi 趙巍 and 石春讓 2005) have assumed or inferred that Yan's theory of translation principles was influenced by Tytler's three laws. For example, Fan (2008: 69) noted the near identity of Yan's criteria and Tytler's laws and speculated that: "if two men living in two different countries a century apart, should think alike, that surely means something." Indeed, since Tytler's *Essay on Translation Principles* was published more than 100 years earlier than Yan's 譯事三難, and Yan Fu studied in Britain during 1876–1878 (Shen 沈蘇儒 1998), it is hard to imagine that Yan Fu was not influenced by Tytler in forming his principles. There has not been any confirmed evidence for this speculation (Chen 陳福康 2011; Shen 沈蘇儒 1998), but Wu (伍蠡甫 1980) recalled an anecdote that Yan's student heard from Yan himself that the concepts of *Xìn, Dá,* and *Yǎ* originated from the West, rather than being Yan's own (Wang 王宏印 2003). The fact that Yan attributed his understanding of the translation principles to traditional Confucian scholarship, such as *the Book of Changes* and Confucius' *Annotation on the Book of Changes* (易傳), could be his strategy to acculturate these concepts and to imbue them with an aura of authority for the intended audience, just like his now outdated and odd non sequitur touting of Pre-Han grammar.

Zheng (鄭振鐸 1921) provided the first full translation of Tytler's three laws of translation into Chinese in his article "How to translate literary work," according to Chen (陳福康, 2011). The word "*ease"* in Tytler's third law was translated as 流利 *liúlì* "fluency" by Zheng. Later on, other Chinese scholars have come up with similar versions of translation (e.g., 通顺 *tōngshùn* “readable” by Guo (郭建中 2013)); however, scholars failed to identify the artistic effect of work that Tytler emphasized when describing his third law metaphorically. Mistakenly perceiving Tytler’s “*ease*” as fluency, translation researchers in China have been arguing that Yan Fu’s second principle of *Dá* matches Tytler’s third law, and his third principle of *Yǎ* somehow matches Tytler’s second law because both deal with style (Shen 沈蘇儒 1998; Wang 王宏印 2003; Xu and Xu 徐守平 and 徐守勤 1994; Zhao and Shi 趙巍 and 石春讓 2005). Only Li (李田心 2014), in his “Further Amendment of the Translation of the Third Principle of Tytler’s Principles of Translation,” rightly pointed out that Tytler’s “*ease*” does not mean “readability,” and that *Dá* “expressiveness” is implicitly referenced rather than explicitly stated in Tytler’s third law. Li (李田心 2014) explained “*ease*” as referring to being able to make the readers comfortable and relaxed. However, he, just like the other scholars cited above, failed to identify the extensive use of “grace” and “graceful” to refer to the law (probably partly due to the translation that they read), and failed to directly link Tytler’s third law to *Yǎ*. Nevertheless, Li (李田心 2014) did conclude that this “*ease*” is the final level of attainment in Tytler’s translation laws, in parallel to Yan Fu’s *Yǎ*. Considering Tytler’s repeated emphasis on “gracefulness” in parallel with the “*ease*” of translation work in the third law, and that such a requirement is clearly interpreted by Tytler (1813 [1978]: 213) as more difficult than the “assimilation of manner and style of writing” in the second law, we conclude that Tytler’s third law is actually closer to Yan’s last principle of *Yǎ* than that of *Dá*. On the contrary, *Dá* is, in fact, just a requirement of manner (or form), not related to the artistic value of the work, and definitely not the most difficult level to achieve in literary translation that Yan focuses on. Therefore, we believe it more likely that Yan’s principle of *Yǎ* “elegance” reflects Tytler’s third law, while *Dá* “expressiveness” is a narrower concept loosely related to Tytler’s second law, with Tytler’s definition covering more broadly the transcript of both original manner and style. If “*Xìn-Dá-Yǎ*” is really a translation from Tytler’s three laws of translation as we reckoned, it was translated in a wholesale, that is, in the same number and order.

It should be noted, however, that the concept of *Dá* actually deviates slightly from Tytler’s principles. Yan Fu’s *Dá* means to be expressive in the best manner adapting to the different syntactic and discoursal patterns between the target and source languages. This principle does not require a translator to strictly stick to the style of the original (Li 李田心 2014); rather, Yan even changed the style of the original work in his translation of *Evolution and Ethics* for the purpose of achieving *Dá* (Huang黃忠廉 2016). While Tytler’s second law addresses manner, it focuses more on a resemblance of the original style of the work and the author. Thus, if reading is a conversation between an author and readers, Tytler wants a translation to follow the manner of the original speaker while Yan Fu wants to adopt the manner of readers. This is, in fact, quite understandable given Yan Fu’s daunting task to win over the scholarly readers in China who are only familiar with the classical Chinese writings.

2.4 *Controversial Understandings of the Order of Importance for* Xìn, Dá, Yǎ

As we argued above, Yan's version might be a direct translation from Tytler's three general laws of translation, with the same number and order of principles. Unlike Tytler, who explicitly ranked the three laws in descending order of importance, Yan has been criticized for his vague account of the order of importance for the concepts of *Xìn*, *Dá*, and *Yǎ*, which leads to the numerous debates among translation researchers in China. Because of this, some scholars even hold that Yan's principles are logically confusing (Chang 常謝楓 1981; Huang黄雨石 1988), susceptible to miscomprehension (cf. Fan 2008: 64), and detrimental to translation practice (Chang 常謝楓 1981).

Xìn and Yǎ, which is the most important?

The understanding of the order of importance for "*Xìn-Dá-Yǎ*" is quite divided in the literature. Many including translators and/or philosophers Ai 艾思奇 (1937), Li (李培恩 1935), Liang (梁啟超 1922), Lv (吕博 1998), Fan (范存忠 1978), Zhang 張威廉 (1984, 1988), Zheng (鄭意長 2002), etc., perceive *Xìn* as the most important criteria, *Dá* as the second most important, and *Yǎ* the least important among Yan's three principles (Fan 2008). Chao (趙元任 1969a, b)[3] argued that *Dá* and *Yǎ* are not always valid, as it might be inappropriate to translate a text into an expressive and/or elegant piece of work against its original style. Rather, *Xìn*, with many dimensions that are difficult to attain, should be the fundamental prerequisite for translation. In a famous article that summarized the development of people's understanding of "*Xìn-Dá-Yǎ*" in China, Luo (羅新璋 1983) stated that (1) *Xìn* had gradually been accepted as the fundamental and essential principle, with *Dá* and *Yǎ* viewed as its subordinates, and (2) researchers (e.g., Qian (錢鍾書 1986)) later tended to realize that *Xìn* can incorporate *Dá* and *Yǎ*. An even more extreme opinion is that *Xìn* should be the only criterion for judging good translation since it already incorporates the other two elements (e.g., Yang and Liu 楊自儉 and 劉學雲 2003), although Wang 王宏印 (2003) pointed out that this is an incorrect understanding of Yan Fu's original proposal.

On the contrary, others claimed that *Yǎ* is the most important principle among the three. Munday (2016: 46) pointed out that although "Yan Fu himself generally placed *Xìn* above *Dá* [(Chan 2004: 4–5)]," "he did not always abide by the hierarchy, often privileging *Yǎ*." Views that take *Yǎ* as the most important principle of translation are generally concerned with the translation of literary works. In view of the controversial understandings of Yan Fu's principles, Chen (陳廷祐 1980) expressed that Yan actually regarded *Yǎ* with higher importance than *Xìn* and *Dá*. Similarly, Ye (葉君健 1997) regarded *Xìn* and *Dá* as the lower requirements for translation but took *Yǎ* as the most important, for only when *Yǎ* is fulfilled can a literary work be represented with unique characteristic (or spirit) in the target language. Rethinking translations of Tytler's laws of translation and its influence on Yan Fu, Li (李田心 2014) interpreted that Yan Fu's principle of *Yǎ* is the final requirement and target in literary translation, and to be *Xìn* and *Dá* is to serve the purpose of *Yǎ*. However, there are others who strongly criticized this position. Shen (沈蘇儒 1982; cf. 沈蘇儒 1998: 268) argued that the position that *Yǎ* is the most important criterion can neither be found nor entailed from Yan Fu's original statement. Chang's (常謝楓 1981) observation that pursuing *Yǎ* as the most important principle in literary translation regardless of the author's original style has become a popular trend in the society is probably the most apt summary. That is, regardless of the ongoing academic debate that seems to slightly favor *Xìn,* the general public perceives *Yǎ* as the ultimate criterion. This is attested by the frequent adoption of "*Xìn-Dá-Yǎ"* in other fields and typically with *Yǎ* as the utmost goal, as the Chinese segmentation standard proposed in 1996 in Taiwan (Huang et al. 2017).

Is Yǎ a fixed or changeable concept?

Other more radical positions include one that denies the value of *Yǎ*. Yan's advocacy to fulfill *Yǎ* by using the pre-Han language style was said to hinder the development of the new culture in the society, especially during the New Culture Movement in China. For example, Qu (瞿秋白 1931; cf. Luo and Chen 羅新璋 and 陳應年 2009), in his letter to 魯迅 *Lŭ Xùn* in 1931, raised the concern that *Xìn* and *Dá* can be overridden by *Yǎ* when a translator pursues *Yǎ* with classical Chinese against the dominance of vernacular literature in the 1930s, but his harsh critic is not supported by modern translation researchers (e.g., Wang 王宏印 2003). Scholars (Wang 王秉欽 2017; Wang 王宏印 2003; Wang 王宏志 1999; Wang 王克非 1992; Wu 吳存民 1997) generally regard Yan's performance as an effort to cater to the target readers (intellectuals who were used to reading pre-Han language) at his time and agree that the way to realize *Yǎ* is changeable with time. To be *Yǎ* "elegance" means to achieve the esthetical or artistical value of the work, in the standardized language commonly adopted across the country at the time of translation (Wang 王秉欽 2017).

Other views on the status of Xìn, Dá, and Yǎ.

Other views[4] include considering *Dá* as the most important principle (e.g., Gu 2010; Wang 王宏印 2003), that the three principles are equal (Yang 楊麗華 2011), and that they are incomparable in terms of importance (Fu 傅國強, 1990). For example, Wang 王宏印 (2003) argued that Yan put relatively less emphasis on *Xìn* and *Yǎ* in his theoretical account of translating but mainly pursued "達旨" "*dá-zhǐ"* "expressing primary intention" in his translation practices. The latter can be understood as a way to achieve expressiveness with adaptation techniques in translation (Huang and Chen 黃忠廉 and 陳元飛 2016). Yet, assigning *Dá* the prime importance might be a misinterpretation of Yan's original meaning, as there is no convention in either English or Chinese to rank the second as the most important in a listing of three ordered items; otherwise, the writing would sound quite illogical. In all, interpretations on the importance of *Xìn*, *Dá,* and *Yǎ* are tremendously controversial, but the reasons behind such disputes have not been deeply investigated. Bearing this important issue in mind, we will explore mechanisms for the misunderstandings possibly arising from Yan Fu’s translation of Tytler’s Three General Laws of Translation into the Chinese context.

3 Methodology

Based on the literature review in Sect. 2, we have identified that Yan Fu’s “*Xìn-Dá-Yǎ*” was most likely a translation from Tytler’s Three General Laws of Translation, with *Xìn* mapping the first law, *Dá* the second, and *Yǎ* the third. While controversial interpretations of the order of importance for *Xìn*, *Dá*, *Yǎ* have arisen in translation studies, the predominant view among the general public is of an ascending order of importance. Given the strong bias favoring *Yǎ* from the public, and the lack of explicit assignment of ranking order in the original texts by Yan Fu, it is likely that the ranking order simply received a different default order of importance in Chinese and English. That is, Tytler’s three laws of descending importance, when translated into Chinese, were given the order of ascending importance in a cultural context.

To attest to the above hypothesis, we take a comparable corpus approach to extract the culturally grounded interpretation of the one-two-three order in Chinese and English. We focus in particular on the culturally loaded meanings of the ordinals *first* and *third* in English and 一 *yī* “one; first” and 三 *sān* “three; third” in Chinese. Comparable corpora are sets of text collections in different languages or language varieties, following the same type of criteria. They differ from parallel corpora in that the corpora to be compared are normally independent, whereas the ones in the parallel corpora are typically translated texts and their source texts or translated texts that share the same source texts (Kenning 2010: 487). Rather than having to rely on translated texts, comparable corpora are advantageous in their ability to keep the authenticity of naturally occurring languages, so they are especially useful for unraveling unique linguistic conventions and features in particular languages and have great potential to inform translation studies. For comparison within our study, we have selected the British National Corpus (BNC) (2007) (for English) and the Gigaword Chinese Corpus (Gigaword) (Huang, 2009) (for Chinese), both are large-scale corpora of contemporary language around the later part of the twentieth century. The former contains 100 million words of texts from a balanced source of genres such as the newspapers, fiction, and academic genres, while the latter comprises 831,748 words of newswire data collected from three newspapers in Chinese. We adopt the Sketch Engine (Kilgarriff et al. 2014) as the online query platform to analyze BNC, and the Chinese Word Sketch (CWS) as the platform to analyze GigaWord, the latter being a special version of the Sketch Engine for the Gigaword corpus (Huang et al. 2005). Both platforms provide similar functions to explore the corpora, including Word Sketch, Word Sketch Difference, Thesaurus, Concordance, etc., which makes it very convenient to compare the usages of target equivalent words in English and Chinese in the two corpora.

4 Findings

4.1 The Structural and Cultural Meaning of First in English

Comparing *first* and *third* in BNC through Sketch Engine, we found that *first* is used far more frequently: the relative frequency of *first* in BNC is 1,076.66 per million, which is more than five times the relative frequency of *third* (189.05 per million). Moreover, by applying the thesaurus function, we compared the top 20 synonyms for *first* and *third*, respectively (see Fig. 1). In Sketch Engine, the thesaurus lists are automatically generated based on the percentage of shared collocates in context. While *second* is very close to both *first* and *third* semantically, the word *important* is only included in the top 20 thesaurus list of *first*, not in that of *third*.

To explore deeper if the sense of importance is typically represented by *first* instead of *third* in English, we also applied the "Sketch Difference" function to compare the word *important* with *first* (Fig. 2) and with *third* (Fig. 3), respectively, in terms of the detailed collocation distributions grouped in categories of various grammatical relations in BNC. For the first three categories regarding subjects, the modified nouns and verbs, and the preceding verbs in the results shown in Fig. 2 and Fig. 3, *important* shares many common collocates with *first* (e.g., the subjects *point*, *question*, and *thing*, the modified nouns *thing* and *step*, and the verbs *consider*, *be*, *have*, *say,* and *do*), but shared no common collocate with *third* except *person.* For the collocates in "and/or" parallel relation with the search words, it is

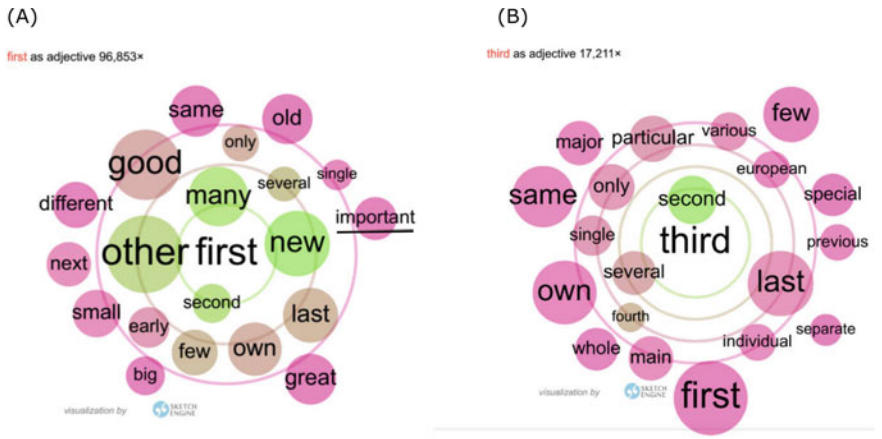
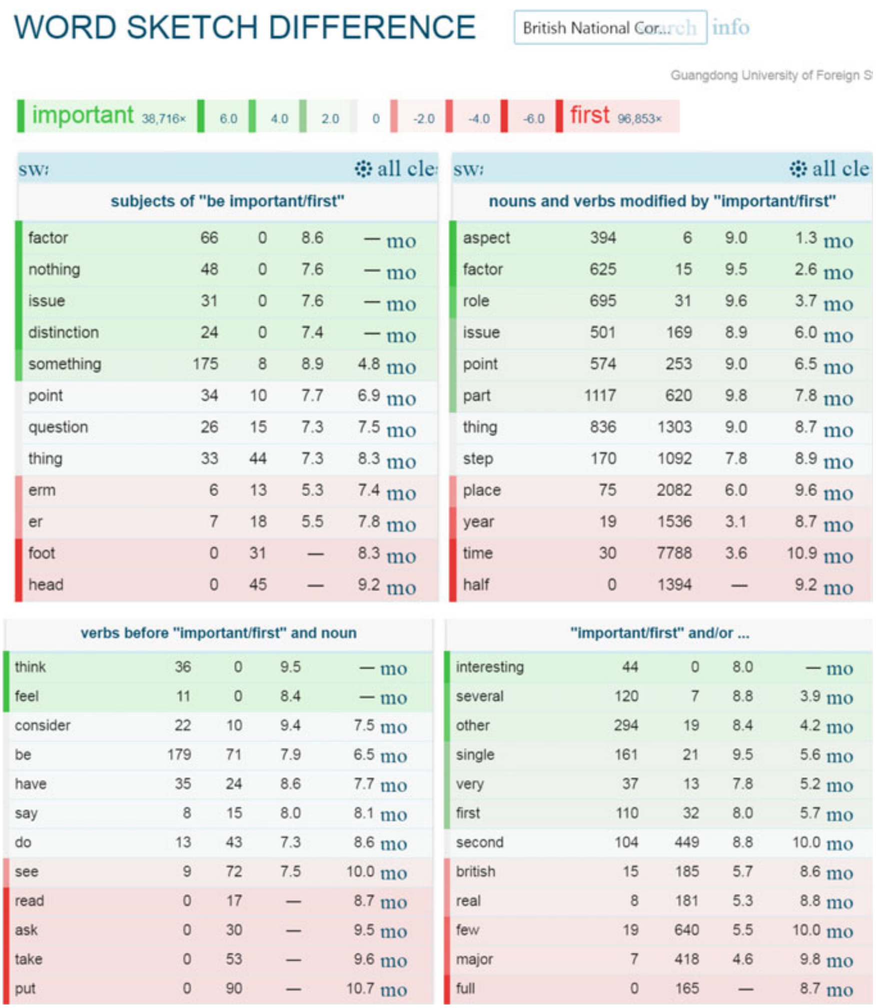


Fig. 1 Comparison of thesaurus results of *first* and *third* in BNC through Sketch Engine. *Note* (A) and (B) show the thesaurus results of *first* and *third* in BNC, respectively. In each figure, the key word *first* or *third* is located at the centre, with its top 20 synonyms or similar words scattering around it. The distance between the center word and the other words indicates their semantic closeness/distance. The word *important* is underlined in (A) for emphasis by the authors



WORD SKETCH DIFFERENCE British National Cor...rch info

Guangdong University of Foreign S

important 38,716× 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 first 96,853×

sw: all cle

subjects of "be important/first"

factor	66	0	8.6	—	mo
nothing	48	0	7.6	—	mo
issue	31	0	7.6	—	mo
distinction	24	0	7.4	—	mo
something	175	8	8.9	4.8	mo
point	34	10	7.7	6.9	mo
question	26	15	7.3	7.5	mo
thing	33	44	7.3	8.3	mo
erm	6	13	5.3	7.4	mo
er	7	18	5.5	7.8	mo
foot	0	31	—	8.3	mo
head	0	45	—	9.2	mo

sw: all cle

nouns and verbs modified by "important/first"

aspect	394	6	9.0	1.3	mo
factor	625	15	9.5	2.6	mo
role	695	31	9.6	3.7	mo
issue	501	169	8.9	6.0	mo
point	574	253	9.0	6.5	mo
part	1117	620	9.8	7.8	mo
thing	836	1303	9.0	8.7	mo
step	170	1092	7.8	8.9	mo
place	75	2082	6.0	9.6	mo
year	19	1536	3.1	8.7	mo
time	30	7788	3.6	10.9	mo
half	0	1394	—	9.2	mo

verbs before "important/first" and noun

think	36	0	9.5	—	mo
feel	11	0	8.4	—	mo
consider	22	10	9.4	7.5	mo
be	179	71	7.9	6.5	mo
have	35	24	8.6	7.7	mo
say	8	15	8.0	8.1	mo
do	13	43	7.3	8.6	mo
see	9	72	7.5	10.0	mo
read	0	17	—	8.7	mo
ask	0	30	—	9.5	mo
take	0	53	—	9.6	mo
put	0	90	—	10.7	mo

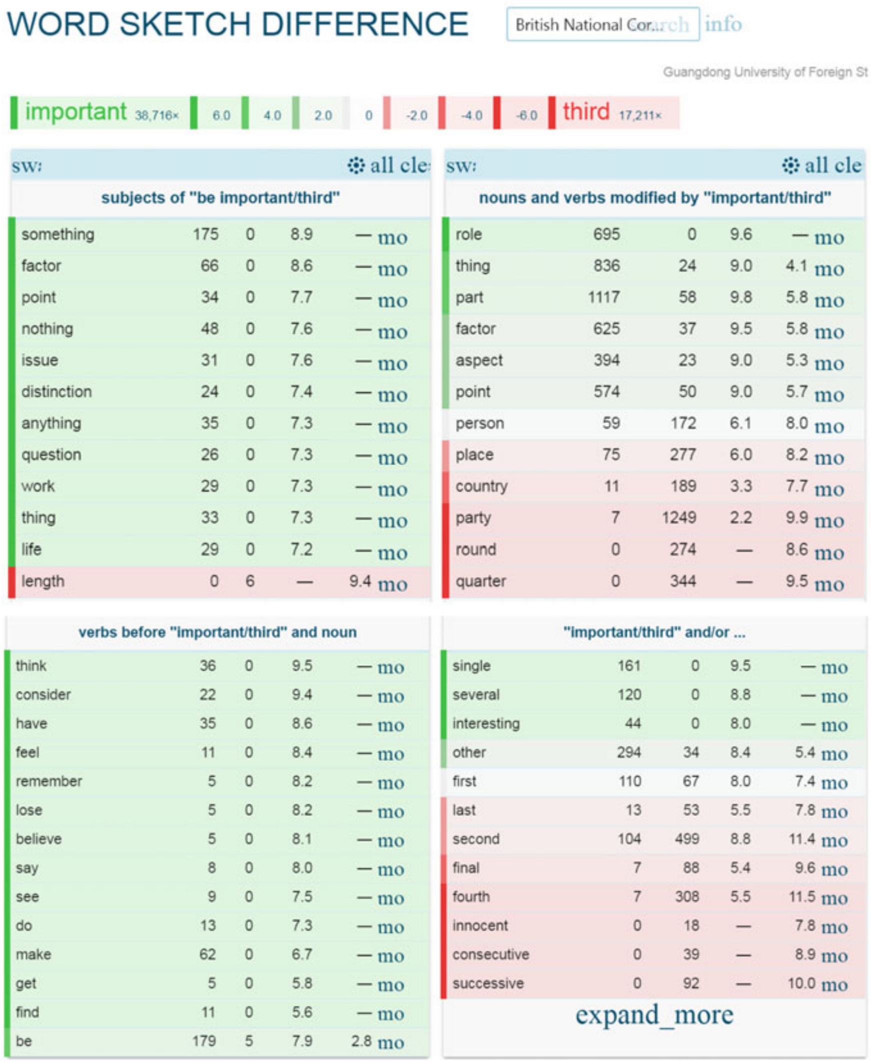
"important/first" and/or ...

interesting	44	0	8.0	—	mo
several	120	7	8.8	3.9	mo
other	294	19	8.4	4.2	mo
single	161	21	9.5	5.6	mo
very	37	13	7.8	5.2	mo
first	110	32	8.0	5.7	mo
second	104	449	8.8	10.0	mo
british	15	185	5.7	8.6	mo
real	8	181	5.3	8.8	mo
few	19	640	5.5	10.0	mo
major	7	418	4.6	9.8	mo
full	0	165	—	8.7	mo

Fig. 2 The sketch difference results for *important* and *first* in BNC

also apparent that *important* is often in an "and/or" relation with *first* (Fig. 2), but not with *last*, *final* (Fig. 3), while *third* can be in parallel with *last* and *final* (Fig. 3). Therefore, we can tell that in English, the *first* is most conventionally associated with importance, while the *last* or *final* item in a list is not.

We also ran a concordance for "*first…most important*" and "*third…most important*" in a sequence of 4–8 words. That is, we set 1–5 tokens between *first/third* and *most important*, thereby excluding expressions of *first most important* and *third most important*. It generated 49 hits (0.44 per million) for the collocation pattern of *first* with *most important* (Fig. 4), but only seven hits (0.06 per million) for *third* with *most important* (Fig. 5).



WORD SKETCH DIFFERENCE British National Cor...rch info

Guangdong University of Foreign St

important 38,716× 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 third 17,211×

sw: all cle

subjects of "be important/third"

something	175	0	8.9	— mo
factor	66	0	8.6	— mo
point	34	0	7.7	— mo
nothing	48	0	7.6	— mo
issue	31	0	7.6	— mo
distinction	24	0	7.4	— mo
anything	35	0	7.3	— mo
question	26	0	7.3	— mo
work	29	0	7.3	— mo
thing	33	0	7.3	— mo
life	29	0	7.2	— mo
length	0	6	—	9.4 mo

sw: all cle

nouns and verbs modified by "important/third"

role	695	0	9.6	— mo
thing	836	24	9.0	4.1 mo
part	1117	58	9.8	5.8 mo
factor	625	37	9.5	5.8 mo
aspect	394	23	9.0	5.3 mo
point	574	50	9.0	5.7 mo
person	59	172	6.1	8.0 mo
place	75	277	6.0	8.2 mo
country	11	189	3.3	7.7 mo
party	7	1249	2.2	9.9 mo
round	0	274	—	8.6 mo
quarter	0	344	—	9.5 mo

verbs before "important/third" and noun

think	36	0	9.5	— mo
consider	22	0	9.4	— mo
have	35	0	8.6	— mo
feel	11	0	8.4	— mo
remember	5	0	8.2	— mo
lose	5	0	8.2	— mo
believe	5	0	8.1	— mo
say	8	0	8.0	— mo
see	9	0	7.5	— mo
do	13	0	7.3	— mo
make	62	0	6.7	— mo
get	5	0	5.8	— mo
find	11	0	5.6	— mo
be	179	5	7.9	2.8 mo

"important/third" and/or ...

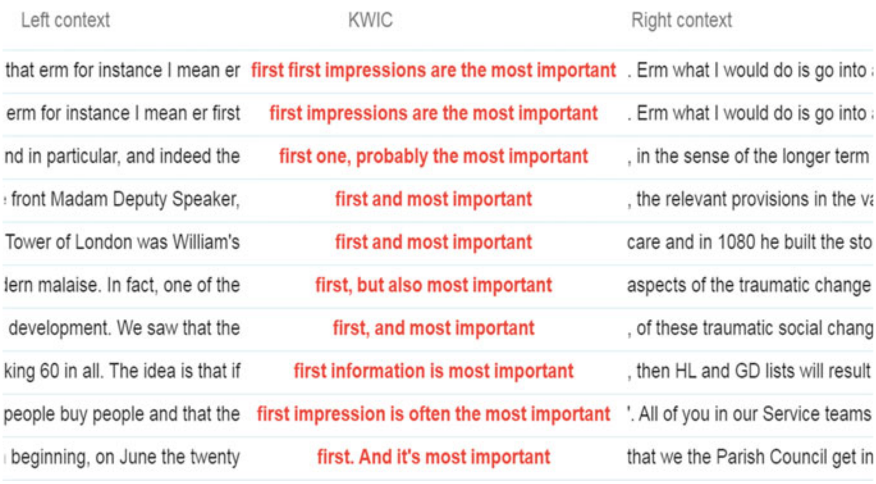
single	161	0	9.5	— mo
several	120	0	8.8	— mo
interesting	44	0	8.0	— mo
other	294	34	8.4	5.4 mo
first	110	67	8.0	7.4 mo
last	13	53	5.5	7.8 mo
second	104	499	8.8	11.4 mo
final	7	88	5.4	9.6 mo
fourth	7	308	5.5	11.5 mo
innocent	0	18	—	7.8 mo
consecutive	0	39	—	8.9 mo
successive	0	92	—	10.0 mo

expand_more

Fig. 3 The sketch difference results for *important* and *third* in BNC

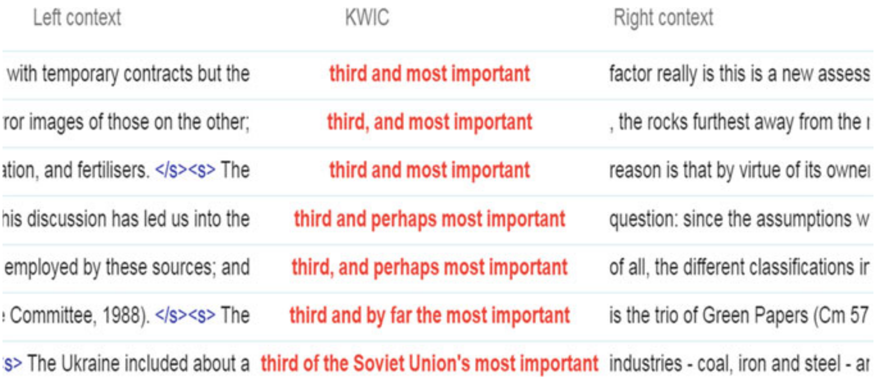
A specific “Word Sketch” (Fig. 6) of the word *first* also showed that it takes strong “and/or” collocation pattern with words indicating importance, such as *major, big, important,* and *foremost.* For example, in the concordance line, “The Fund's aim is first and foremost to secure the interests of developed countries,” the fixed phrase *first and foremost* reflects that the ordinal *first* has a conventionalized indication of being the most prominent.

Our corpus-based investigation through Thesaurus, Word Sketch Difference, Concordance, and Word Sketch showed strong evidence for the structural and



Left context	KWIC	Right context
that erm for instance I mean er	first first impressions are the most important	. Erm what I would do is go into
erm for instance I mean er first	first impressions are the most important	. Erm what I would do is go into
nd in particular, and indeed the	first one, probably the most important	, in the sense of the longer term
front Madam Deputy Speaker,	first and most important	, the relevant provisions in the va
Tower of London was William's	first and most important	care and in 1080 he built the sto
lern malaise. In fact, one of the	first, but also most important	aspects of the traumatic change
development. We saw that the	first, and most important	, of these traumatic social chang
king 60 in all. The idea is that if	first information is most important	, then HL and GD lists will result
people buy people and that the	first impression is often the most important	'. All of you in our Service teams
beginning, on June the twenty	first. And it's most important	that we the Parish Council get in

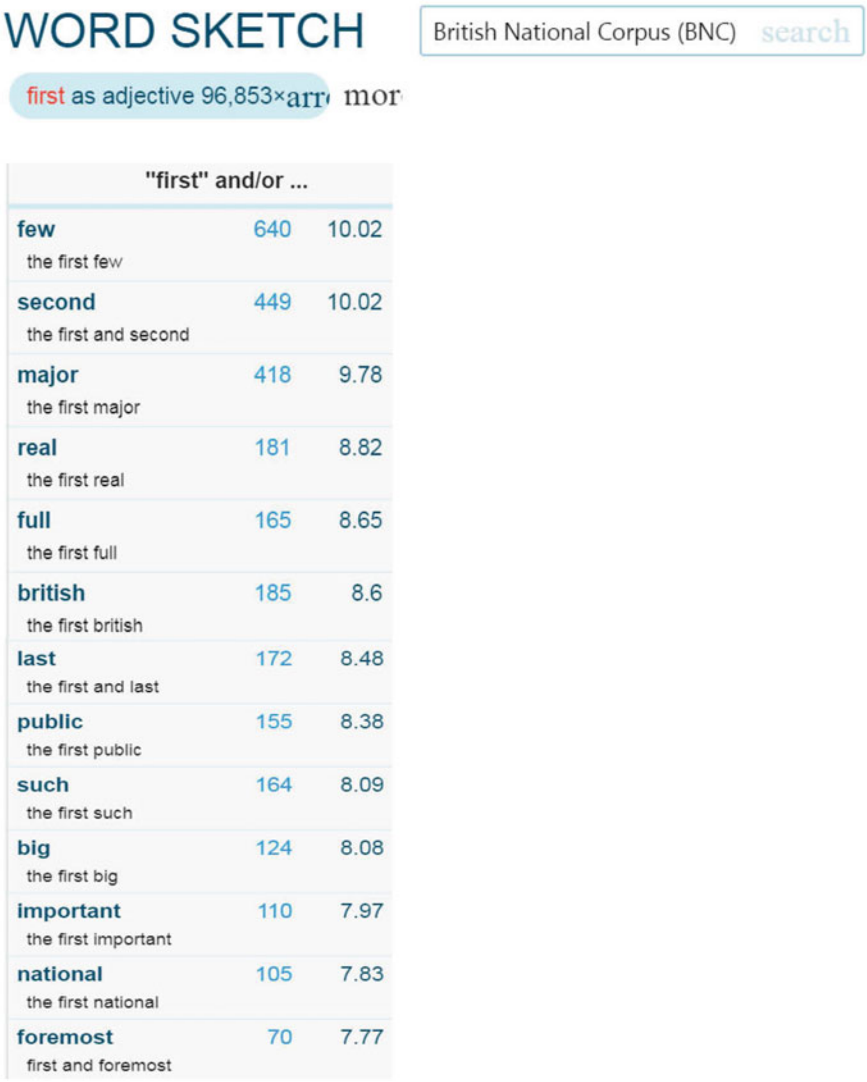
Fig. 4 Concordance results for "*first…most important*" in BNC (the first ten lines)



Left context	KWIC	Right context
with temporary contracts but the	third and most important	factor really is this is a new assess
ror images of those on the other;	third, and most important	, the rocks furthest away from the
ation, and fertilisers. </s><s> The	third and most important	reason is that by virtue of its owne
his discussion has led us into the	third and perhaps most important	question: since the assumptions w
employed by these sources; and	third, and perhaps most important	of all, the different classifications ir
Committee, 1988). </s><s> The	third and by far the most important	is the trio of Green Papers (Cm 57
s> The Ukraine included about a	third of the Soviet Union's most important	industries - coal, iron and steel - ar

Fig. 5 Concordance results for "*third…most important*" in BNC

cultural meaning of the ordinal *first* in English. As we demonstrated, English has strict rules of assigning *first* as the most important. This can also be evidenced by the general practice of "counting down" a list from the largest number and with the most important (*first*) being the last. For example, on the website listverse.com, which regularly publishes lists of interesting facts in human knowledge, the published top ten lists are always presented from the tenth to *first*. So is the listing of best movies in 2019 by *Time* (https://time.com/5737103/best-movies-2019/), in which the best movie is numbered with *One* but put at the bottom of the list. The most important issue is that *first* always refers to the most important/prominent, regardless of the sequence of the list; hence, *first* has to come last when the situation requires the most important item to be presented last. This convention shows that



WORD SKETCH

British National Corpus (BNC) search

first as adjective 96,853×arr mor

"first" and/or ...		
few the first few	640	10.02
second the first and second	449	10.02
major the first major	418	9.78
real the first real	181	8.82
full the first full	165	8.65
british the first british	185	8.6
last the first and last	172	8.48
public the first public	155	8.38
such the first such	164	8.09
big the first big	124	8.08
important the first important	110	7.97
national the first national	105	7.83
foremost first and foremost	70	7.77

Fig. 6 The Word Sketch results for *first* in BNC

the primary meaning of ordinal numbers when listing in English is importance. We will show in the next section that, in contrast, there is no strict conventional rule for the *first* to be most important in Chinese, and it seems that the primary meaning of ordinal numbers when listing in Chinese is just temporal ordering, with importance established otherwise.

4.2 *The Structural and Cultural Meaning of* Sān *in Chinese*

While *first* has the unique meaning of being the most prominent in English, the numeral 三 *sān* "three, third" also has its unique structural and cultural meaning in Chinese. Studies in symbolic meanings of Chinese numerals have described *sān* as a "complete" number, representing highest maturity and completeness (Shu 舒志武 2004). For example, in 三巡 *sān-xún* "three rounds, typically referring to the rounds of wine serving," 三鞠躬 *sān-jūgōng* "to bow three times," and 三顧茅廬 *sān-gù-máolú* "to pay three visits to the thatched cottage (of Zhuge Liang); to sincerely and repeatedly request someone to take up a post," the etiquette activity is not completed until the third iteration is completed. Furthermore, in 三思而後行 *sān-sī-ér-hòu-xíng* "thrice-think-and-then-act; to consider carefully before taking action," 舉一反三 *jǔ-yī-fǎn-sān* "raise-one-infer-three; to infer many other cases from one instance," *three* refer to the highest level of efforts (and not just the listing of exactly three things/times). This symbolic meaning might have originated in the Taoist culture in China, as Lao Tzu believed "道生一, 一生二, 二生三, 三生万物" "*Dào shēng yī, yī shēng èr, èr shēng sān, sān shēng wànwù*" "The Tao produces unity, unity produces duality, duality produces trinity, and the triad produces all things" (*Tao Te Ching*, Chapter 42, cf. Yu 2015: 10). Hence, *three* is used in the Chinese culture to represent "wholeness and fulfillment, to which nothing can be added" (Chevalier and Gheerbrant 1996: 993; cf. Yu 2015: 10). This sense of final completeness in *three* gives the number a more prominent meaning.

To further test if *three* indicates prominence in the modern time, we calculated the Mutual Information (MI) values for the collocation of *first/third* with *most important* in BNC and 第一/ 第三 with 最重要 in Gigaword, both setting 1–5 tokens between the node word and the collocating phrase. MI value, which measures the salience of collocation, is counted based on the algorithm $MI = \log_2 \frac{f(n,c) \times T}{f(n) \times f(c)}$. The MI values shown in Table 1 are no higher than three, except for the collocation of *first* with *most important*, but the general tendency is clear: *most important* is more closely collocated with *first* in comparison to *third* in BNC, but 最重要 *zuì-zhòngyào* "most important" is more closely collocated with 第三 *dì-sān* "third" in comparison to 第一 *dì-yī* "first" in Gigaword. The results suggest that *first* is more saliently associated with prominence in English and 第三 *dì-sān* "third" receives more prominence in Chinese.

Next, we ran corpus query language in CWS to do a more specific and refined concordancing for 第三 *dì-sān* "third" and 第一 *dì-yī* "first" in collocation with 最重要 *zuì-zhòngyào* "most important" on its right, setting 1–5 tokens in between and excluding cases in which units of person, party, volume, time sequence, or location (i.e., "人|夫人|主席|任|位|卷|黨|次|階段|季|季度|屆|號||回合|天|度|期|聲|步|所|道") immediately follow 第三/第一. After further manual filtering to rule out the instances meaning "the first important" or "the third important" (as such an expression does not precisely indicate that the *first/third* item is at the same time the most prominent one in a listing), we retrieved 22 examples (0.026 per million) in

Table 1 Comparing collocation salience for *first/third* with *most important* in BNC and 第一/ 第三 with 最重要 in Gigaword

Node Word (n)	*first*	*third*	第一	第三
Collocating word/phrase (c)	*most important*	*most important*	最重要	最重要
Collocation pattern	*first* {1–5 tokens} *most important*	*third* {1–5 tokens} *most important*	第一 {1–5 tokens} 最重要	第三 {1–5 tokens} 最重要
Corpus	BNC	BNC	Gigaword	Gigaword
MI Value	3.066720463	2.769083985	1.257074897	1.628994375
T(corpus size)	96,134,547	96,134,547	831,748,000	831,748,000
f(n,c)	49	7	74	37
f(n)	120,958	21,239	617,977	238,772
f(c)	4648	4648	41,671	41,671

Note f(n) = frequency of the node word, f(c) = frequency of the collocating word/phrase, f(n, c) = frequency of the co-occurrence of the node and collocating word/phrase, T(corpus size) = total number of words in the corpus

which 第三 *dì-sān* “the third,” meaning to mention thirdly in order of presence, is emphasized at the same time as the most important item in a listing (Fig. 7). On the contrary, we found much fewer cases (12 instances, 0.014 per million) for 第一 *dì-yī “first”* being emphasized as the most important point among other ones in a listing (Fig. 8), and a more careful check of contexts showed that some of the cases were uttered by a speaker working in the West (e.g., 王治郅 *Wáng Zhìzhì*, an NBA player) or translated from a foreigner’s speech. Comparing this finding with the occurrences of “*first/third…most important*” in BNC, it is evident that *sān* has much more prominence in the Chinese culture: The item ranked last in a hierarchy of three tends to be the most important in Chinese, whereas in English it is the *first* that should always be the most important.

5 From Tytler to Yan Fu: Three Principles of Translation Shaped by Translation

5.1 Translating Principles of Translation: Cross-Cultural Perspective

In Sect. 2, we speculated that Yan Fu’s principles of “*Xìn-Dá-Yǎ*” were most likely a translation from Tytler’s Three General Laws of Translation in exactly the same number and sequential order, with *Xìn* matching the first law, *Dá* relating to the second, and *Yǎ* mapping the third. Moreover, because Yan (1898) did not explicitly state the order of importance for these three principles as Tytler did, his translation resulted in controversial understandings in the relations among *Xìn, Dá,* and *Yǎ*.

混為一談 」 。</p><p>	第三 個 也 是 最 重要	的 原因 ， 依
進步 的 台灣 ，	第三 也 是 最 重要	的 一 點 ， 建
一 談判 。</p><p>	第三 個 、 也 是 最 重要	的 理由 是 中
的 基地 。</p><p>	第三 項 也 是 最 重要	的 因素 是 ，
多 得多 。</p><p>	第三 ， 也 就是 最 重要	的 ， 向來 廣
展現 的 成熟度 ；	第三 也 是 最 重要	的 ， 中國 大
就是 偷偷 入境 ；	第三 ， 也 是 最 重要	的 ， 無異 承
、 第二 協商 ，	第三 提供 服務 ， 目前 最 重要	的 是 要 兩
能 完成 。</p><p>	第三 、 目前 最 重要	的 工作 是
兩性 關係 。</p><p>	第三 項 也 是 最 重要	的 部份 ， 那
早日 明朗 。</p><p>	第三 、 特調會 最 重要	的 任務 ， 是
廉價 的 勞動力 ；	第三 ， 最 重要	的 ， 有 一 套
能 如願 。</p><p>	第三 ， 最 重要	的 一 點 是
根本 上 增強 。	第三 個 層次 ， 也 是 最 重要	的 層次 ， 就是
不復 返 。</p><p>	第三 個 原因 ， 也 是 最 重要	的 原因 ， 是
的 戰略 任務 。	第三 ， 發展 高科技 ， 人才 最 重要	。 資金 和 設備
於 應試教育 。	第三 、 最 重要	的 原因 是
科學 的 價值 ；	第三 也 是 最 重要	的 一 點 是
世界 名將 。</p><p>	第三 ， 也 是 最 重要	的 一 點 ， 就是
檢疫 工作 。</p><p>	第三 方面 、 也 是 最 重要	的 工作 ， 就是
正常 的 生活 ；	第三 ， 也 是 最 重要	的 ，美沙酮 維持
經 。</p><p>其中	第三 點 被 認為 是 最 重要	， 就 他 參觀

Fig. 7 Concordance results of "第三 … 最 重要" in the Gigaword Chinese Corpus

People in China tend to view *Xìn* 信 "faithfulness" as an easily attained baseline of translation, 達 *Dá "*expressiveness" as an intermediate, advanced level, and 雅 *Yǎ* "elegance" as the ultimate goal. The idea that *Yǎ* is more important than *Dá* and *Dá* is in turn more important than *Xìn* has been debated, but it has been held on to as the "given" in Chinese translation studies (especially literary translations), leading to a "detrimental trend" (cf. Chang 常謝楓 1981) in the society to privilege *Yǎ* (elegance) as the most important principle in translation practice.

結果 所 顯示出 的	第一 個 也 是 最 重要	的 事實 ， 是
有利 的 條件 。	第一 、 也 是 最 重要	的 是 ： 台灣
訓練 中心	第一 件 事 最 重要	的 就是 體能 和
成立 的 新政府	第一 件 而且 最 重要	的 事 ， 是 要
的 改建 不 是	第一 要務 ， 最 重要	的 是 要 改善
明顯 的 好處 。	第一 也 是 最 重要	的 ， 此 一 修正案
四 大 項 工作 ，	第一 項 是 最 重要	也 是 最 難 的
。</p><p> 我	第一 項 也 是 最 重要	的 義務 ， 是
的 數 項 議題 ，	第一 項 選擇 是 最 重要	的 ， 因為 美國
為期 9 天 飛行 的	第一 項 ， 也 是 最 重要	的 一 項 任務
要 做 的 「	第一 件 事情 ， 也 是 最 重要	的 事情 ， 就是
個 要素 。 </p><p>	第一 是 思想 。 這 是 最 重要	的 ， 也 王治郅

Fig. 8 Concordance results of “第一 … 最重要” in the Gigaword Chinese Corpus

In this sense, Tytler’s original principles of translation, in which the first law (i.e., Yan’s *Xìn)* was undoubtedly ranked as the most important among the three laws, was “shaped” by the translation into a version with reverse order of importance in the Chinese context. But, why and how did this mis-transformation occur? We assumed that this problem might be due to a failure in Yan’s translation of the itemized content to include the ideas regarding their internal order/relations, and such an assumption has been corroborated by our findings in Sect. 4. Based on a comparable corpora approach, we revealed substantial cultural differences in the structural and cultural meanings of *first* and *third* in English and Chinese. While ordinal numbers in Chinese are used in temporal/mention order, in English they are used in order of importance, with the *first* always being the most prominent. Chinese tends to adhere to the convention of *first* being the first presented, and there is a tendency to have ascending importance, with the *third* being the most important. Due to such culturally bounded interpretations, the fact that “the first principle” (i.e., *Xìn* “faithfulness”) is the most important in English was missed in Yan’s translation, and the third law (i.e., *Yǎ* “elegance”) was often held as the highest regardless of Yan’s own translation practice or people’s perception and implementation of Yan’s translation principles.

5.2 Translation as a Multi-Brain Activity

We observed earlier that Yan Fu's divergent interpretation of the second law/*Dá* from Tytler's is probably due to his taking a reader's perspective, instead of the writer's perspective by Tytler. To further explore this insight and to extend it to the broader theory of translation, we would like to borrow the multi-brain framework that has recently been advocated in neuro-cognitive studies (Schoot et al. 2016). The multi-brain perspective has been shown to provide important insights to the understanding of human interactive cognition, such as conversation. This can be extended to the paired writing and reading. Critically, the classical cognitive model assumes that speaking and listening involve encoding/decoding, which is the same process in reverse. This is called the single-brain perspective, as it assumes that it is the same (kind of) processor with the same mechanism to deal with coding and decoding. The simplistic single-brain perspective ignored the potential differences between the two distinct brains of the speaker/writer versus listener/reader. In addition, in conversation, the production and comprehension roles alternate to create even more complex interaction (what one says can influence what one hears, and vice versa).

Through examining this framework, we can see that translation involves three interactive brains: the author of the source text, the translator, and the readers of the translated text. In particular, it involves two pairs of two-brain activities (author-translator, translator-reader), each with the third playing a potentially critical role. The discussion of translation principles, which possibly include both Tytler's and Yan's, takes a single-brain perspective of the translator. That is, the laws/principles are designed to guide the translator when they are trying to optimize paraphrasing of what they were told. Yan Fu's three translation principles assumed, as indicated by his invoking Confucian classics on writing, that translation is another special scholarly writing activity. The translator attempts to paraphrase what they understand of the source text as (1) faithfully, (2) expressively, and (3) as elegantly as possible.

However, can today's translator work exclusively within one's brain? Our simple study showed that even the translation of simple concepts such as the ordinals *first* and *third* needs to accommodate culturally grounded interpretation differences. That is, the translator's brain must interact with the writer's brain to get the intended and complete meaning (i.e., the thinking process of the author's writing), and the reader's brain to know how he or she reads and interprets the translation task. It also needs to mediate the collective cultural and social brains of the two languages in terms of the collective interpretation of certain texts, i.e., how the same text will be read or interpreted differently in different languages.

The multi-brain perspective also brings more nuance to how to interpret and evaluate the translation principles. For instance, should *Xìn* "faithful" be understood as faithful to the author's original text, faithful to the translator's reading of the text, or faithful in the sense that the reader can understand the text faithfully? The same set of questions can be asked of the other two principles. However, once we consider the potential answers to these questions, it is obvious that implementing such principles from the translator's perspective is not optimal. For instance, Yan's choice of the style of Pre-Han language being graceful could not be made from the author's perspective and is not necessarily shared by the readers. Given this concern, we can simply perceive the multi-brain model of translation as a translator mediated two-brain model between the author and the reader. That is, the meaning generated from the brain of the author and permeated with the socio-cultural contexts of the writing, must be rendered to be read and interpreted as closely as possible by the brains of the intended readers, given the socio-cultural context of the reading of the translated text. In other words, a translator must simultaneously mimic how the author's and the readers' brains work, not just by themselves but also in their respective linguistic and socio-cultural contexts.

6 Conclusion

In this paper, we argued that Yan Fu's "*Xìn-Dá-Yǎ*" was most likely a translation from Tytler's Three General Laws of Translation. We also revealed that the internal relations among the three translation principles were shaped by the parallel translation into Chinese without taking into consideration the cultural context of the default ranking. This led to people's controversial views on "*Xìn-Dá-Yǎ"* for over a century, and their common misunderstanding of *Yǎ* as the most important principle in China. Comparing *first/third* in BNC and 一/三 in the Gigaword Chinese Corpus, it is evident that differences in the structural and cultural meanings of these ordinal numbers in English and Chinese led to the mis-transformation of his translation. We conclude that translation is a multi-brain activity situated in cultural contexts, concerning the translator's mediation of the original ideas and their adaption of meanings to map the target audience's culture. A translator's herculean task of mediating two brains in two different linguistic contexts is almost impossible and often results in distortion at one end or the other. A more nuanced view of the translation principles should be taken into consideration due to this complex interaction. A possible alternative would be to focus instead on the quality of the translator's mediation, such as following the single criterion of information quality (Huang and Wang 2020).

End Notes

1. Tytler's third edition of *Essay on the Principles of Translation* was originally released in 1813. It was reprinted as a new edition in 1978 by John Benjamins B.V with an introductory article by Jeffrey F. Huntsman.
2. There have been many printed versions for 天演論. We refer to the one printed by沔陽盧氏慎始基齋*Miǎn yáng lú shì shèn shǐ jī zhāi* in 1898, as it was the first formally released full version of 天演論 that included Yan Fu's 譯例言 *Yì-lì-yán* "Tranlsator's Preface."

3. Chao (趙元任 1969a) and Chao (1969b) are the English and Chinese versions of the same article, with the English version (1969a) briefer than the Chinese (1969b).
4. Based on our translation of Yan's arguments in *Yì-lì-yán*, we believe that Yan Fu claimed *Dá* as more important than *Xìn,* but did not explicitly put *Dá* ahead of *Yǎ* in terms of importance in this translator's preface, and there might be deep socio-cultural reasons for his vague and elusive interpretation of the relationship among the three principles. Besides, the difference in the subject matter could also account for the divergent interpretations for the importance of translating principles by Tytler and Yan Fu. As pointed out by one of the anonymous reviewers, Tytler's principles were designed to discuss literary translation, while Yan Fu's translations were mainly social, economic, and philosophical works. It is possible that different subject matter requires translators to do their job differently. Due to the limit of space, we only focus on the debates on whether *Xìn* or *Yǎ* is the primary principle in this paper. The other views regarding the order of importance for *Xìn*, *Dá*, and *Yǎ* will be further examined and reported separately in our future work.

Acknowledgments This study is supported by the Youth Foundation of Humanities and Social Sciences, Ministry of Education, China ("The construction of an ontology-based English-Chinese bilingual termbank in respirology").

References

Ai, Siqi. 艾思奇. 1937. 談翻譯. 语文 1(1).

Chan, T.L. 2004. *Twentieth-century Chinese translation theory: Modes, issues and debates*. Amsterdam/Pheladelphia: John Benjamins.

Chang, X. 常謝楓. 1981. 是"信", 還是"信、達、雅"? 外語教學與研究 (*Foreign Language Teaching and Research*) (4): 66–68.

Chao, Y.R. 1969a. Dimensions of fidelity in translation, with special reference to Chinese. *Harvard Journal of Asiatic Studies* 29: 109–130.

Chao, Y.R. 趙元任. 1969b. 論翻譯中信, 達, 雅的信的幅度. 中央研究院歷史語言研究所集刊, 1–13. http://www2.ihp.sinica.edu.tw/file/3935bvbdcwf.pdf

Chen, F. 陳福康. 2011. 中國譯學史. 上海 (Shanghai): 上海外語教育出版社 (Shanghai Foreign Language Education Press).

Chen, T. 陳廷祐. 1980. 英文漢譯技巧. 北京 (Beijing): 外語教學與研究出版社 (Foreign Language Teaching and Research Press).

Chevalier, J., and A. Gheerbrant. 1996. *A dictionary of symbols*. New York: Penguin.

Fan, C. 范存忠. 1978. 漫談翻譯. 南京大學學報(哲學社會科學版) (*Journal of Nanjing University (Philosophy and Social Sciences)*) (3): 86–95.

Fan, S. 2008. Ever since Yan Fu and his criteria of translation. In *Translation: Theory and practice, tension and interdependence*, ed. M.L. Larson. Amsterdam/Philadelphia: John Benjamins.

Fu, G. 傅國強. 1990. 對"信、達、雅"說的再思考. 中國科技翻譯 (*Chinese Science & Technology Translators Journal*) (4): 1–6.

Gu, L. 2010. Foreword: “Xin Da Ya” in translation and virtue. *Journal of Chinese Philosophy* 37 (4): 655–659. https://doi.org/10.1111/j.1540-6253.2010.01610.x.

Guo, J. 郭建中. 2013. 泰特勒翻譯三原則中譯辨正. 中國翻譯 (*Chinese Translators Journal*), (3), 68–70.

Huang, C.R. 2009. Tagged Chinese gigaword version 2.0. LDC2009T14. Linguistic Data Consortium, Philadelphia. https://catalog.ldc.upenn.edu/LDC2009T14.

Huang, C.R., and X. Wang. 2020. From faithfulness to information quality: On 信 in translation studies. In *Key issues in translation studies in China*, ed. L. Lim and D. Li, 111–142. Singapore: Springer.

Huang, C.R., S.K. Hsieh, and K.J. Chen. 2017. *Mandarin Chinese words and parts of speech: A corpus-based study*. London: Routledge.

Huang, C.R., A. Kilgarriff, Y. Wu, C.M. Chiu, S. Smith, P. Rychly, K.J. Chen et al. 2005. Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the fourth SIGHAN workshop on chinese language processing*, 48–55.

Huang, Y. 黃雨石. 1988. 英漢文學翻譯探索. 西安 (Xi’an): 陝西人民出版社(Shaanxi People’s Publishing House).

Huang, Z. 黃忠廉. 2016. 達: 嚴復翻譯思想體系的靈魂——嚴復變譯思想考之一. 中國翻譯 (*Chinese Translators Journal*) 37(1): 34–39.

Huang, Z. 黃忠廉, and Yuanfei Chen 陳元飛. 2016. 从达旨术到变译理论. 外語與外語教學 (*Foreign Languages and Their Teaching*) (1): 98–106.

Hui-Chih, Yu. 2015. A comparative study of the meanings of numbers in English and Chinese cultures. *Intergrams* 16(1). Retrieved from http://benz.nchu.edu.tw/～intergrams/intergrams/161/161-yu.pdf

Kenning, M.M. 2010. What are parallel and comparable corpora and how can we use them. In *The Routledge handbook of corpus linguistics*, ed. A. O’Keeffe and M. McCarthy. Abingdon, UK: Routledge. https://doi.org/10.4324/9780203856949.ch43

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, and V. Suchomel. 2014. The sketch engine: Ten years on. *Lexicography* 1 (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Li, P. 李培恩. 1935. 論翻譯. 之江學報 (4).

Li, T. 李田心. 2014. 泰特勒翻譯三原則的漢譯再辨正——兼論嚴復三字原則和泰特勒三原則乃異曲同工之作 (Further amendment of the translation of the third principle of Tytler ’ s principles of translation). 樂山師範學院學報 (*Journal of Leshan Normal University*) 29 (11): 46–49.

Liang, Q. 梁啟超. 1922. 佛典之翻譯. In 梁啟超 (Ed.), 飲冰室合集·專集, *Collection 14* (p. 64). 上海 (Shanghai): 中華書局 (Zhonghua Book Company).

Luo, X. 羅新璋. 1983. 我國自成體系的翻譯理論 (Our country’s translation theory: A system of its own). 翻译通讯 (*The Translator’s Notes*) (7): 9–13.

Luo, X. 羅新璋, and Y. Chen. 陳應年 eds. 2009. 翻譯論集 (*An anthology of essays on translation*) (Revised ed). 北京 (Beijing): 商務印書館 (The Commercial Press).

Lv, B. 呂博. 1998. 也談翻譯的標準和原則. 中國翻譯 (*Chinese Translators Journal*) (3): 3–5.

Munday, J. 2016. *Introducing translation studies: Theories and applications*. Routledge.

Qian, Z. 錢鍾書. 1986. 譯事三難. In 錢鍾書 (Ed.), 管錐編 (2nd ed., p. 1161). 上海 (Shanghai): 中華書局 (Zhonghua Book Company).

Qu, Q. 瞿秋白. 1931. 魯迅和瞿秋白關於翻譯的通信–瞿秋白的來信. 十字街頭 (1).

Rener, F.M. 1989. *Interpretatio: Language and translation from Cicero to Tytler*. Amsterdam: Rodopi.

Schoot, L., P. Hagoort, and K. Segaert. 2016. What can we learn from a two-brain approach to verbal interaction? *Neuroscience and Biobehavioral Reviews* 68: 454–459.

Shen, S. 沈蘇儒. 1982. 論“信、達、雅.” 編譯參考 (2).

Shen, S. 沈蘇儒. 1998. 論信達雅: 嚴復翻譯理論研究 (*On “Xin-Da-Ya”: The Chinese principle of translating*). 北京 (Beijing): 商务印书馆 (The Commercial Press).

Shu, Z. 舒志武. 2004. 数词“三”的文化意义分析. 华南农业大学学报(社会科学版) (*Journal of South China Agricultural University (Social Science Edition)*) 4(2): 1–3.

The British National Corpus, version 3 (BNC XML Edition). 2007. *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*. http://www.natcorp.ox.ac.uk/.

Tytler, A.F. 1790 [1907]. *Essay on the principles of translation*, 1st ed. London: Dent.

Tytler, A.F. 1813[1978]. *Essay on the principles of translation*, 3rd ed. Amsterdam: John Benjamins B.V.

Wang, K. 王克非. 1992. 論嚴復《天演論》的翻譯. 中國翻譯 (*Chinese Translators Journal*) (3): 6–10.

Wang, Hongzhi. 王宏志. 1999. 重釋 “信達雅”: 二十世紀中國翻譯研究. 上海 (Shanghai): 東方出版中心 (Orient Publishing Centre).

Wang, Hongzhi. 王宏印. 2003. 中国传统译论的经典阐释: 从道安到傅雷. 武漢 (Wuhan): 湖北教育出版社 (Hubei Education Press).

Wang, B. 王秉欽. 2017. 近現代中國翻譯思想史. 上海 (Shanghai): 华东师范大学出版社 (East China Normal University Press).

Wu, Lifu. 伍蠡甫. 1980. 伍光建的翻譯. In 伍光建 (Ed.), 伍光建翻譯遺稿. 北京 (Beijing): 人民文學出版社 (People’s Literature Publishing House).

Wu, C. 吳存民. 1997. 論“信達雅”的有機完整性——兼評譯論中的一種錯誤傾向. 中國翻譯 (*Chinese Translators Journal*) (5) : 39–41.

Xu, Shouping. 徐守平, and Shouqin. Xu. 徐守勤. 1994. “雅”義小論——重讀《天演論·譯例言》. 中國翻譯 (*Chinese Translators Journal*) (5) : 3–5.

Yan, F. 嚴復. 1898. 天演論. 沔阳 (Mianyang): 沔阳卢氏慎始基斋 (Miǎn yáng lú shì shèn shǐ jī zhāi).

Yang, L. 楊麗華. 2011. 中國近代翻譯家研究. 天津 (Tianjin): 天津大學出版社 (Tianjin University Press).

Yang, Z. 楊自儉, and X. Liu. 劉學雲 eds. 2003. 翻譯新論: 1983–1992 (2nd ed.). 武漢 (Wuhan): 湖北教育出版社 (Hubei Education Press).

Ye, J. 葉君健. 1997. 翻譯也要出“精品.” 中國翻譯 (*Chinese Translators Journal*) (1): 30–31.

Zhang, W. 張威廉. 1984. 怎樣提高我們文學翻譯的質量?. In 翻譯研究論文集 (1949–1983) (p. 465). 北京 (Beijing): 外語教學與研究出版社 (Foreign Language Teaching and Research Press).

Zhang, Y. 張英倫. 1988. “信、達、雅”芻議. 《瞭望》周刊 (*Outlook Weekly*) (11).

Zhao, W. 趙巍, and C. Shi. 石春讓. 2005. 比較譯學的個案研究引發的思考 ——從嚴復的“信達雅”與泰特勒的三原則說起. 外語學刊 (*Foreign Language Research*) (5): 96–101.

Zheng, Y. 鄭意長. 2002. 近代翻譯思想的演進. 天津 (Tianjin): 天津古籍出版社 (Tianjin Ancient Books Press).

Zheng, Z. 鄭振鐸. 1921. 譯文學書的方法如何? (How to translate literary work?) . 小說月報 (*The Short Story Magazine*) 12(3).

Website Resources

Chinese Word Sketch (CWS). http://wordsketch.ling.sinica.edu.tw/.
The Sketch Engine. http://www.sketchengine.co.uk/.

Chu-Ren Huang (PhD, Cornell; DHC, Aix-Marseille) is a Chair Professor at the Hong Kong Polytechnic University. He has published 25 book or edited volumes, 30 online or licensable language resources, 136 journal articles, and 125 book chapters. His main research areas include computational and corpus linguistics, digital humanities, lexical semantics, and ontology. His papers have appeared in *Cognitive Linguistic Studies, Computational Linguistics, Corpus Linguistics and Linguistic Theories, Humanities and Social Sciences Communications, IEEE*

Access, Journal of Chinese Linguistics, Journal of Quantitative Linguistics, Knowledge-Based Systems, Language and Linguistics, Language Resources and Evaluation, Lingua, Lingua Sinica, Linguistics, Natural Language Engineering, PLoS One, SAGE Open, among others.

Xiaowen Wang is an Associate Professor of Applied Linguistics and the director of the Research Center for English Education and Linguistic Studies in the School of English Education, Guangdong University of Foreign Studies, Guangzhou, China. She also serves as the associate editor of the *The International Journal of English for Specific Purposes*. Currently, she is doing her doctoral study under the supervision of Professor Chu-Ren Huang in the Faculty of Humanities, the Hong Kong Polytechnic University. Her research interests cover pragmatics, English for medical purposes, lexical semantics, corpus linguistics, computational linguistics, discourse analysis, and translation.

Going to Understand 柴? Evidence and Significance of Metonymic Chains in Chinese/English Translation

Zi-yu Lin

Abstract Based on three researches using big data (Chinese/English corpora) and small data (translation of a particular Chinese character), and assisted by other diachronic and synchronic records, this article attempts to establish the central argument that metonymic chains do exist in Chinese/English Translation practice. By metonymic chains in translation, it is meant that a chain of correlatively motivated metonymic extensions or inferences that are found in the multiple target language renditions of the source language original, and these semantic extensions or inferences are in compliance with human metonymic operations. The first study is on the Chinese term 明白 [míng bái] and its English equivalents. The second research study investigates the relationship between the English verbal and grammatical phrase *be going to* and its numerous translated Chinese counterparts. The third is a small part of a meticulous analysis of the English translations of Wang Wei's 鹿柴 [lù chái/zhài]**,** with the focus on the English translations of the Chinese character 柴. The findings show that many variations in Chinese/English translation cannot be a simple matter of unmotivated randomness or arbitrariness in style or diction. Rather, they are the demonstrations of different links in a metonymic chain that is justifiable by the cognitive rationale and can be uncovered when adequate diachronic and synchronic data are examined from the embodiment and frame semantics perspectives. In the translation practice, the metonymic conceptual movements are more fundamental than metaphorical projections, which are often found to be made by smaller metonymic operations in a chain. Based on the curves depicting the one-to-many relations between the source language originals and the target language renderings that constitute the metonymic chain, it is found that the Pareto Principle offers a close mathematical approximation of the data. In Chinese/English translation practice and studies, this means, about 20% of the data is able to account for about 80% of the total translation outputs. Pedagogically and theoretically, therefore, this 20% deserves our special attention.

Z. Lin (✉)
Macao Polytechnic Institute, Macao, China
e-mail: zylin@ipm.edu.mo

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_9

Keywords Chinese/English translation · Bilingual corpora · Metonymic chain · Pareto principle · Embodiment · Frame semantics · Metonym · Metaphor

1 Introduction

This article is based on three researches using big data (Chinese/English corpora), small data (translation of a particular Chinese character), and other diachronic and synchronic records to establish the argument that metonymic chains (转喻链 [zhuǎn yù liàn]) do exist in Chinese/English translation.

Metonymy is a cognitive process that allows us to use one well-understood aspect of an entity to stand for the thing as a whole, or for some other aspect of it, or for the entity to which it is very closely related (Gibbs 1994, p. 11). As Littlemore (2018) rightly points out, metonyms function as shortcuts for our language, thoughts, and communication. For instance, a personal name can function as an effective shortcut to the reference of a particular human being. At a more theoretical and abstract level, according to Barcelona (2002, p. 246), a metonym is a mapping of a source domain to a target domain, both of which are in the same functional domain and are linked by a pragmatic function, whereby the target is mentally activated.

A classic definition of metonymic chains was proposed in Barcelona (2005, pp. 328–331), which refers to a "direct or indirect series of conceptual metonymies guiding a series of pragmatic inferences." During these metonym-based pragmatic inferences, metonym X can trigger the inference of metonym Y when X is the main factor responsible for leading the comprehender to Y. Alternatively, a metonym X can also facilitate the inference of Y when X is able to provide part of the conceptual material that leads to Y. A typical example is that the nominal phrase "gas pump" can trigger the inference of a "gas station" through the SALIENT PART OF THE WHOLE FOR THE WHOLE metonym, because the gas pump can be the main factor that guides the comprehender toward the gas station. Then, a gas station in turn can facilitate the inference of a car, which in turn can further facilitate the inference of the insurance industry. Although the car is not the only factor with regard to insurance service, it does provide part of the conceptual context for it in the comtemporary society. By the metonymic chain in this study, it is meant that a chain of such correlated semantic extensions or inferences found in the multiple target language translations that are triggered or facilitated step by step by the conceptual materials in the source language original, and these semantic extensions or inferences are in compliance with human metonymic operations.

The first research project is the study on the Chinese term 明白 [míng bái] and its English equivalents, in an effort to address issues on how abstract concepts like UNDERSTAND are expressed within a language and across languages and why they have happened the way that has been. The second investigates the relationship between the English verbal and grammatical phrase *be going to* and the numerous Chinese counterparts it has, which collectively traverse a wide range of meanings, including movement, intention, future, and modality, to clarify the seemingly chaotic interchangeability between *be going to* and its Chinese counterparts. The third is a detailed analysis of the English translations of the Chinese character 柴 in 鹿柴 [lù chái/zhài], which is the title of a poem authored by Wang Wei.

The compelling evidence and interesting findings of the metonymic chain discovered in the 明白 [míng bái] project provided the enlightening and guiding principles in examining the data in the *be going to* and 鹿柴 [lù chái/zhài] projects, where the one-to-many relations are also strikingly similar. The outcomes of these efforts demonstrate that even with different sizes or sets of data, variations in translation can be accounted for by metonymic chains. In other words, these synonymous variations are motivated and linked by metonymic inferences.

In addition, the 明白 and *be going to* studies attempt to address the following three questions: Firstly, what are the English equivalents of the Chinese term 明白 and why it is so? Reversely, given the semantic/grammatical continuum that *be going to* can represent, how does the Chinese language manage to cover this continuum? Secondly, in these two pairs of the one-to-many relations (i.e. 明白 vs. many English equivalents, and *be going to* vs. many Chinese equivalents), how does "many" become the semantic equivalents of "one" in Chinese/English translation? Thirdly, is there a mathematical model that can approximate these relations and what theoretical and pedagogical implications could such a model bring to us?

Juxtaposing the naming elements of the three projects in English and Chinese i.e. "going to", "understand" and "柴", we arrive at the beginning portion of the title for this article: *Going to Understand* 柴.

2 Research Methodologies

2.1 The 明白 *Project*

The 明白 project examines the English equivalents of the Chinese lexical item 明白 [míng bái], utilizing such databases as

The Chinese-English Sentence Aligned Bilingual Corpus. Published by the Chinese Linguistic Data Consortium (CLDC) in 2007, it provides experimental data for the modeling and analysis of bilingual models based on statistics. It also affords samples of real text labels for extracting word pairs and phrase pairs for machine translation and information search among different languages. *The Chinese-English Sentence Aligned Bilingual Corpus* purchased by the Macao Polytechnic Institute Library contains one million pairs of sentences (cf. http://shachi.org/resources/1215).

Chinese Text Project (CTP) (中国哲学书电子化计划 [zhōnggúo zhéxué shū diànzǐhuà jìhuá]). This is an online open-access digital library that makes pre-modern Chinese digital texts available to readers and researchers all around the world, featuring a gold mine of valuable diachronic information (cf. ctext.org). Many of the Chinese texts are accompanied by their authentic English translation.

This continuously developing website is one of the rare resources offering free, powerful information for Chinese historical linguistics research.

CC-CEDICT. The objective of the CEDICT project was to create an online, downloadable (as opposed to searchable-only) public-domain Chinese-English dictionary (cf. https://cc-cedict.org/wiki/). This dictionary is integrated into the search and statistics tool home-developed for the 明白 project to provide the initial identification, retrieving, and frequency calculation of the English equivalents.

The Oxford English Online (OED) Premium Collection. This gigantic database presents the definitive record of the English language and offers the best resources for the etymological analysis of English. It is an indispensable tool for studying the semantic evolution of English lexical items.

百度百科 ([bǎi dù bǎi kē] *The Baidu Encyclopedia*, cf. http://baike.baidu.com/). Claiming to be the world's largest encyclopedia in Chinese, this vibrantly growing database advocates the noble principle that all people share knowledge freely and equally. For the research in this article, it is used as the second source for fact verification and data collection, especially for the issues in Chinese language studies.

The Chinese-English Sentence Aligned Bilingual Corpus was delivered to the Library of Macao Polytechnic Institute without any search and statistical capabilities. The data are found in the following format (without *Pinyin* and the gloss):

(1) 然而，在许多国家设备陈旧，对确保数据质量及其及时转递都造成了严重问题, 一旦设备出现故障, 又出现如何处理遗漏的资料问题。

[ránér, zài xǔduō guójiā shèbèi chénjiù, duì quèbǎo shùjù zhìliáng jí qí jíshí zhuǎndì dōu zàochéng le yánzhòng wèntí, yīdàn shèbèi chūxiàn gùzhàng, yòu chūxiàn rúhé chǔlǐ yílòudí zīliào wèntí

but in many country equipment old, for assure data quality and timely transformation all make serious problem, once equipment have trouble, also occur how process missing data problem]

However, obsolete equipment in many countries poses severe problems for ensuring data quality and their timely transmission, as well as for coming to grips with missing information when equipment breaks down.

Therefore, it is essential that a searching tool be developed so that the desired data can be accurately and efficiently retrieved and presented. With the help from Mr. Terence Chi Ip Tai, Head of the Systems and Client Services of Macao Polytechnic Institute Library, a searching and statistical program as follows was designed and developed:

Through this tool, *the Chinese-English Sentence Aligned Bilingual Corpus* is queried step by step with regard to 明白 until satisfactory data are obtained.

2.2 The Be Going to Project

Through personal communication, the author was allowed the privilege of using the English/Chinese parallel corpus developed by Lu Wei of Xiamen University, China. At the time of my accessing, it contained 215,713 parallel English/Chinese sentences, 3,290,670 English word tokens, and about 5,370,429 Chinese character tokens. The search string is "*going to*" and all of the 765 valid hits thus generated are used. The acquired data are filtered through a home-grown computer program, whereby a number of individual Chinese characters or character strings are keyed in to parse the data and sort them out.

For both of the 明白 and *going to* projects, careful diachronic and synchronic analyses are conducted for the representative sentences and lexical items retrieved from the bilingual databases. The aim is to better understand and explain the drastic variations in translation renderings. Finally, frequency curves on the "many" parts are constructed and analyzed to find out a likely statistical model.

2.3 The 鹿柴 Project

The data source is from Weinberger, Paz, and Wang (1987), which lists 19 different English translations of a poem authored by Wang Wei and titled 鹿柴 [lù chá/zhài]. The diversified translation renderings form the typical one-to-many relationship between the Chinese original and its English translations. This provides us with a good sample with valuable clues to explore the grounds for the drastic variations in translating Tang Poetry into English. Specifically, the focus of the discussion and analysis here is on how one Chinese character 柴 is translated.

3 The Data and Findings

3.1 Findings and Analyses of The明白 Project

Based on the search and statistical tool described in Fig. 1 and after the further manual operation to consolidate the data, 明白 is found to have appeared 737 times in the *Chinese-English Sentence Aligned Bilingual Corpus*, with 49 different English equivalents, as in Table 1.

Below are some of the typical data:

(2) 明白**:** understand

重要的是岛国居民自己也明白这是一个不可避免的步骤，因为唯一现实的选择就 是共同前进。

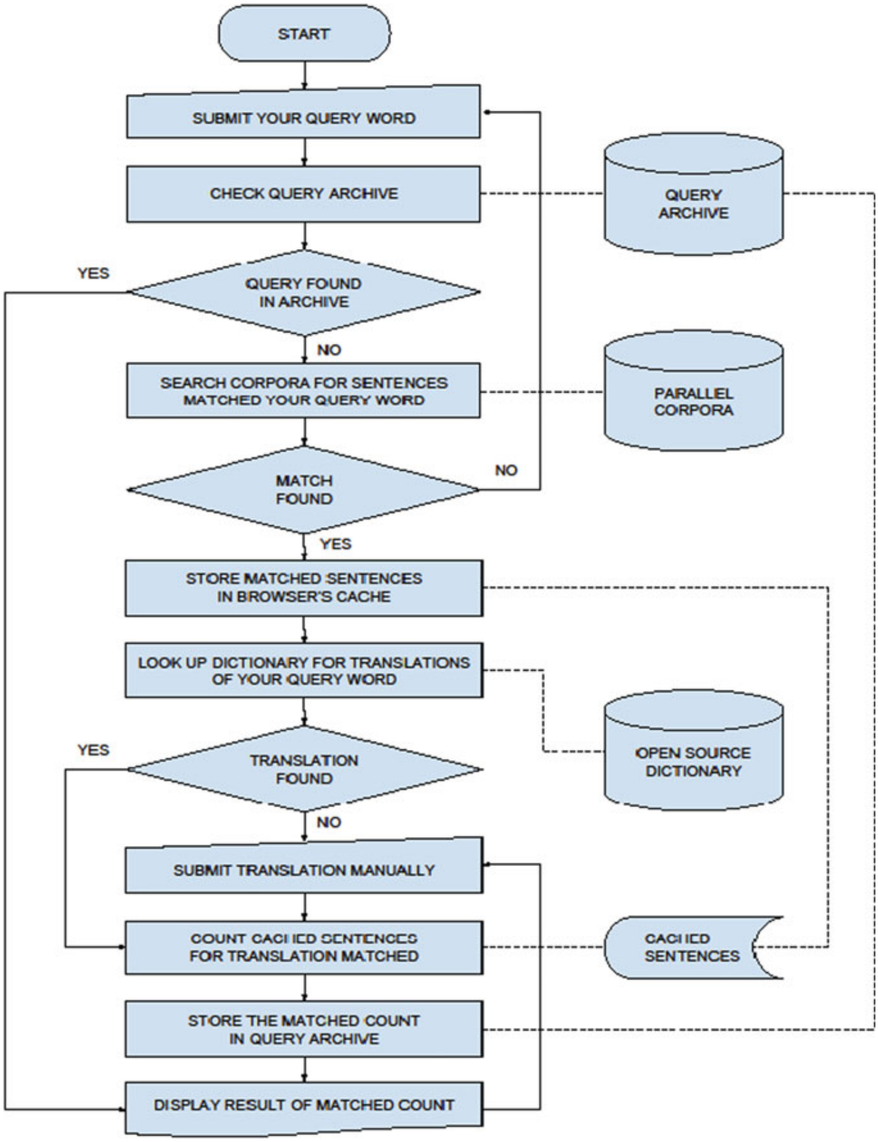

START
SUBMIT YOUR QUERY WORD
CHECK QUERY ARCHIVE
QUERY ARCHIVE
YES
QUERY FOUND IN ARCHIVE
NO
SEARCH CORPORA FOR SENTENCES MATCHED YOUR QUERY WORD
PARALLEL CORPORA
NO
MATCH FOUND
YES
STORE MATCHED SENTENCES IN BROWSER'S CACHE
LOOK UP DICTIONARY FOR TRANSLATIONS OF YOUR QUERY WORD
OPEN SOURCE DICTIONARY
YES
TRANSLATION FOUND
NO
SUBMIT TRANSLATION MANUALLY
COUNT CACHED SENTENCES FOR TRANSLATION MATCHED
CACHED SENTENCES
STORE THE MATCHED COUNT IN QUERY ARCHIVE
DISPLAY RESULT OF MATCHED COUNT

Fig. 1 The Searching and Statistical Tool Home-developed for the 明白 Project

zhòngyàodí shì dǎogúo jūmín zìjǐ yě míngbái zhè shì yīgè bù kě bìmiǎnde bùzhòu, yīnwèi wéiyī xiànshíde xuǎnzé jìu shì gòngtóng qiánjìn。

important be island inhabitant self also understand this is a inevitable step, because only realistic alternative be together forward

It is important that the islanders themselves also understand that this is an inevitable step, because the only realistic alternative is to march forward together.

Table 1 明白 and its English equivalents

Serial number	English equivalents of 明白	Frequency
1	understand(ing)/understood	340
2	clear(ly)	124
3	know/knowledge	50
4	see	34
5	express	29
6	plain	28
7	aware(ness)	17
8	appreciate/appreciation	15
9	learn	11
10	obvious	11
11	explicit	9
12	realize	8
13	show	5
14	get the idea	4
15	find out	4
16	specific(ally)	3
17	intelligent(ly)	3
18	recognize	3
10	believe	2
20	has noted	2
21	committed	2
22	flatly	2
23	finally reached the place where it truck her	2
24	convinced	2
25	get the message	2
26	clarify/clarity	2
27	find their way	1
28	evident	1
39	apparently	1
30	comprehend/comprehension	1
31	grasp	1
32	familiar	1
33	spell out	1
34	perceive	1
35	make out	1
36	self-illustrative	1
37	self-explanatory	1
38	stark	1
39	unmistakably	1
40	profess	1

(continued)

Table 1 (continued)

Serial number	English equivalents of 明白	Frequency
41	surmise	1
42	reason out	1
43	dawn upon	1
44	bear in mind	1
45	hammer into our heads	1
46	catch your meaning	1
47	this is a clue to	1
48	give the reader a sense of	1
49	has the notion	1
Total		**737**

(3) 明白**:** clear

联柬权力机构向柬埔寨当局明白表示，他们必须对各少数民族提供充分的保护 。

liánjiǎn quánlì jīgòu xiàng jiǎnpǔzhài dāngjú míngbái biǎoshì , tāmén bìxū

duì gè shǎoshù mínzú tígōng chōngfēndí bǎohù

UNTAC to Cambodian authority clear express, they must to all minority provide adequate protection

UNTAC has made it clear to the Cambodian authorities that they are obliged to provide adequate protection to ethnic minorities.

(4) 明白: know/knowledge

全世界的农民们都明白季节的重要性和永恒性。

quán shìjiè de nóngmínmén dōu míngbái jìjié de zhòngyàoxìng hé yǒnghéngxìng

whole world POSS peasant all know season POSS importance and immutability

Farmers all over the world know the importance and immutability of the seasons. (POSS = Possessive)

The data findings presented so far have partially addressed the first question posed earlier, namely, what are the English equivalents of the Chinese term 明白? Such a finding, however, is by and large descriptive and mechanical, because it does not explain why the Chinese term明白is able to afford so many English equivalents, a question that calls for a more insightful theoretical exploration diachronically and synchronically. To do so, we have to dig deep to reach the roots of this Chinese term in Classical Chinese, the written Chinese language that had been in dominant use from the fifth century B.C. until the beginning of the twentieth century, uncovering the original literal meanings of the two component Chinese characters 明 and 白 and examining their subsequent semantic extensions.

Simply put, we want to know how 明白 eventually evolved to mostly mean UNDERSTAND through a historical linguistics probe.

A careful research using the *Chinese Text Project* and 百度百科 *(The Baidu Encyclopedia)* reveals the etymological sophistication of the character 明 in the literature of Classical Chinese. First, 明 is a compounded ideograph made of 日 ([rì], the sun) and月 ([yuè], the moon), both of which emit light. As a result, the original meaning of 明 was light. From this point on, a semantic network has been developed, as the data in Table 2 show (unless indicated otherwise, the English translations are retrieved from the *Chinese Text Project*).

Evidently, the light of the Sun and the Moon, as symbolized by the Chinese character 明, is bright and brilliant, and therefore provides the physical condition for humans to see with their naked eyes. Later on, the semantic contents of 明 ramified in several directions, mainly metonymically. Through the CONDITION FOR PHYSICAL ABILITY metonymy, there emerged EYESIGHT, and SEE CLEARLY, where LIGHT triggers EYESIGHT, and good illumination triggers SEE CLEARLY. Then, CLEAR, OBVIOUS, UNDERSTAND, and MAKE WISE are facilitated through the CAUSE FOR RESULT metonymy. When one can see something, it becomes obvious. When something becomes clear to him, he then can understand it. Once he understands, he can be wiser. Again, a metonymic operation refers to an association between two entities in one conceptual frame so that one entity can stand for the other (Evens and Green 2006, p. 167), with various relations in between (Bredin 1984; Kövecses 2013). In Contemporary Cantonese, which has inherited a great deal from Classical Chinese semantically and phonologically, 明 is still single-handedly used as a verb to mean UNDERSTAND, as in

(5) 明唔明?

ming4 ng6 ming2

clear not clear

Do you understand?

With regard to the semantics of 白 [bái], this pictographic character was first found in the oracle bone inscriptions with the shape of sunlight shooting up and down to denote the brightness of the sun (cf. hydcd.com, a compressive site for different kinds of Chinese dictionaries). The brightness of the sun was perceived as white during the daytime. Therefore, one of the earliest uses of this character refers to daylight, as in

(6) 秋爲白藏。《爾雅·釋天》(403–221 B.C)

qīu wéi bái cáng

fall be white store

Table 2 Semantic network of character 明 in classical Chinese literature

The Meanings of 明 in the Literature of Classical Chinese	Data Sources	Examples
A combination of the light from the Sun and the Moon	《周易•系辞 下》(9 C. B.C)	日往则月来, 月往则日来, 日月相推而明生焉。 rì wǎng zé yuè lái, yuè wǎng zé rì lái, rì yuè xiàng tūi er míng shēng yān The sun goes and the moon comes; the moon goes and the sun comes; the sun and moon thus take the place each of the other, and their shining is the result
Bright, brilliant	《詩經•雞鳴》(11–6 C. B.C)	東方明矣、朝既昌矣。 dōng fāng míng yǐ、 cháo jì chāng yǐ The east is bright; the court is crowded
Clear, obvious	《公孫龍子•白馬論》(320–250 B.C.)	可與不可, 其相非明。 kě yǔ bù kě, qí xiàng fēi míng Acceptable and unacceptable are clearly in opposition to each other
Eyesight	《孟子•梁惠王》(468–376 B.C.) 《禮記•檀弓上》(5th C.–221 B.C.)	明足以察秋毫之末 。 míng zú yǐ chá qīu háo zhī mò My eyesight is sharp enough to examine the point of an autumn hair 子夏喪其子而喪其明。 zǐ xià sàng qí zǐ er sàng qí míng When Zi-xia was mourning for his son, he lost his eyesight
See clearly	《荀子•勸學》(Circa 313–238 B.C.)	目不能兩視而明, 耳不能兩聽而聰。 mù bù néng liǎng shì er míng, er bù néng liǎng tīng er cōng The eye cannot look at two objects and see either clearly; the ear cannot listen to two things and hear either distinctly (Knoblock, 1988, p.139)
Straightforwardly, clearly	《孟子•梁惠王上》(372–289 B.C.)	願夫子輔吾志, 明以教我。 yuàn fū zǐ fǔ wú zhì, míng yǐ jiào wǒ I wish you, my Master, to assist my intentions. Teach me clearly
Clarify, make clear	《墨子•小取》(468–376 B.C.)	夫辯者, 將以明是非之分。 fū biàn zhě, jiāng yǐ míng shì fēi zhī fēn “Distinguishing” will be used to make clear the distinction between so and not so
Understand	《墨子•尚賢下》(468–376 B.C.)	我以此知天下之士君子, 明於小而不明於大也。 wǒ yǐ cǐ zhī tiān xià zhī shì jūn zǐ, míng yú xiǎo er bù míng yú dà yě Then I know the gentlemen understand only trifles and not things of significance

(continued)

Table 2 (continued)

The Meanings of 明 in the Literature of Classical Chinese	Data Sources	Examples
Wise	《道德經》 (Circa 5th C. B.C.)	不自見, 故明。 bù zì xiàn, gù míng He is free from self-display, and therefore he shines/is wise
Make wise	《道德經》 (5th C. B. C.)	古之善為道者, 非以明民, 將以愚之。 gŭ zhī shàn wéi dào zhě, fēi yĭ míng mín, jiāng yĭ yú zhī The ancients who showed their skill in practising the Dao did so, not to enlighten the people, but rather to make them simple and ignorant

During the daylight in autumn, the harvest was stored.

Based on the semantic contents of light, 白 is also found to mean MENTALLY CLEAR and UNDERSTAND, as in

(7) 禮義不加於國家, 則功名不白。《荀子·天论》(316–235 B.C.)
lĭ yì bù jiā yú gúo jiā, zé gōng míng bù bái
ritual not set up to state, then honor rank not clear understand
If rituals are not established for a country, the scholarly honor and official ranks cannot be clearly understood.

The compound 明白 was formed early, and has been polysemous with numerous meanings ranging from the concrete to the abstract concepts, such as bright/brilliant/daytime/clear, obvious/clarify/make clear/understand/straightforwardly/clearly/clean/wise/make wise/eyesight/good eyesight/see clearly. Some typical uses of this compound include

(8) 此皆生於法明白易知而必行。《商君書·定分》(403–221 B.C.)

cĭ jiē shēng yú fă míng bái yì zhī er bì xíng

this all originate from law clear easy know and apply

All this originates from the fact that the law is clear, easy to know, and strictly applied.

(9) 夫明白於天地之德者, 此之謂大本大宗, 與天和者也。《莊子·天道》

(403–221 B.C.)

fū míng bái yú tiān dì zhī dé zhě, cĭ zhī wèi dà běn dà zōng, yŭ tiān hé zhě yě

alas clear understand at heaven earth POSS virtue thing, this POSS call great root great origin, with heaven harmonize AFF (POSS = Possessive, AFF = Affirmative)

The clear understanding of the virtue of Heaven and Earth is what is called “The Great Root”, and “The Great Origin”—they who have it are in harmony with Heaven.

(10) 王冕看書, 心下也著實明白了。 《儒林外史》(1749)

wáng miǎn kàn shū, xīnxià yě zhǎo shí míng bái le

Wang Mian read book, mind also solid understand/see clear PERF (PERF = Perfective)

Wang Mian studied and began to see things clearly.

Hence, 明 and 白 originated from meanings that are relevant to light. In the later developments, both acquired other semantic contents. For 明 in particular, the semantic changes have been very diversified, covering a spectrum from the physical phenomena, namely, from LIGHT to the physiological ability of SEEING, then, to the cognitive capabilities of KNOWING and UNDERSTANDING. The relationship between LIGHT and SEEING can also be considered a PART AND WHOLE metonymical relation, because light is part of the conditions for a naked eye to be able to see. Consequently, the Chinese phrase 失明 ([shī míng], lose light) means losing one's eyesight or going blind, and 复明 ([fù míng], restore light) means regaining one's eyesight or being able to see again.

As such, the Chinese character 明 has therefore completed a remarkable semantic journey to become a word that eventually means UNDERSTANDING, only after a chain of metonymic extensions, as in the schema of (11):

(11) LIGHT → SEEING → UNDERSTANDING/KNOWING

In recent neural researches, it is found that although seeing and thinking are carried out by different parts of the brain, they also often interact intimately via feedforward and feedback interactions to give rise to conscious visual percepts (Carsetti 2004, p. 29). Such interactions form the neuron circuitry for the metaphor “THINKING IS SEEING,” and further “UNDERSTANDING IS SEEING,” which metaphorically maps our knowledge about vision onto the domain of understanding and knowing, causing words meaning SEEING to extend their meanings to UNDERSTANDING and KNOWING (cf. ICSI 2020). In fact, (11) represents a typical correlation-based metaphor, which emerges from frame-like mental representations through the metonymic stages whereby “one of the elements of a frame-like mental structure is generalized (schematized) to a concept that lies outside the initial frame in a different part of the conceptual system. The generalization process leads to sufficient conceptual distance between the initial and the new frame on which metaphors can be based” (Kövecses 2013). According to Evans and Green (2006, p. 211) and Evans (2007, p. 85), a frame or domain refers to a knowledge structure that is represented at the conceptual level and held in long-term memory relating. Words deleted and changed elements and entities associated with a particular culturally embedded situation from human experience. In essence, (11) can be argued to be a metonymic chain where the concept in one frame, i.e. LIGHT, is generalized to a new frame, i.e. SEEING, later from the frame for SEEING to those for UNDERSTANDING and KNOWING. This means that conceptually, several metonymic steps can lead to the achievement of a longer distance conceptual projection across frames or domains, which is the essential condition for making a conceptual metaphor.

The verb 明白 in Chinese is not a loner. *The Oxford English Dictionary Online (OED) Premium Collection* undoubtedly shows that the original meaning of the English word *clear,* the second most frequently found English equivalent of 明白, is also closely related to light, as data in Table 3 indicate (cf. "clear, adj., adv., and n.``. OED Online. June 2020. Oxford University Press. https://www-oed-com.rpa.library.ipm.edu.mo/view/Entry/34078?rskey=oUQYzO&result=1&isAdvanced=false (accessed August 07, 2020).

We can see that the semantic extension path of *clear* is similar to that of 明白 presented in (11).

The similar etymological paths of 明白 and *clear* show that because we humans possess the same physiological characteristics and functions endowed by our shared natural environments on the Earth, we have a broad common cognitive basis for our languages and thoughts. This constitutes the foundation of communication across

Table 3 Semantic contents and extensions of "clear"

Original or extended meanings	Literature sources	Examples
(1) Of light, color, things illuminated. a. *orig.* Expressing the vividness or intensity of light: Brightly shining, bright, brilliant	1297 *R. Gloucester's Chron.* (1724) 416	Ther come..a leme swythe cler & bryʒte
(2) Of the day, daylight, etc.: Fully light, bright; opposed to *dusk* or *twilight*. *arch*	*c*1320 *Sir Beues* 755	A morwe, whan hit was dai cler, Ariseþ kniʒt and squier
(3) Of a vision, conception, notion, view, memory, etc.: Distinct, unclouded, free from confusion	1398 J. Trevisa tr. Bartholomew de Glanville *De Proprietatibus Rerum* (1495) ii. v. 32	Bryghte and clere knowynge of god
(4) Of the faculty of discernment: That sees, discerns, or judges without confusion of ideas	1340 *Ayenbite* (1866) 24	Clier wyt, wel uor to understonde
(5) Of words, statements, explanations, meaning: Easy to understand, fully intelligible, free from obscurity of sense, perspicuous	*a*1400 (a1325) *Cursor Mundi* (Vesp.) l. 11,615	Þan com þe propheci al cler
(6) Of persons: Having a vivid or distinct impression or opinion; subjectively free from doubt; certain, convinced, confident, positive, determined	1604 S. Hieron *Preachers Plea* in *Wks.* (1620) I. 500	I am cleere in it, that many then in that darkness did..'See day at a very little hole'

languages and second language acquisition. The nature of our thoughts and the way we understand meaning in language are closely tied to our bodies when we feel and act in the world. This is the central thesis of the embodiment revolution (Bergen 2012, p.7; Feldman and Narayanan 2004; Gibbs 2005). In the cases of 明白 and *clear*, the environmental factor is LIGHT, which allows us to see, discern, and eventually understand. Once we can understand, we gain knowledge and intelligence, and we are able to, among others, appreciate (cf. Lakoff and Johnson 2010; Lakoff 2015). The data from the Chinese/English bilingual corpora simply lend us more compelling evidence to confirm the argument based on the embodiment hypothesis.

Having studied the metonymic extension of 明白 and *clear*, we come to the second question concerning the relationship between Chinese 明白 and its English equivalents. We want to know how this one-to-many relationship (i.e. 明白 vs. many English equivalents) in Chinese/English translation managed to emerge.

Notice that (11) ends at UNDERSTANDING, which, however, is not the last stage for the semantic extension of 明白. Rather, UNDERSTANDING has functioned as the rendezvous for various synonymous semantic units, lexical or phrasal, to converge on, and that allows 明白 to be associated with more lexical varieties in English. Each of them comes from its own origins and has preserved its own inherited connotations. The cognitive mechanisms that build the association include metonymy, the conceptual small mover, and metaphor, the larger resultant cross-domain conceptual projection that is, nevertheless, often initiated by the continuing metonymic pushes. One of the typical examples is UNDERSTANDING IS GRASPING. According to *OED*, since Old English, one of the semantic equivalents of UNDERSTANDING is COMPREHENDING, which in turn originated from words describing physical actions, such as *seize*, *grasp*, *lay hold*, and *catch*. In this metaphorical extension, the source domain is about the physical action of grasping an object, which is interpreted as an idea. Metonymically, therefore, if you can get hold of a notion, you may understand the idea in it. In this sense, the grasper is seen as the understander, and the objects grasped, as the ideas understood (Dancygier and Sweetser 2014, p. 28). Consequently, failing to grasp becomes failing to understand. With the concept of GRASP comes into the scene, we now have a semantic scenario as in Fig. 2:

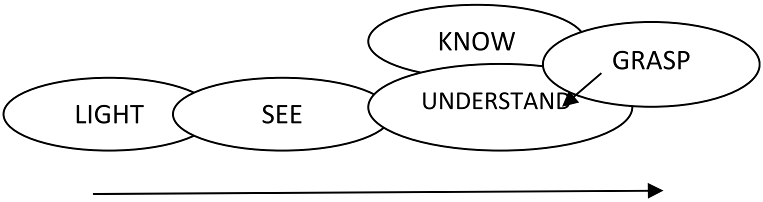


Fig. 2 Semantic rendezvous of LIGHT and GRASP at UNDERSTAND

In order to show that such a semantic rendezvous exists, we should be able to find examples in which the Chinese 明白 is equal to grasping or related meanings in English. From the data that we gleaned, this is exactly the case, as in

(12) 他不明白, 也不可能明白從他話裡單獨抽出來的字的意義。

tā bù míngbái, yě bù kěnéng míngbái cóng tā huàlǐ dāndú chōuchūlái de zì de yìyì

he not understand, also not possible understand/grasp from his word single extract out POSS word POSS meaning (POSS = Possessive)

He did not understand, and could not grasp the significance of words taken apart from the sentence.

(13) 對不起, 我沒明白你的意思。

dùibùqǐ, wǒ méi míngbái nǐde yìsī

sorry I not understand/catch your meaning

I'm sorry I didn't catch your meaning.

Notice that in (12), 明白 is rendered into *grasp*, and in (13), into *catch*, which suggests that a successful comprehender has to be a successful catcher of the meaning emitted from its giver in communication.

The concept GRASPING facilitates to open up a new semantic horizon for 明白, because further metonymic ramifications from GRASPING can bring about rich lexical possibilities that 明白 may hitherto not have had in the English translations. For instance, since an idea can be treated as an object that can be grasped, it then should be able to be given, as is often seen in the exchange of objects in human activities. In fact, the data carrying such meanings are also available from the *Chinese-English Sentence Aligned Bilingual Corpus*, such as

(14) 為了讓讀者明白其重要性, “八國集團”(G8)這個詞後面通常緊跟著世界上“領先”丶“最富有”丶“最大”或“最重要”經濟體的描述。

wèile ràng dúzhě míngbái qí zhòngyàoxìng,“ bā gúo jítuán”(g8) zhègè cí hòumiàn tōngcháng jǐngēn zhe shìjièshàng“ lǐngxiān” 丶 “ zùi fùyǒu” 丶 “ zùi dà” hùo“ zùi zhòngyào ” jīngjìtǐ de miáoshù

for allow reader understand/give reader sense its importance, G8 this word behind often tight follow PRO world advanced, richest, biggest, or most important economy body POSS description (PRO=Progressive, POSS=Possessive)

To give readers a sense of their importance, the words “Group of Eight” are usually followed by “leading”, “richest”, “largest” or “most important” economies in the world.

In (14), 明白 is rendered into “to give somebody something”.

The English *give* is a di-transitive verb, involving moving the possession of an object from person A to person B. It is therefore logical that this movement can be generated by the verbs or phrases that originally indicate physical forces causing objects to move, whereby an idea, now seen as an object, is moved into the possession of another human, whose head can function as the receiving container. In our data, this kind of movement is expressed by *hammer…into* and *strike*, as in

(15) 哈尼夫先生(巴基斯坦)(以英語發言): 主席先生, 我願感謝你作出不懈努力,並耐心地推遲作出決定的時間,你終於使我們明白,我們今年不可能有第三個議程項目。

hānífū xiānshēng (bājīsītǎn) (yǐ yīngyǔ fāyán) : zhǔxí xiānshēng , wǒ yuàn gǎnxiè nǐ zùochū bùxiè nǔlì , bìng nàixīndì tūichí zùochū juédìng de shíjiān , nǐ zhōngyú shǐ wǒmén míngbái , wǒmén jīnnián bù kěnéng yǒu dìsāngè yìchéng xiàngmù

Hanif mr. (Pakistan) (with English speak) chair mr. I would thank you make tireless effort, and patiently delay make decision POSS time, you finally make us understand/hammered into our head, we this year impossible have third agenda item (POSS = Possessive)

Mr. Hanif (Pakistan): I wish to thank you, Sir, for your tireless efforts and for your patience in delaying your decision, which you have finally hammered into our heads, that we cannot have a third item on the agenda this year.

Here, 明白 is rendered into *hammer something into something*. The verb *hammer* itself is a metonymic extension from the noun *hammer* through the OBJECT FOR ACTION metonymy. The force that causes the object to move originates from hitting the object with a hammer. In other words, the action of hammering or striking can be well triggered by the object hammer.

The example that is involved with *strike* is

(16) 然後, 她覺得自己必須把錢存進銀行以保安全, 這樣發展下來, 到了最後, 她終於明白了, 享受十全十美的生活的大門還沒有打開。

ránhòu, tā juéde zìjǐ bìxū bǎ qián cúnjìn yínháng yǐ bǎo ānquán, zhèyàng fāzhǎn xiàlái dào le zuìhòu, tā zhōngyú míngbái le, xiǎngshòu shíquánshíměi de shēnghuóde dàmén háiméiyǒu dǎkāi

but she feel self must BA money deposit into bank for safety, this develop downward, reach final, she finally understand/strike, enjoy perfect life POSS big door yet not open (BA = Ba Construction) (POSS = Possessive)

Then she found she must put her money in the bank for safety, and so moving, finally reached the place where it struck her that the door to life's perfect enjoyment was not open.

The result of a give-take transaction is that the taker becomes the possessor of an idea or a notion, which in English can be expressed by the verbal phrase *to have something*, as in

(17) 在聯邦儲備委員會(Fed) 1月份緊急降息和貝爾斯登(Bear Stearns)被迫嫁與他人的時候, 華爾街明白了政府不會聽任事情一敗塗地。

zài liánbāng chǔbèi wěiyuánhùi (Fed) 1 yuèfèn jǐnjí jiàngxī hé bèiersīdēng (bear stearns) bèipò jiàyǔ tārén de shíhòu, huáerjiē míngbái le zhèngfǔ bùhùi tīngrèn shìqíng yībàitúdì at Fed January emergent lower interest and Bear Stearns forced to marry other POSS time, Wall Street understand (have the notion) PERF government will not allow things get bad (POSS = Possessive, PERF = Perfective)

Along with the Federal Reserve's emergency interest-rate cut in January and the shotgun wedding of Bear Stearns, Wall Street has the notion the government won't let things get too, too bad.

Going along the pattern of having something, 明白 now expresses the idea of "having the notion."

In addition, as some objects can be manufactured by human efforts, ideas can also be made through similar causal actions. In this case, the making process facilitates the outcome of making, which metonymically allows such phrases as *make out* to mean understanding, as in

(18) "我總會和您的丈夫爭論; 我不明白, 他為什麼要去作戰。"皮埃爾向公爵夫人轉過身來毫無拘束地 (年輕男人對年輕女人交往中常有的這種拘束) 說道。

wǒ zǒng hùi hé nínde zhàngfū zhēnglùn; wǒ bù míngbái, tā wèishénme yào qù zuò zhàn。" píāiě xiàng gōngjué fūrén zhuàn gùo shēn lái háowú jūshù de (niánqīng nánrén dùi niánqīng nǚrén jiāowǎng zhōng cháng yǒu de zhèzhǒng jūshù) shūo dào

I often argue with your husband I not understand/make out, he why want fight, Pierre princess turn COMP body COMP no restriction ADV (young men to young women communication inside often have POSS restriction) say

(COMP=Completive, ADV=Adverbial)

"I'm still arguing with your husband; I can't make out why he wants to go to the war," said Pierre, addressing the princess without any of the affectation so common in the attitude of a young man to a young woman.

From the examples we have examined, we may construct the following semantic network with the converging rendezvous at UNDERSTAND, as in Fig. 3, which just shows only one part of the metonymic chain extensions found in the English equivalents of 明白.

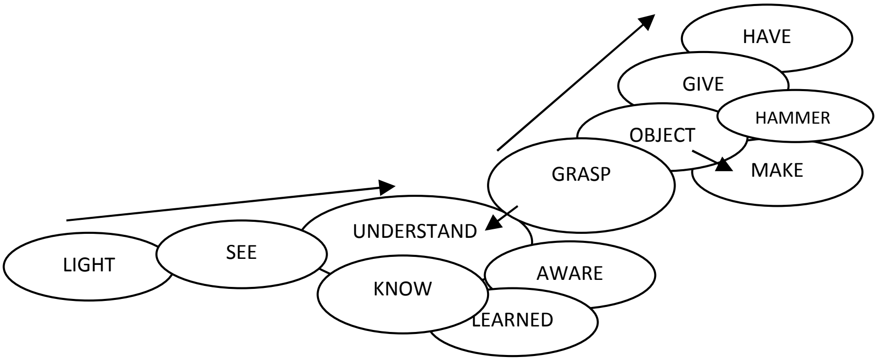


Fig. 3 Partial metonymic Chain extensions related to 明白

Hence, the semantic extension from 明 and 白 (LIGHT) to HAVE A SENSE has traversed a long distance and is punctuated by several directional changes in between. Nevertheless, the cognitive links connecting these notions do exist. These links are established by the driving force of metonym in human cognition, which is rightfully depicted by Radden and Kövecses (1999) as a cognitive process in which one conceptual entity provides mental access to another conceptual entity. In turn, accumulated metonymic extensions in a chain can make up a metaphorical projection across conceptual domains, which builds up the relationship between originally distant notions.

3.2 A Brief Account of the Story of "be Going to" and Its Chinese Equivalents

The lexical or grammatical meanings expressed by the Chinese equivalents of *be going to* can be classified into the following six categories. They include the following.

1. Physical Movement

The apparent movement senses are expressed, such as 去 ([qù], *go*), 到 ([dào], *go, arrive*), and 上 ([shàng], *ascend, go, attend*). The movement can be toward a concrete physical location or, more abstractly, to an event or occasion. In addition to the horizontal movement, the shift can be vertical in some Chinese idiomatic expressions. The following is an example:

(19) As chance would have it he was going to London as well and was able to give me a lift.

趕巧他也去倫敦, 所以能載我一程。(Location)

gǎnqiǎo tā yě qù lúndūn, suǒyǐ néng zǎi wǒ yī chéng
Coincidentally he also go London so able take me one leg

2. Intention + Movement

The combination of intention and movement senses, such as 要去 ([yàoqù] *want* + *go*), 打算去 ([dǎsuanqù], *plan* + *go*), and 会去 ([huìqù]: *intend* + *go*), 将 ([jiāng], *will* + v), as in

(20) "I'm going to town," she said.

"我要去城裡," 她說。

wǒ yào qù chéng lǐ tā shuō

I want go town inside she say

3. Stronger Intention Sense

Stronger intention senses, such as 要 ([yào], *want*]), 会 ([huì], *will*), 打算 ([dǎsuan], *plan*), *intend*], 准备 ([zhǔnbèi], *prepare*, *intend*), and 想 ([xiǎng]: *think*, *want*). In these situations, the English *be going to* is usually followed by a verb that does not indicate a movement and so are its Chinese counterparts, as in

(21) “Get down on your knees,” said the genie, “for I'm going to kill you.”

“跪下,”魔鬼說, “因為我要殺死你。”.
guì xià móguǐ shuō yīnwei wǒ yào shāsǐ nǐ
kneel down genie say because I want kill you

4. Futurity and Modality

Be going to indicating the grammatical senses of futurity and modality:

(22) A new subject is going to be given next week.

下星期將給一個新課題。 (Future)
xià xīngqī jiāng gěi yī gè xīn kètí
next week FUT give one CLF new subject
(FUT = Future or modal, CLF = Classifier)

(23) By all accounts, he is going to resign.

據說, 他將辭職。 (Possibility)
jùshuō tā jiāng cízhí
allege he FUT resign (FUT = Future or modal)

(24) It’s going to rain.

要下雨了。(Prediction)
yào xià yǔ le
FUT fall rain LE (FUT = Future or modal, LE = sentence final LE)

5. Combination of Modal/Future Grammatical Grams (Cf. Bybee et al. 1999, p. 2)

Combinations of the grams, such as 将会 ([jiānghuì], future gram + future gram] and 将要 ([jiāngyào], future gram + future gram) in the following:

(25) Some were going to be hanged in the next few days.

有些犯人將要在以後的幾天中被絞死。 (Future)

yǒuxiē fànrén jiāng yào zài yǐhòu de jītiān zhōng bèi jiǎosǐ

some prisoners FUT FUT at afterwards POSS several day inside PASS hang die

(FUT = Future, POSS = Possessive, PASS = Passive)

(26) It's going to rain tomorrow.

明天將要下雨。 (Prediction/Possibility)

míng tiān jiāng yào xià yǔ

tomorrow FUT FUT rain (FUT = Future or Modal)

6. Collocations of Temporal Adverbs with Grams

The collocations of temporal adverbs with the grams, as in 就会 ([jiùhuì], *right now* + future gram), and 肯定会 ([kěndìnghuì], *certainly* + future gram), as in

(27) I was going to pay the money back as soon as I saw you.

我一見到你就會還那筆錢的。 (Immediate future)

wǒ yī jiàndào nǐ jiù huì huán nà bǐ qián de

I once see you immediately FUT return that CLF money AFF

(FUT = future, CLF = classifier, AFF = affirmative)

(28) Milan is going to win the cup for sure.

米蘭隊肯定會贏得這個錦標賽。 (Strong prediction)

mǐlán duì kěndìng huì yíngdé zhègè jǐnbiāosài

Milan team certainly FUT win this tournament (FUT = Modal)

Basically, the following semantic continuum/metonymic chain connoted in the English phrase *be going to* are rendered overt by different Chinese translations, according to the contexts, as in

(29) Movement toward a goal → Movement + Intention → Future + Modality

However, most of the Chinese future and modality markers originated from volitional verbs, such as 要,想, 打算, which follow another path of grammaticization to change into future/modality markers:

(30) Volition or Desire → Intention → Future + Modality

It is at the **Intention, Future, Modality** (i.e. the semantic rendezvous) that the movement verbs meet with the volition verbs, and they become exchangeable in translation, as in Fig. 4.

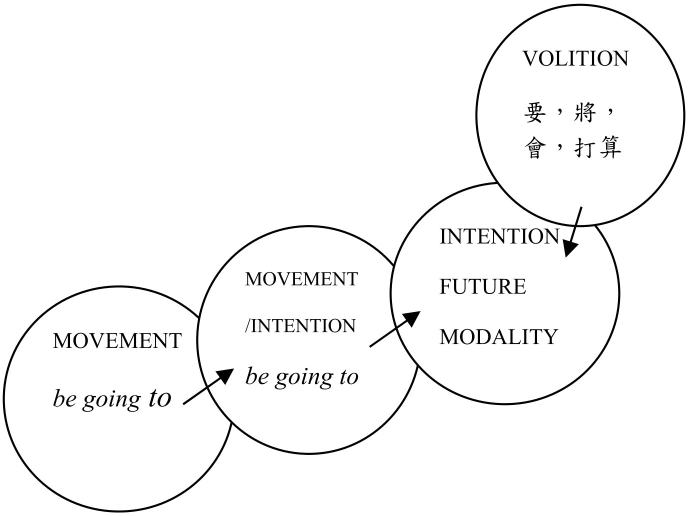


Fig. 4 Movement and Volition Meeting at Intention, Future, and Modality

That explains why we can find so many Chinese equivalents for *be going to*, as in Table 4, in which we have 21 different tokens that appear a total of 765 times in the bilingual corpus.

3.3 The Mathematical Model for the One-To-Many Relationship in English/Chinese Translation

The discovery of metonymic chains in the English/Chinese translation data and the unveiling of the one-to-many relations in the 明白 and *be going to* projects lead us to the inquiry about whether there could be a mathematical model that can capture the gist of the data. This is the third question we have posed.

Based on Table 4 and Table 1, respectively, the one-to-many relations in both projects constitute the two curves in Figs. 5 and 6.

Figures 5 and 6 bear considerable similarity to the typical Pareto Curves as presented in Fig. 7.

This suggests that Pareto Curves could probably approximate the “one-to-many” relationships we have been discussing.

The original Pareto principle, or the 20/80 rule, claims, among others, that about 20% of the population controls about 80% of the wealth. This principle is found extensively true in many social behaviors, such as the circulation of a library collection. It is often the case that only 20% of a collection could satisfy 80% of the library circulation needs. In terms of the Chinese future/modality grams used to indicate *be going to*, the 20/80 distribution can be arrived through the following

Table 4 Be going to and its Chinese equivalents

Serial number	Chinese equivalents of "be going to"	Frequency n	n/765 × 100% (%)
1	將 [jiāng: future gram]	136	18
2	要 [yào: to want, future gram]	130	17
3	去 [qù: to go]	119	16
4	會 [huì: be able to, future gram [possibility/ prediction]]	116	15
5	打算 [dǎsuan: to plan, to intend]	91	12
6	準備 [zhǔnbèi: to prepare, to intend]	23	3
7	到 [dào: to go, to arrive]	21	3
8	上 [shàng: to ascend, go, attend]	20	3
9	要去 [yàoqù: to want + to go]	19	2
10	想 [:xiǎng to think, to want]	18	2
11	將會 [jiānghuì: future gram + future gram]	12	2
12	將去 [jiāngqù: future gram + to go]	9	1
13	打算去 [dǎsuanqù: to intend + to go]	8	1
14	能 [néng: be capable of]	7	1
15	將要 [jiāngyào: future gram + future gram]	7	1
16	正要 [zhèngyào: just + future gram]	6	1
17	要到 [yàodào: future gram + to go]	5	1
18	快要 [kuàiyào: fast + future gram]	5	1
19	打算到 [dǎsuandào: to intend + to go]	4	1
20	正打算 [zhèngdǎsuan: just + to intend]	3	0
21	可以 [kěyǐ: can]	2	0
22	就會 [jiùhuì: right now + future gram]	2	0
23	想要 [xiǎngyào: to think + to want]	1	0
24	肯定會 [kěndìnghuì: certainly + future gram [possibility/prediction]]	1	0

calculation: 20% of the 24 Chinese equivalents translated from *be going to* is about 5, and these five are listed in Table 5.

These five Chinese equivalents constitute 78% (18% + 17% + 16% + 15% + 12%) of the total frequencies of 24 types, which shows what Fig. 4 presents is a Pareto curve in nature. Therefore, the human behavior in translating English *be going to* into Chinese is appropriately described by the 20/80 Pareto principle. Pedagogically, in our English/Chinese translation instruction regarding *be going to*, we may inform students that it is important to learn these five Chinese equivalents, because they represent 78% of the translation outputs from English. For a fuller discussion of the pedagogical issues related to *be going to* from the perspective of semantic similarity across languages, see Lin (2013).

For the 明白 data, the statistics is also close: 20% of the English equivalents is about 10, as presented in Table 6.

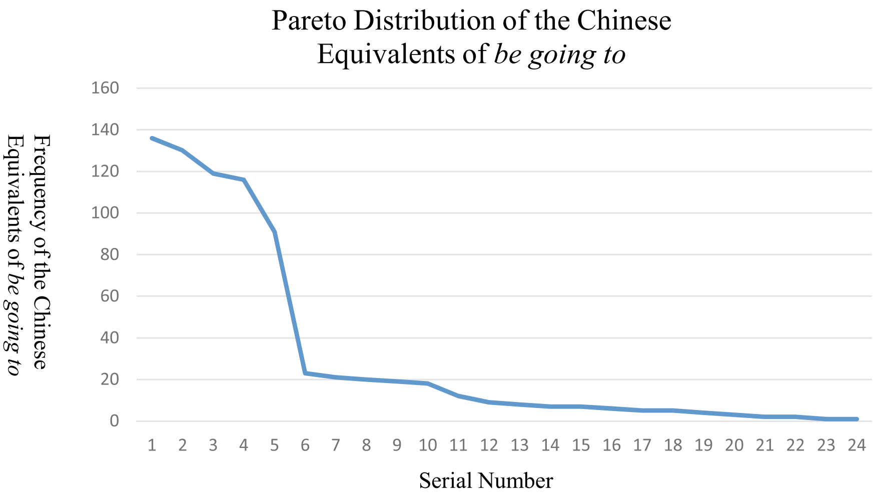


Fig. 5 Pareto Distribution of the Chinese Equivalents of *be going to*

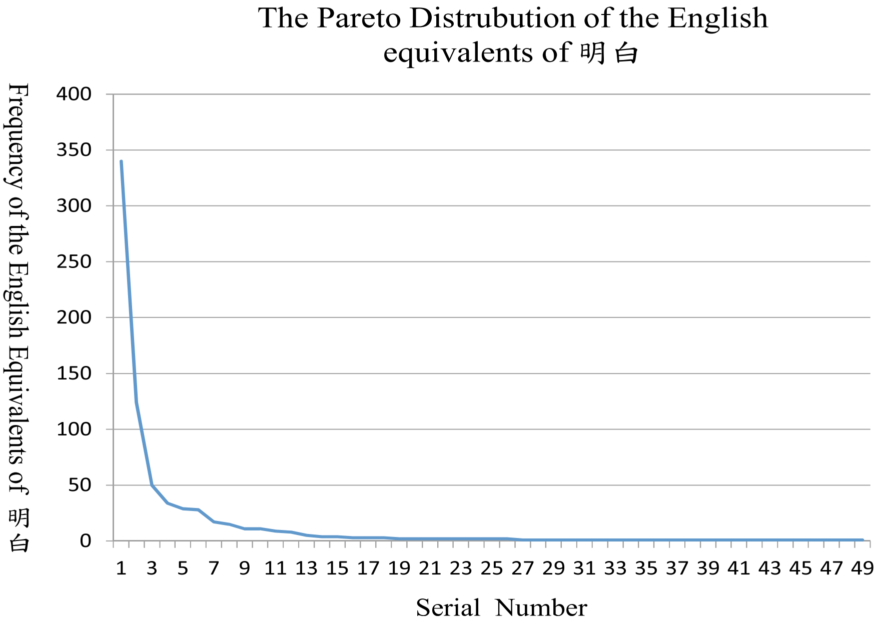


Fig. 6 Pareto Distribution of the English Equivalents of 明白

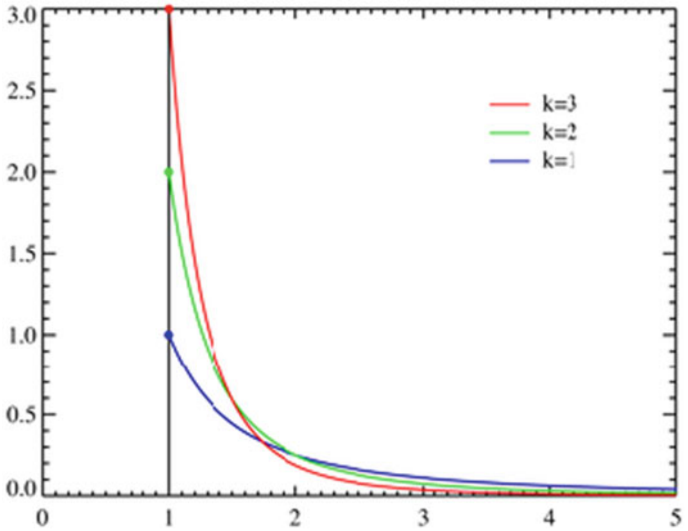


Fig. 7 Typical Pareto Curves

Table 5 20 of the Chinese Equivalents Translated from the English *be going to*

The Chinese Equivalents of *be going to*	Frequency	%
將 ([jiāng]: future gram)	136	18
要 ([yào], to want, future gram)	130	17
去 [qù: to go]	119	16
會 ([huì]: to be able to, future gram [possibility/prediction])	116	15
打算 ([dăsuan]: to plan, to intend)	91	12
Total %		78

Table 6 20% of the English Equivalents of 明白

The English Equivalents of 明白	Frequency	%
understand(ing)/understood	340	46
clear(ly)	124	17
know(n)	50	7
see/saw	34	5
express	29	4
plain	28	4
aware(ness)	17	2
appreciate/appreciation	15	2
learn	11	1
obvious	11	1
Total %		89

Namely, 10, or 20%, of the total English equivalents of 明白can cover up 89% of the translated results. The instructional value of this finding is also significant: our students can be advised to pay special attention to these ten English equivalents from among at least 49 options that have been found.

In short, it is clear that the Pareto principle is able to provide us with a mathematical approximation of the data we have examined. The translated equivalents with lower frequencies, nevertheless, are by no means insignificant. Since they often display a higher degree of rarity, they can add to the semantic, stylistic, and rhetorical richness of the translation.

3.4 English Translations of Wang Wei's 鹿柴 *[Lù Chái/zhài]: Small Data and the Metonymic Chain*

Metonymic chains in English/Chinese translation not only exist in the bilingual corpora of considerable sizes but also can be found from small data, such as the translation of a poem or even the translation of a single word. This can be seen in the various English translations of the Chinese character 柴 in 鹿柴, the title of a poem authored by Wang Wei.

Wang Wei (701–761) is a well-known landscape poet whose works are vivid with painterly visions and often imbued with rich Buddhist connotations. One of his landscape poems is 鹿柴 [lù chái/zhài], the Chinese original, *Pinyin*, and gloss are given in (31):

(31) 鹿 柴 [lù chái/zhài]

deer firewood
空 山 不 見 人 [kōng shān bú jiàn rén]
empty mountain not see person
但 聞 人 語 響 [dàn wén rén yǔ xiǎng]
only hear person speech sound
返 景 入 深 林 [fǎn yǐng rù shēn lín]
return scene enter deep forest
復 照 青 苔 上 [fù zhào qīng tāi shàng]
again illuminate green moss top

Two typical English translations of this poem are (32) by Egan (as cited in Cai 2012, p. 207) and (33) by Byanner and Kiang (as cited in Weinberger et al. 1987, p.10):

(32) The Deer Fence

On the empty mountain, no one is seen.
But the sound of voices is heard.

Returning: light enters the deep forest.
Again: it shines on the green moss.

(33) DEER-PARK HERMITAGE

There seems to be no one on the empty mountain….
And yet I think I hear a voice,
Where sunlight, entering a grove,
Shines back to me from the green moss.

In fact, Weinberger, Paz, and Wang (1987) list 19 different English translations of this famous poem. The diversified renderings found from such a one-to-many relationship provide us with the valuable clues to explore and examine the rationale of the drastic variations in Tang Poetry translation. For the sake of the discussion of metonymic chains in English/Chinese translation, we focus on the English translations of the title鹿柴. A more comprehensive study of the English translations of the whole poem and related research in translating Chinese poetic works can be found in Lin (2017, 2018).

While the Chinese character 鹿 unequivocally means deer in the title 鹿柴, our curiosity is with the English translations of 柴, which has been rendered into several different English terms, such as fence or park + hermitage. Now the question is whether there exists a semantic footing in the Chinese original that licenses multiple English translated versions.

According to the earliest Chinese dictionary 說文解字 (*Shuo Wen Jie Zi*: Explaining Characters) compiled (circa 100 -121 A.D.) by Xu Shen, 柴 means "small and scattered tree branches or twigs", or firewood. The metonymic operation involved in rendering "small and scattered tree branches or twigs" into a fence is the MATERIAL CONSTITUTING AN OBJECT FOR THE OBJECT metonymy, which is one of the PART FOR WHOLE relations. Specifically in this metonymy, because 柴 can be weaved into a fence, then the material 柴, namely, twigs, which make up the fence, is used to stand for the fence (cf. Wang, 2000, p. 483).

Interestingly, using 柴 ([chái]: small and scattered tree branches or twigs) to stand for "fence" has motivated this character to gain the other pronunciation [zhài], which is a homophone of the Chinese character 寨 [zhài], meaning "stockade" or "circumvallation". In Chinese philology, one of the traditions is that when two characters have the same pronunciation, one of them can often be borrowed to mean the other. In our case, [zhài] is the pronunciation borrowed from the stockade 寨 [zhài] for the twig 柴 [chái], due to the semantic extension from its original meaning "twig/firewood" to "fence".

The metonymic extension that started from the original meaning of 柴 does not stop at the fence. As given by OED, a fence is a structure made of wood or wire supported with posts that is put between two areas of land as a boundary, or around a garden/yard, field, etc. to keep animals in, or to keep people and animals out.

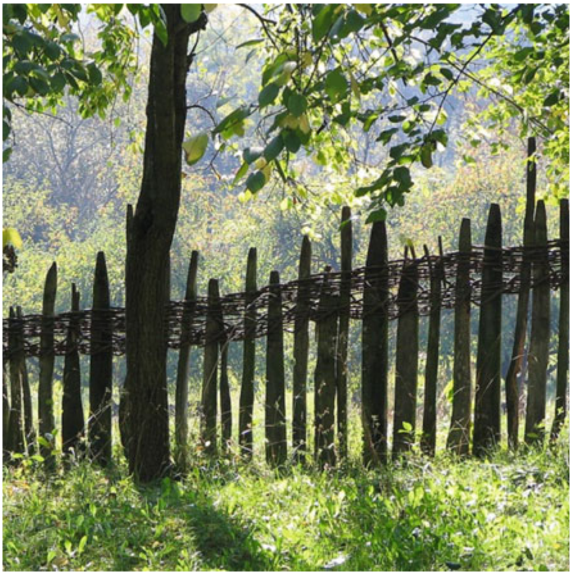


Fig. 8 A Fence Made of Twigs

Namely, a fence made of twigs can look like an object depicted in Fig. 8 (retrieved from https://commons.wikimedia.org/w/index.php?sort=relevance&search=Wattle+fence&title=Special:Search&profile=advanced&fulltext=1&advancedSearch-current=%7B%7D&ns0=1&ns6=1&ns12=1&ns14=1&ns100=1&ns106=1#/media/File:Wattle_fence,_West_Serbia.jpg).

One of the most salient physical properties of a fence is its enclosure that marks the boundaries, which, through metonymic association, facilitates the possibility to render 柴 into an enclosure in cognition and in translation.

In fact, "enclosure" is exactly the English translation by C.J. Chen and Michael Bullock (1960), Wai-lim Yip (1972) (as cited in Weinberger et al. (1987)), and Xu Yuanzhong (Xu et al. 1988, p. 87), where 鹿柴 is rendered into "The Deer Enclosure" and "Deer Enclosure", respectively. Admittedly, salience is very much a matter of the beholder's evaluation and therefore it is often subjective (Littlemore 2018, p.24).

Moreover, several translations rendered 柴 into "park", as in "The Deer Park" (H.C. Chang, 1977), "Deer Park" (G.W. Robinson, 1973), and "The Deer Park" (Soame, Jenyns, 1944) (as cited in Weinberger et al. (1987)), where the concept ENCLOUSURE again functions as the salient property facilitating such diction. OED tells us that a park refers to "any large enclosed piece of ground, usually comprising woodland and pasture, attached to or surrounding a manor, castle, country house, etc. and used for recreation, and often for keeping deer, cattle, or sheep", and "in extended use: [it is] an enclosed piece of ground for pasture or cultivation; a field, a paddock" ("park, n.'. OED Online. June 2020. Oxford

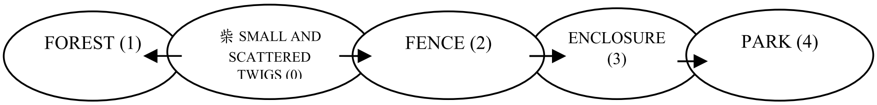


Fig. 9 The Metonymic Chain in Translating 柴 into English

University Press. https://www-oed-com.rpa.library.ipm.edu.mo/view/Entry/137946?rskey=RUiSMM&result=1&isAdvanced=false (accessed August 14, 2020)).

It is interesting to note that 柴 is also translated into "forest" as in "Deer Forest Hermitage" by Chang and Walmsley (1958) (as cited in Weinberger et al. (1987)). This is a PART FOR the WHOLE metonymy that renders 柴 back to its original source. After all, the firewood is a part, or a produce, of a forest. Therefore, from the data we have analyzed, we can construct a metonymic chain to describe and explain the four different English translations of 柴, as in Fig. 9.

where the following metonymic operations are at work:

(34) PART FOR WHOLE (0 to1)
MATERIAL FOR OBJECT (0 to 2)
SALIENT PROPERTY OF A CATEGORY FOR THE WHOLE CATEGORY (2 to 3), (3 to 4).

The English translations of 柴 demonstrate that even from small data, the metonymic chain can be pieced together, based on the work of different individuals. This provides the evidence that although each individual employs the metonymic operation in the way that he deems fit, when these operations are examined collectively, a chain that logically connects these metonymic operations can appear. Moreover, when the conceptual shifts driven by a chain of metonymic operations are lumped together, such as from step (0) to step (4) in (34), we find a metaphor is made as a result of the long-distance conceptual projection supported by the metonymic movements in between. Hence, translation data from multiple sources give us the opportunity to detect the cognitive processes in action during the translation process.

4 Conclusion

Metonymic chains are a powerful driving force in human thinking, as has been abundantly borne out by studies in semantic extension and grammaticalization, and now they are attested by the variations in English/Chinese translation outputs.

From a wider perspective, variations in Chinese/English translation can broaden the angle of frame semantics, which concerns exploration and establishment of the knowledge structure that is needed in the understanding of a particular word or related sets of words (Evans 2007, p. 192). In this study, in order to fully understand the semantic contents of 明白, *be going to*, and 柴 in Chinese/English translation, we have to study the knowledge structures around these words and phrases. The renditions in the target language actually substantiated the new knowledge structures woven together by metonymic chains, as in Figs. 3, 4, and 9. On the other hand, the semantic contents of many words and phrases are so volatile or protean that they can subtly shift meanings in different contexts of use (Evans 2009, p. xi). In a given context, a particular word or phrase often affords limited epistemic cues to multiple encyclopedic semantic frames. An epistemic clue refers to "any information that a hearer derives from his own knowledge and beliefs that then helps him determine the speaker's intended target" (Talmy 2018, p. 11). Take *be going to* as an example, it has afforded the movement frame, the futurity frame, and the modality frame in different contexts.

In Chinese/English translation, each segment of the metonymic chains can be considered as a building block of a newly composited knowledge structure, each carrying the semantic flavor brought out from the source frame the linguistic expression originally belongs to. For instance, "hammer …into" as an equivalent in translating 明白 in (15) has a flavor of the HAMMER AS A TOOL frame, which is obviously different from "have the notion" in (17), which are related to the POSSESSION frame and the CONCEPT frame. It is often in this manner that a translator would select his own portion of nuances that he deems appropriate in the translation process, which is usually considered contextually salient by the translator. When these "salient" portions are examined collectively on the basis of bilingual corpora, many phenomena hitherto unnoticed become overt. The emergence of the new knowledge structures metonymically chained together is one of them, which would not have been observable if data remain scattered or are only looked at individually. The discovery of metonymic chains from big data corpora sheds light on the examination of small data sets, where the validity of the chains still holds. Furthermore, with metonymic chains in place, many metaphorical projections in translation can be seen as having been caused by the metonymic conceptual movements and therefore can be better justified in translation practice. For instance, the semantic inference from TWIG to PARK in Fig. 9 is a long-distance conceptual projection between two distinct conceptual domains, a cognitive process that is typically characterized by conceptual metaphor. Without the metonymic chains in between, the projection would appear far-fetched. In short, the researches on 明白, *be going to*, and 柴 in this article are consolidated into a joint endeavor to discover and discuss these metonymic and metaphorical connections in the cognitive process of translation.

Pedagogically, the high-frequency tokens, such as *understand, clear* for 明白, and將 ([jiāng]: future gram) and 要 ([yào]: to want, future gram]) for *be going to*, should be first taught to simultaneous interpretation students who must gain the fastest access in their memory to the proper terms in the target language, not only because of its wider use and easier acceptance by the audience, but also because of the Conserving Effect (Bybee 2007, p. 10), which refers to the fact that repetition strengthens memory representations of linguistic forms, and makes them more accessible than lower frequency tokens. In simultaneous interpretation in particular, fast accessibility of a term matters. On the other hand, the low-frequency equivalents in the target language, such as *hammer something into somebody's head* for 明白, can be discussed later. Yet, low-frequency equivalents can offer more stylistic diversities and often carry extra exotic semantic nuances. In a small-data situation, such as the English translation of 柴, although rendering frequencies are statistically less significant, they can still be ranked as a matter of factual choices made by previous translators.

Translation studies should be constantly informed by researches in other fields so that the rationale behind the intricate linguistic phenomena produced in translation can become clearer. Like many other human behaviors, translation outputs cannot be a mass of random, chaotic choices. Rather, they are manipulated by the rule-governed cognitive hands often hidden behind an opaque screen. This had been the difficulty caused by the lack or scarcity of systematic data. With the increasing availability of large parallel corpora, the opacity is being lifted. Translation study scholars should and can find what these rules are, and show how they operate. With their own data and findings obtained from large-scale bilingual corpora, they can either endorse or question the novel conclusions or hypotheses proposed by the neighboring academic disciplines, such as cognitive linguistics and neurolinguistics.

References

Barcelona, A. 2002. Clarifying and applying the notions of metaphor and metonymy within Cognitive Linguistics: an update. In *Metaphor and metonymyin comparison and contrast*, ed. R. Dirven and R. Pörings, 207–277. (Cognitive Linguistics Research 20). Berlin/New York: Mouton de Gruyter.

Barcelona, A. 2005. The multilevel operation of metonymy in grammar and discourse, with particular attention to metonymic chains. In *Cognitive linguistics: Internal dynamics and interdisciplinary interaction*, ed. R.M. Ibanez, and F.J. Sandra Pena Cervel, 313–352. Berlin: Mouton de Gruyter.

Bergen, B.K. 2012. *Louder than words: The new science of how the mind makes meaning*. New York, NY: Basic Books.

Bredin, H. 1984. Metonymy. *Poetics Today* 5 (1): 45–58. https://doi.org/10.2307/1772425.

Bybee. J., R. Perkins, and W. Pagliuca. 1999. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago & London: The University of Chicago Press.

Bybee, J.L. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Cai, Z. 2012. *How to read Chinese poetry*. New York: Columbia University Press.

Carsetti, A. 2004. *Seeing, thinking and knowing: Meaning and self-organization in visual cognition and thought*. Dordrecht: Springer.

Dancygier, B., and E. Sweetser. 2014. *Figurative language*. Cambridge: Cambridge University Press.

Evans, V. 2007. *A glossary of cognitive linguistics*. Edinburgh: Edinburgh University Press.

Evans, V. 2009. *How words mean: Lexical concepts, cognitive models, and meaning construction*. Oxford: Oxford University Press.

Evans, V., and M. Green. 2006. *Cognitive linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Feldman J., and S. Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain Lang* 89 (2): 385–392. doi:https://doi.org/10.1016/S0093-934X(03)00355-9.

Gibbs, R.W. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge: Cambridge University Press.

Gibbs Jr., R. 2005. Embodiment in metaphorical imagination. In *Grounding cognition: The role of perception and action in memory, language, and thinking*, ed. D. Pecher, and R. Zwaan, 65–92. Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/CBO9780511499968.004.

ICSI [International Computer Science Institute, University of California, Berkeley]. 2020. *Metaphor: KNOWING IS SEEING* . https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:KNOWING_IS_SEEING.

Konblock, J. 1988. *Xunzi: A translation and study of the complete works*. Stanford: Stanford University Press.

Kövecses, Z. 2013. The metaphor–metonymy relationship: Correlation metaphors are based on metonymy. *Metaphor and Symbol* 28 (2): 75–88.

Lakoff, G., and M. Johnson. 2010. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books.

Lakoff, G. (Presenter). 2015, March 14. *How brains think: the embodiment hypothesis*. International Convention of Psychological Science. YouTube, https://www.youtube.com/watch?v=WuUnMCq-ARQ.

Lin, Z. 2013. Conceptual similarities in languages—Evidence from English *be going* to and its Chinese counterparts. In *Research in Chinese as a second language*, ed. I. Kecskés, 235–255. De Gruyter Mouton.

Lin, Z. 2017. Mental simulation and translation: An analysis of the cognitive motivations in the English translation of 天静沙 思秋. *Intercultural Communication Studies: ICS. XXVI* 2: 145–156.

Lin, Z. 2018. Variations in Tang poetry English translation—A cognitive rationale. In *Macao Language and culture research (2017)*, ed. H. I. Lei, 292–309. Macao: Macao Polytechnic Institute.

Littlemore, J. 2018. *Metonymy: Hidden shortcuts in language, thought and communication*. Cambridge: Cambridge University Press.

Radden, G., and Kövecses, Zoltán. 1999. Towards a theory of metonymy. *Metonymy in Language and Thought*. 17–59. https://doi.org/10.1075/hcp.4.03rad.

Talmy, L. 2018. *The Targeting system of language*. The MIT Press.

Wang, L. 王力. 2000. 王力古汉语字典*Wang Li gu hanyu zidian* [A dictionary of Classical Chinese by Wang Li]. Beijing: Zhonghua shu ju.

Weinberger, E., O. Paz, and W. Wang. 1987. Nineteen ways of looking at Wang Wei: How a Chinese poem is translated, 10. Kingston, R.I: Asphodel. Retrieved August 18, 2020 from http://www.jonvonkowallis.com/readers/CHIN5910/178-206-Eliot_Weinberger_&_Octavia_Paz-Nineteen_Ways_of_Looking_at_Wang_Wei.pdf.

Xu, Y., P. Lu, and J. Wu. 1988. *Tang shi san bai shou xin yi: Ying Han dui zhao = 300 Tang poems, a new translation : English-Chinese*. Beijing: Zhongguo dui wai fan yi chu ban gong si.

Zi-yu Lin is director of MPI-Bell Centre of English, and professor of the School of Languages and Translation, Macao Polytechnic Institute. He was a tenured associate professor at Seton Hall University, New Jersey, U.S.A. He earned his Ph.D. in linguistics from the State University of New York at Buffalo under the guidance of Prof. Joan L. Bybee.

Mind the Source Data! Translation Equivalents and Translation Stimuli from Parallel Corpora

Mikhail Mikhailov

Abstract Statements like ‘Word X of language A is translated with word Y of language B’ are incorrect, although they are quite common: words cannot be translated, as translation takes place on the level of sentences or higher. A better term for the correspondence between lexical items of source texts and their matches in target texts would be translation equivalence (Teq). In addition to Teq, there exists a reverse relation—translation stimulation (Tst), which is a correspondence between the lexical items of target texts and their matches (=stimuli) in source texts. Translation equivalents and translation stimuli must be studied separately and based on natural direct translations. It is not advisable to use pseudo-parallel texts, i.e. aligned pairs of translations from a ‘hub’ language, because such data do not reflect real translation processes. Both Teq and Tst are lexical functions, and they are not applicable to function words like prepositions, conjunctions, or particles, although it is technically possible to find Teq and Tst candidates for such words as well. The process of choosing function words when translating does not proceed in the same way as choosing lexical units: first, a relevant construction is chosen, and next, it is filled with relevant function words. In this chapter, the difference between Teq and Tst will be shown in examples from Russian–Finnish and Finnish–Russian parallel corpora. The use of Teq and Tst for translation studies and contrastive semantic research will be discussed, along with the importance of paying attention to the nature of the texts when analysing corpus findings.

Keywords Parallel corpora · Translation equivalents · Interlingual correspondences · Corpora in lexicography · Word alignment

M. Mikhailov (✉)
Languages Unit, Tampere University, Tampere, Finland
e-mail: mikhail.mikhailov@tuni.fi

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_10

1 Introduction

Electronic corpora are used nowadays in almost every field of linguistic research, and they are especially popular in lexicography (see e.g. Ooi 1998; Krishnamurthy 2008; Walter 2010; Hanks 2012; Kilgarriff 2013), at least when talking about monolingual corpora and projects involving only one language. In recent years, comparable and parallel corpora have also become one of the main sources of data in contrastive and translation studies. 'Translation is a source of perceived similarities across languages. Most linguists working in the field have either explicitly or implicitly made use of translation as a means of establishing cross-linguistic relationships' (Johansson 2007: 3). In spite of all this, multilingual corpora do not seem to be used on a large scale for compiling bilingual dictionaries; they remain for the time being only a secondary source of data if they are used at all. Why is this the case?

The possibilities of extracting bilingual lists of translation equivalents from parallel corpora have been discussed since the 1990s (Tiedemann 1997, Tiedemann 1998, Čmejrek and Cuřín 2001, Danielsson 2003, Kraif 2003, Garabík and Dimitrova 2015, Čermák 2019: 99–100). Many researchers consider parallel corpora a promising source of data for multilingual lexicography (Sinclair 2001, Teubert 2001, Kenning 2010, Kenny 2001, Štichauer and Čermák. 2016, Doval and Sánchez Nieto 2019, Zakharov and Bogdanova 2020). At the same time, one must admit that this resource presents far more challenges compared to using corpora for compiling monolingual dictionaries (Mikhailov and Cooper 2016: 149–154, Salkie 2008, Salkie 2002, Perdek 2012, Kubicka 2019, Tarp 2020), and therefore, comparable corpora are often considered a more realistic alternative (see e.g. Gamallo 2019).

The crucial problem of parallel corpora is that they are much smaller in size than monolingual corpora, and they will never be very large. While the TenTen corpora at Sketch Engine have passed the milestone of 10 G words, even the largest parallel corpora are only approaching the range of 1 G words for some common pairs of languages. Europarl, a parallel corpus of European Parliament debates, contains data in 21 languages of the EU, and it currently has the size of about 50 M tokens per language (Koehn 2005; Tiedemann 2012; https://opus.nlpl.eu/Europarl.php). The UN Parallel Corpus has about 500 M tokens per each of the six languages of the United Nations (ar, en, fr, es, ru, zh) (Ziemski et al. 2016). The ParaCrawl project is crawling parallel tests from the web and has succeeded in collecting data for over 40 language pairs. The largest ParaCrawl corpora are the French–English corpus, with over 1 G tokens, and German–English and Spanish–English corpora, which have close to 1 G tokens (Bañón et al. 2020).

The reason for the relatively modest sizes is that, although almost all types of texts are occasionally translated, only a limited number of genres are translated on a regular basis. These are news, technical instructions and user manuals, tourist brochures, political speeches, legal texts (remember that the famous Rosetta stone had a text of a decree by Ptolemy V inscribed in Ancient Egyptian and Ancient Greek as parallel texts), religious texts (e.g. the Bible) and fiction. Even these sources of data are not as inexhaustible as monolingual texts. Only a small proportion of fiction books is translated, and only documentation for imported products is translated. Likewise, only news from international news agencies is regularly translated. Many other text types—private letters, local news, financial documents, textbooks for schools—are not translated under normal circumstances, unless a special need arises (e.g. evidence for a trial at a court of law). Documents, contracts, agreements and the international letter exchange of state bodies and international companies are often translated, but most of these documents are not available to the general public. Thus, the amount of natural parallel texts is always incomparable to the amount of monolingual texts circulating in the community. For world languages and for languages with great numbers of speakers, the amount of parallel texts is much larger than for languages of lesser diffusion, and it is clear that for pairs of geographically distant minority languages (e.g. Gaelic-Irish and Kunama, Uyghur and Maltese) natural parallel texts are practically non-existent. Apart from the issue of the availability of the data, aligning parallel texts presents a serious technical challenge that slows down the whole process of compiling a parallel corpus. Large projects use fully automated aligning with some percentage of inevitable misalignments (see e.g. Koehn 2005, Bañón et al. 2020). Because of these issues, bilingual parallel corpora cannot be as large as monolingual corpora. Furthermore, parallel corpora are not available for every language pair, every text type and every topic.

Emilia Kubicka notes that ‘scholars dealing with translation studies have repeatedly pointed out the gap between traditional bilingual dictionaries and actual textual reality, and called for the creation of translation dictionaries which reflect the actual linguistic equivalents used by translators’ (Kubicka 2019: 75–76). At the same time, it is important to understand that a bilingual dictionary must supply equivalents for any word of any register, even if texts in which some of these words typically occur are seldom or never translated. Unfortunately, parallel corpora would not provide data for all words because of their limited size and restrictions in structure. For this reason, unlike monolingual corpora for monolingual lexicography, parallel corpora will never become a dominating source of data for multilingual lexicography. They will always be an additional resource, to be checked out using monolingual data.

At this point, a salient question arises. In some cases, we can suggest that a word *x* from a text in the language *A* has an equivalent *y* in our native language without consulting dictionaries or parallel corpora. How do we manage to do it? Obviously, we do not have an ‘internal parallel corpus’. What we might have in our brains are phrases in our native language that might be used in similar contexts or situations, i.e. a kind of ‘internal comparable corpus’. This means that comparable corpora have better perspectives as a source of interlingual equivalents compared to parallel corpora. Unlike parallel texts, comparable texts can be found for any text type and for almost any topic. However, comparable corpora cannot be aligned, and therefore, there is no straightforward way of searching for lexical correspondences. Although researchers actively develop methods of extracting interlingual equivalents from comparable corpora (Delpech 2014, Grabowski 2018, Terryn et al.

2020), such tools are not yet widely available. At the current state of technologies, comparable corpora are mostly used for reference purposes, e.g. to check out translation equivalents found in a parallel corpus or a dictionary.

In spite of its limited usability as a tool for the lexicographer, the parallel corpus can still be a very useful source of data for contrastive and typological studies. It is much more convincing to study authentic examples rather than the eternal *John killed Mary* or *The cat is on the mat* with their do-it-yourself translations into other languages. In his book, Stig Johansson shows multiple case studies from different areas of contrastive studies that benefit from the use of parallel corpora: times of the day, *love/hate*, *to spend time*, *to seem*, *well*, etc. A parallel corpus makes it possible to compare frequencies, and thus to detect translationese, to find equivalents used by translators and evaluate their popularity and usability (Johansson 2007). Authentic examples from published translations offer new opportunities for the development of this direction in linguistics, but like any research data, parallel texts require accuracy in use. One must keep in mind, however, that those 'naturally born' authentic examples, as opposed to artificial examples from the top of a linguist's head, do not appear in the texts for the sake of becoming an illustration of a certain linguistic phenomenon in a scholarly publication, but are instead a result of natural communication activities. The translator does not try to convey a meaning of repeated or interrupted action, the indefiniteness of the object, diminutives, etc. per se from the source text: the translator's mission is to transmit a message in another language.

Statements like 'Word *x* of language *A* is translated with word *y* of language *B*' are not quite correct from a linguistic perspective (a detailed explanation of this issue will be provided in the beginning of Sect. 2). In spite of this, we can sometimes read such statements in linguistic literature (see e.g. Ramón and Labrador 2008, Baños 2013, Dobrovol'skij and Pöppel 2016, Pöppel 2018, Zalizniak et al. 2018, Claire and El-Farahaty 2019). Of course, most of the authors use the term 'translation' as a shortened version of 'the item that appears as a representative of the word *x* when translating segments containing *x* into another language', and they understand the difference between translating and choosing a suitable lexical element when translating. Josep Marco uses three terms for this phenomenon: 'translation', 'translation solution' and 'translation correspondence'o (Marco 2019). In any case, the term 'translation' used for interlingual lexical correspondences is confusing. It downgrades the translation process to a mechanical substitution of elements where a parallel text is considered a set of pairs of matching sentences and not translations performed by a human with certain skills and training at a certain moment of time in a certain place and for a certain audience.

In this chapter, the interlingual lexical correspondences will be discussed from the viewpoint of the translation process. The following issues will be addressed:

- To what extent do translation equivalents from parallel corpora correlate with equivalents from bilingual dictionaries?
- How important is the direction of a parallel corpus for looking up translation equivalents?

- Do words of all grammatical classes have translation equivalents?

The data used in the study will be the Russian–Finnish and Finnish–Russian parallel corpora of fiction texts, ParRus and ParFin. Both corpora are composed of full texts and include works by different authors and translations by different translators. For some works, more than one translation is available. Works from different historical periods are included. Corpora of fiction texts represent language for general purposes, and these data are, therefore, suited to our study. ParRus and ParFin are different in size and are not identical in composition because of the natural asymmetry of literary translation activities in these two very different cultures. As a result, the two corpora do not form a bidirectional corpus, but they can still be used for comparing Russian–Finnish and Finnish–Russian data. More detailed information on the composition of ParRus and ParFin can be found in Mikhailov and Härme (2015) and Härme and Mikhailov (2016).

2 Translation Versus Translation Equivalent

The term ‘translation’ is overused in linguistic literature. This term often appears in contexts like ‘Word *x* is translated with the word *y*’ or ‘Word *x* is not translated’, etc. Strictly speaking, the expression ‘translation of the word *x* to language A’ is not correct, because translation is ‘conversion of writing or speech from one language to another’ (Danesi 2000, s.v. translation), i.e. only communicative-level units can be called translations, and the lowest appropriate unit would be an utterance. Kenny (2011) examines the concept of the translation unit from different points of view and shows that it is not connected to single words in the text, but rather at least to phrases or patterns. For intertextual interlingual matches of lower levels (word, grammatical form, morpheme), it is better to use other terms, for example, ‘translation correspondence’, ‘translation equivalent’, ‘lexical correspondence’, etc. (cf. Kraif 2002).

To study correspondences between source and target texts, two functions, Tr (translation) and Teq (translation equivalence), can be defined. To make the explanation more simple, fictional examples will be used.

Tr(m, sl, tl): translation Tr of the message *m* from the language *sl* to the language *tl*.

> Tr(‘John killed Mary’, en, ru) -> {‘Džon ubil Mèri’, ‘Džon pogubil Mèri’, ‘Džon zagubil Mèri’, ‘Džon – ubijca Mèri’, …}

Teq(u, sl, tl): translation equivalent Teq of the lexical unit *u* of the language *sl* in the language *tl*.

> Teq(‘John’, en, ru) -> {‘Džon’, ‘Ioann’, ‘Ivan’, …}

Obviously, Teq is a reoccurring lexical correspondence, and it does not cover all possible word alignments that can be discovered in parallel texts. Teqs should be more or less compatible semantically. For example, Russian words *on* 'he' or *čelovek* 'person' should not be included in the list of Russian Teqs of the English personal name *John*, although they might be used for translating messages containing the word *John*.

It is quite obvious to a linguist that when translating message *m* between languages *la* and *lb*:

$$\mathrm{Tr(m, la, lb)} \neq \mathrm{Tr(Tr(m, la, lb), lb, la)}$$

This means that the back translation of a message is not likely to reproduce the same message.[1] The Teq function is also irreversible, i.e.

$$\mathrm{Teq(u, la, lb)} \neq \mathrm{Teq(Teq(u, la, lb), lb, la)}$$

It is very important to understand that translations have a direction from source language to target language. Consequently, parallel corpora also have a direction: they can be uni- or bidirectional. If a corpus is bidirectional, it is necessary to define subcorpora including texts with required directions of translation.

In addition to 'natural' parallel texts, where original source texts are paired with their direct translations, there are indirect translations, where the translation is performed via a third language. This happens sometimes with translations of fiction when it is difficult to find a translator with the required pair of languages (or for other reasons). For example, all works by Chinghiz Aitmatov, a renowned Kyrgyz author of the Soviet period, were translated into Finnish from Russian, including his early works, which were originally written in the Kyrgyz language. In multilingual environments, it is possible to obtain pseudo-parallel texts, where both paired texts are translations from a third language. For example, most EU documents are available in all the official languages of the European Union, and it is, therefore, possible to obtain parallel texts for language pairs like Lithuanian and Greek, Maltese and Danish, etc. However, these parallel texts will be pseudo-parallel, because the texts are translated from another language, most likely, from English. It is obvious that in most cases, one should avoid using indirect translations and pseudo-parallel texts.

So, if Russian translation equivalents for Finnish words are to be found, direct translations from Finnish to Russian are required, not translations from Russian to Finnish. The latter will not yield Russian translation equivalents, but the Russian translation stimuli of Finnish words. (In everyday life, one can say *Your father is just like you*, but it is clear that this statement does not look quite natural). As for lexical correspondences acquired from pseudo-parallel texts or indirect translations; they cannot be interpreted in terms of the translation of this pair of languages. McEnery and Xiao note that the direction of translation is important for corpus-based contrastive studies (McEnery and Xiao 2007, 2010), and it is worth adding that it is equally important in lexicography.

[1] This is true even for machine translation: the result of back translation is often different from the initial source language message.

Let us take a simple example from our data. Finnish–Russian dictionaries register for the Finnish word *sauna* ‘bath’ two Russian Teqs, *sauna* and *banja*, while Russian–Finnish dictionaries suggest for the Russian word *banja* ‘bath’ only one Finnish Teq, *sauna*.

Teq(‘sauna’, fi, ru) -> {‘sauna’, ‘banja’}

Teq(‘banja’, ru, fi) -> {‘sauna’}

The first Russian Teq for *sauna* is a borrowing from Finnish. We can assume, therefore, that if we look up Russian translation equivalents for the Finnish word *sauna* in real-life translations from Finnish to Russian, we would find mostly examples with the word *sauna*, because it is a Finnish culturally-bound word and would be more appropriate for texts about Finland (as most texts in Finnish are expected to be). If we build a reverse parallel concordance for the Finnish word *sauna* in a Russian–Finnish corpus, we are likely to get both *sauna* ‘sauna’ and *banja* ‘Russian bath'. The word *banja* would be used as a general word for any bath or to refer to the Russian traditional bath, while the word *sauna* would refer only to the Finnish bath. For this reason, one can expect that the word *banja* would be more common than the word *sauna*.

This hypothesis was not, however, fully confirmed in authentic material: the parallel concordances from corpora of literary texts yield slightly different results (see Tables 1 and 2). In the Finnish–Russian corpus, the equivalent *banja* gets an unexpectedly high frequency, and only separate querying of two subcorpora—the ‘pre-war’ = ‘before 1945’ and ‘post-war’ = ‘after 1945’[2]—makes it clear that the Finnish borrowing *sauna* means in Russian a ‘modern’, ‘urban’, electrical Finnish bath, and therefore, in Russian translations of works by Aleksis Kivi, Juhani Aho and other classical authors of Finnish literature, the word *sauna* is rare and the equivalent *banja* is used instead. As for reverse concordancing in the Russian–Finnish corpus, the word *sauna* occurs on the Russian side only once, and it means ‘Finnish sauna’: all the other examples have *banja* ‘Russian bath’.

This example demonstrates that the direction of the corpus matters: a search in a corpus containing translations in both directions would yield unreliable results, a search in the wrong direction is likely to lead to wrong conclusions, and the use of indirect translations and pseudo-parallel texts would distort the picture even more. In the example with the Russian equivalents for the Finnish word *sauna*, a search in Russian–Finnish texts would give us an impression that *banja* is the only Russian equivalent for the Finnish word *sauna*, which would be incorrect, and only a carefully organised search in the Finnish–Russian corpus would show that there are two translation equivalents—*sauna* and *banja*—and the choice depends on the cultural context.

[2] In Finland, like in many other countries of Europe, the processes of urbanisation and industrialisation accelerated after the end of World War II, and the whole way of living changed.

[3] The sign ∞ is used to mean ‘other equivalents’.

Table 1 Matches for the Finnish word *sauna* in the Finnish–Russian parallel corpus

Matches	Before 1945	After 1945	Total result
banja	139	67	**206**
sauna	12	140	**152**
∞[31]	18	9	**27**
Total result	169	216	385

Table 2 Matches for the Finnish word *sauna* in the Russian–Finnish parallel corpus (reverse concordancing)

Matches	F
banja	242
sauna	1
∞	9
Total result	252

3 Translation Equivalent Versus Translation Stimulus

The example from the previous section demonstrates that a reverse parallel concordance is not the same thing as a parallel concordance. A reverse parallel concordance does not tell us about translation equivalents, but about the language units of the source text that provoke the use of certain units in translation. Let us call this dependence **translational stimulus**. Translational stimulus Tst (u, sl, tl) is a function, the reverse to the function Teq. It is obvious that

$$\mathrm{Teq}(\mathrm{w}, \mathrm{la}, \mathrm{lb}) \neq \mathrm{Tst}(\mathrm{w}, \mathrm{la}, \mathrm{lb}),$$

although the resulting sets usually do have an overlap. This was just demonstrated in the example with the word *sauna*.

In order to have a closer look, let us take a more complex example—the Finnish Teq for the Russian word *volosy* ‘hair’. This time, the concordances are much longer: over 900 examples in the Russian–Finnish corpus and over 600 in the Finnish–Russian one. Fortunately, it is not necessary to read all the examples and mark equivalents manually. Smaller concordances can be handled in Excel by means of applying filters to a table and group annotation. Very large tables can be processed in R by running relatively simple scripts that match examples for substrings and assign relevant equivalents to each example.

After checking these two large parallel concordances, we have Tables 3 and 4. Surprisingly, the lists of Finnish correspondences and their rank places coincide in both tables, although the normalised frequencies (ipm = instances per million tokens) vary substantially.

In the Teq list, the first equivalent, *hiukset*, outmatches all the remaining candidates, while in the Tst list, the second stimulus, *tukka*, closely follows the first

[4] The sign ∅ means the omission of the unit in the corresponding segment.

Table 3 Finnish Teq for the Russian word *volosy* ‘hair’ (Russian–Finnish corpus)

Fi	F	ipm
hiukset ‘hair’	578	161.20
tukka ‘hair’	195	54.38
karvat ‘bristle’	35	9.76
hius(-) ‘hair’	33	9.20
kihara ‘curly’	18	5.02
pää ‘head’	12	3.35
jouhi ‘horsehair’	5	1.39
∞	19	5.30
ø[41]	19	5.30
Total Result	911	254.08

Table 4 Finnish Tst for the Russian word *volosy* ‘hair’ (Finnish–Russian corpus)

Fi	F	ipm
hiukset ‘hair’	314	201.74
tukka ‘hair’	234	150.34
karvat ‘bristle’	22	14.13
hius(-) ‘hair’	20	12.85
kihara ‘curly’	14	8.99
pää ‘head’	4	2.57
jouhi ‘horsehair’	2	1.28
∞	8	5.14
ø	21	13.49
Total Result	639	410.54

equivalent. The phenomenon can be explained by the interference of the source language during translation in the Russian–Finnish data. Obviously, the Finnish translators subconsciously choose for the Russian pluralia tantum *volosy* a Finnish pluralia tantum *hiukset*, although there is another equivalent, *tukka*, which is as good, but is a singularia tantum. This case shows that if only the Teq is checked, one can possibly overlook a good suggestion. Still, it would not be a good idea to mix the two sets of data.

More substantial differences between Teq and Tst can be seen after analysing parallel concordances for the Russian verbs *pokupat'*and *kupit',* 'to buy'. The two verbs make an aspect pair[5]: the first verb is imperfective and has the meaning of a habitual, incomplete and repeated action of buying, while the second is a perfective verb and has the meaning of a completed action. The aspectual differences are not only grammatical, but also semantic, which results in the use of different translation equivalents, as can be seen in Tables 5 and 6. The Tst list is shorter, and the difference in frequencies is visible to the naked eye.

Again, we have to admit that the Tsts from the reverse concordances give some idea about lexical correspondences in the languages in question. As in the previous example with the noun *volosy* 'hair', some interference with the Russian originals can be noticed: among the Finnish equivalents for the Russian perfective verb *pokupat',* the second place is occupied by the Finnish verb *ostella* 'to shop' with quite a high frequency. This verb has the additional semantics of recurring action and is more frequent in Russian translations than in non-translated Finnish, e.g. in the fiTenTen2014 corpus hosted at Sketch Engine—it has a frequency of 4.28 ipm. The list of Tsts for these verbs (Table 5) does not contain *ostella*. This list, however, provides us with two good suggestions that are not in the Teq list: *hankkia* 'obtain' and *saada* 'get'.

It is important to understand that Tsts do not reflect the real translation processes. However, unlike Teqs, Tsts are not subject to interference and can help to eliminate such lexemes. Jurkiewicz-Rohrbacher distinguishes between translation equivalents, which work only in the direction of translation, and functional equivalents, which work both ways (Jurkiewicz-Rohrbacher 2019: 110–111). In our case, comparing Teqs and Tsts does not produce inverse correspondences, but helps to filter out the equivalents that are influenced by the source language. Tsts would, therefore, be useful for contrastive and typological studies. Nevertheless, the researcher should understand the difference between Teqs and Tsts, look up Teqs and Tsts separately, and purposefully use Tsts to detect asymmetry in the lexical systems of the two languages.

[5] Russian verbs belong to one of two aspects: the perfective (which sees the situation as a single whole (Comrie 1976: 16)) or the imperfective (which refers to general facts, or to continuing or repeated events). Perfective verbs have two tense forms: the past and future simple. Imperfective verbs have three tense forms: the past, present, and future complex (which is formed with the auxiliary byt' 'to be' + infinitive). Gerunds of perfective verbs are in the past tense, while gerunds of imperfective verbs are in the present tense. Perfective verbs can only form past participles, while imperfective verbs form both present and past participles. The Russian language does not have a perfect aspect (which should not be confused with the Russian perfective). Verbs with close meaning belonging to different aspects form so-called aspect pairs. These paired verbs can replace each other in different contexts. Still, they are different lexemes, not forms of the same word. For details, see (RG 1980: pp. 1384–1387, 1490–1498).

Table 5 Finnish Teq for *pokupat' / kupit'*(Russian-Finnish corpus)

kupit'			pokupat'		
Fi	F	ipm	Fi	F	ipm
ostaa 'to buy'	657	183.24	ostaa 'to buy'	179	49.92
hankkia 'to obtain'	7	1.95	ostella 'to do shopping'	24	6.69
lahjoa 'to bribe'	6	1.67	kauppa 'store, N'	3	0.84
saada 'to get'	6	1.67	∞	5	1.39
maksaa 'to pay'	4	1.12	∅	2	0.56
ostella 'shop, N'	3	0.84			
∞	12	3.35			
∅	8	2.23			
Total Result	701	195.51	Total Result	213	59.41

Table 6 Finnish Tst for *pokupat' / kupit'*(Finnish–Russian corpus)

kupit'			pokupat'		
Fi	F	ipm	Fi	F	ipm
ostaa 'to buy'	402	258.27	ostaa 'to buy'	134	86.09
hankkia 'to obtain'	41	26.34	hankkia 'to obtain'	5	3.21
saada 'to get'	15	9.64	hakea 'to seek'	3	1.93
hakea 'to seek'	6	3.85	∞	6	3.85
ottaa 'to take'	4	2.57	∅	7	4.50
∞	18	11.56			
∅	16	10.28			
Total Result	503	323.16	Total Result	155	99.58

4 Does Any Word Have Translation Equivalents?

When talking about translation equivalents, it is also important to understand whether all lexemes can have translation equivalents. In corpus linguistics, aligning parallel texts at the word level, so-called word alignment, is practiced (Tiedemann 2004; Östling and Tiedemann 2016). The purpose of such alignment is to find the maximum number of matches between the words of aligned sentences. The starting point of the algorithm is an assumption of the presence of a potential match for any token.

Let us illustrate word alignment in a simple Russian sentence, *Ja čitaju knigu s babuškoj*, and its English and Finnish translations, *I am reading a book with grandma* and *Luen kirjaa mummon kanssa* (see Figs. 1 and 2).

It is clear even from these simple examples that some tokens of the source sentence have no correspondence in the translations and some may correspond to more than one token in the target text. Even for the tokens that can be aligned, there

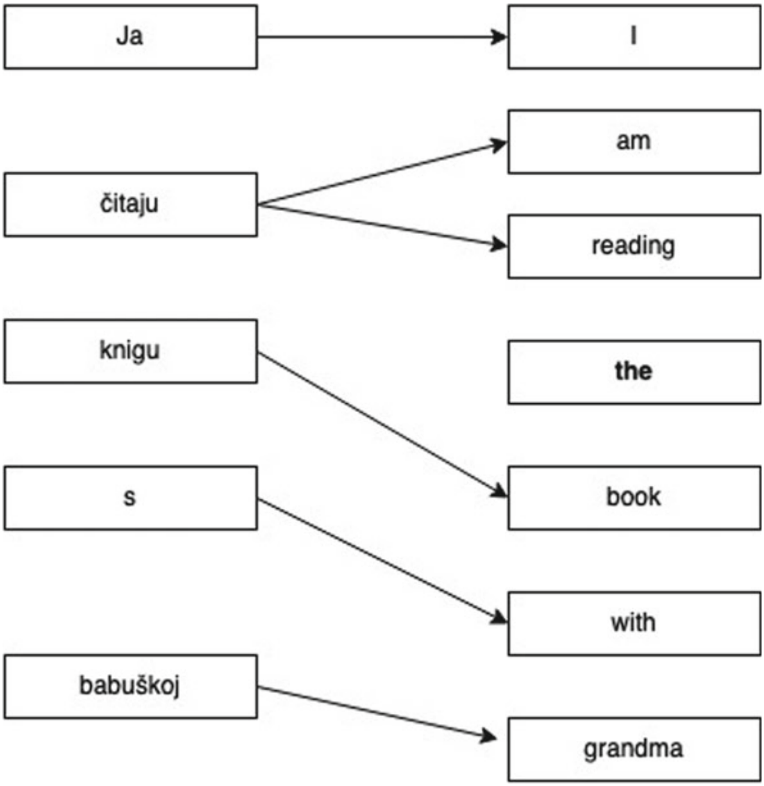


Fig. 1 Word alignment: A Russian-English example

are doubts whether they are indeed ‘translated’ and whether ‘translation equivalent’ would be the correct term here. Are the tokens *with* and *kanssa* Teq for the Russian preposition *s* ‘with’? As we know, the choice of preposition often depends on the noun, cf. ru *Petr v škole* - > *Petr is at school* and *Petr v komnate* - > *Petr is in the room*, where the Russian preposition *v* ‘in’ corresponds with the English preposition *at* in the first sentence and *in* in the second sentence.

To check whether translation equivalence and translation stimulation are applicable for function words, I looked up the Finnish correspondences for the Russian conjunction *hotja* ‘although’ in the Russian–Finnish corpus. This time, the search was performed on the texts starting from the middle of the twentieth century. The results of the search can be found in Table 7.

The reverse search for translation stimuli in the Finnish–Russian corpus provides a very similar list of correspondences (Table 8). Interestingly, the conjunction *hotja* is much more frequent in translations into Russian than in original Russian texts; the difference in relative frequencies is almost triple. The frequencies of Tsts descend more smoothly than the frequencies of Teqs, where *vaikka* ‘although’ clearly dominates. From the statistics in Table 7, we can see that the conjunction *vaikka*

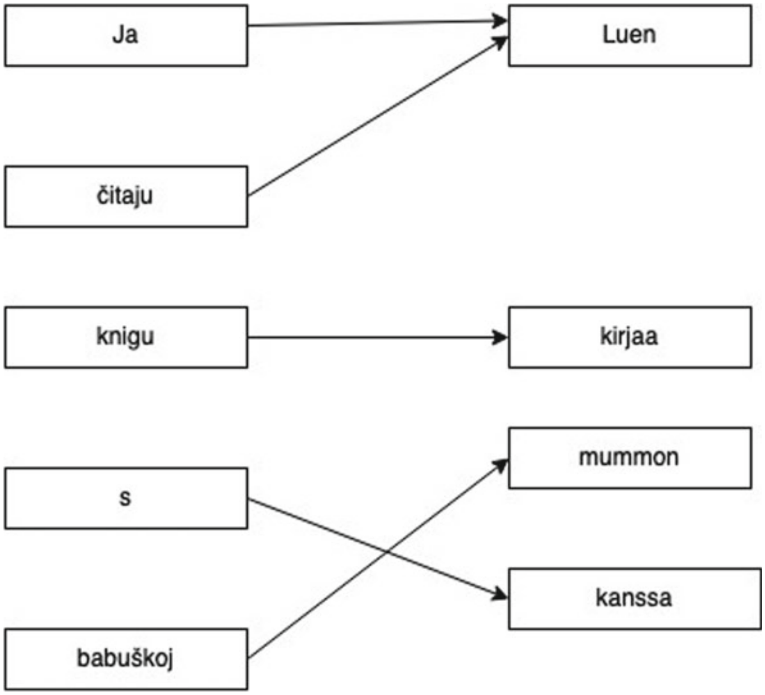


Fig. 2 Word alignment: A Russian–Finnish example

Table 7 Finnish correspondences for the word *hotja* 'although' (Russian–Finnish data)

Teq	F	ipm
vaikka 'although'	510	336.14
edes 'even'	49	32.3
tosin 'indeed'	26	17.14
ainakin 'at least'	24	15.82
mutta 'but'	21	13.84
joskin 'although if'	13	8.57
huolimatta 'in spite of'	6	3.95
vaan 'though'	6	3.95
kuitenkin 'still'	5	3.3
paitsi 'except'	1	0.66
∞	42	27.68
Total number of examples	703	463.35

'although' is the absolute favourite: 71% of the contexts are translated into Finnish using this conjunction, and this corresponds with the recommendations of the Russian–Finnish dictionaries. The Finnish–Russian data (Table 8) also have *vaikka*

Table 8 Finnish correspondences for the word *hotja* ‘although’ (Finnish–Russian data)

Tst	F	ipm
vaikka ‘although’	1003	850.82
mutta ‘but’	119	100.95
edes ‘even’	55	46.66
ainakin ‘at least’	46	39.02
kuitenkin ‘still’	31	26.3
vaan ‘though’	19	16.12
tosin ‘indeed’	13	11.03
huolimatta ‘in spite of’	8	6.79
joskin ‘although if’	7	5.94
paitsi ‘except’	7	5.94
∞	165	139.97
Total number of examples	1473	1249.51

as the main correspondence for *hotja* with 68% of all examples. However, in this data *mutta* ‘but’, *edes* ‘even’, *ainakin* ‘at least’, and *kuitenkin* ‘still’ are more visible and have much higher frequencies than in Table 7.

The remaining part of the list contrasts the Teq statistics for the content words in the previous section: many of the matches are not only unlikely to appear in bilingual dictionaries, but are not even conjunctions.

To get a better understanding of what is going on, let us have a look at few examples:

(1) К чему этот насмешливый тон? При чем тут “наследники”? **Хотя** жена действительно … (Пастернак Б.Л., Доктор Живаго) (‘What is this mocking tone for? What do the ‘heirs’ have to do with this? **Although** the wife indeed…’)

Miksi tuollainen pilkallinen sävy? Mitä tekemistä tässä on perillisillä? **Tosin** vaimo todellakin … (transl. J. Konkka.) (‘Why such a mocking tone? What do the ‘heirs’ have to do with this? **Really** the wife indeed…’)

(2) Вы **хотя** бы отдаленно представляете себе, о чем говорите? (Маринина А., За все надо платить)

(‘Do you understand **at least** approximately, what you are talking about?’)

Onko teillä harmainta**kaan** käsitystä siitä mitä te puhutte? (transl. O. Kuukasjärvi)

(‘Do you have **any** slight idea of what you are talking about?’)

(3) Он все-таки **хотя** и очень милый, но странный. (Улицкая Л., Сквозная линия)

(‘**Although** he is nice, still he is strange’)

Kaikesta rakastettavuudestaan **huolimatta** hän oli kovin omituinen mies. (transl. A. Pikkupeura)

(‘**In spite of** all his loveability, he is a very strange man’)

Table 9 Finnish correspondences for the word *nu* 'well, so' (Russian–Finnish data)

Teq	F	ipm
no 'well'	1742	1148.16
niin 'so'	155	102.16
mutta 'but'	95	62.62
entä 'and'	71	46.8
ja 'and'	71	46.8
nyt 'now'	67	44.16
mikä/mitä 'what'	52	34.27
hyvä 'good'	48	31.64
voi 'oh'	41	27.02
sitten 'than'	35	23.07
kyllä 'yes'	25	16.48
vaikka 'although'	23	15.16
jo 'already'	21	13.84
siinä 'there'	20	13.18
oikein 'really'	16	10.55
vain 'only'	15	9.89
hei 'hi'	12	7.91
totta 'true'	12	7.91
ihan 'really'	5	3.3
∞	213	140.39
Total number of examples	2739	1805.29

In example (1), the structure of the translation is more or less similar to that of the source text, but in examples (2) and (3), the translators changed the syntax and the correspondences for *hotja* are not easy to find.

We get an even more contradictory picture for the Finnish correspondences of the Russian particle *nu* 'well, so' (Table 9).

The length of the list speaks for itself, as it demonstrates that there are no exact correspondences (cf. Salkie 2002) for the Russian particle *nu* in Finnish texts. The dominating *no* 'well' covers only about 30% of cases, and it is mainly used when translating sentences with *nu* in the initial position. The remaining Teq are all so different that it is even hard to imagine how all these Finnish words could correspond to the same Russian word.

The inverse parallel concordance from the Finnish–Russian data quite expectedly also yields a long vague list of correspondences (see Table 10). It is worth noting that this time particle *nu* is much more frequent in the texts originally written in Russian.

Checking some contexts with *nu* from the Russian–Finnish data again demonstrates changes in the syntax of the translations.

Table 10 Finnish correspondences for the word *nu* ‘well, so’ (Finnish–Russian data)

Tst	F	ipm
no ‘well’	668	566.65
niin ‘so’	84	71.26
mutta ‘but’	45	38.17
nyt ‘now’	42	35.63
ja ‘and’	35	29.69
sitten ‘than’	30	25.45
voi ‘oh’	28	23.75
entä ‘and’	27	22.9
mikä/mitä ‘what’	27	22.9
jo ‘already’	25	21.21
kyllä ‘yes’	21	17.81
hyvä ‘good’	15	12.72
vaikka ‘although’	12	10.18
siinä ‘there’	8	6.79
ihan ‘really’	7	5.94
hei ‘hi’	5	4.24
vain ‘only’	5	4.24
totta ‘true’	4	3.39
oikein ‘really’	3	2.54
∞	109	92.46
Total number of examples	1200	1017.93

(4) **Ну** да где тут думать, поезд-то уж близко, думать некогда. (Пастернак Б.Л., Доктор Живаго)
(‘**So** when would you think, the train is already close, no time to think’)
Vaikka eihän siinä ollut ajattelemisen aikaa, juna oli jo lähellä. (transl. Juhani Konkka)
(‘**Anyway** there was no time for thinking, the train was already close’)

(5) **Ну**, скажем, в театр? (Булгаков М.А., Театральный роман)
(‘**Well**, for example to a theatre?’)
Sanotaan **nyt** vaikka teatteriin? (transl. Esa Adrian)
(‘Shall one say **now** for example to a theatre?’)

(6) Дядя Толя книжку принес старинную. Называется “Заветные сказки‶. Старинные сказки русские, необработанные. Там такие тексты, **ну** точно как бабушка выдает. (П. Санаев. Похороните меня за плинтусом)
(‘Uncle Tolja has brought a book, an old one. It is called ‘The Secret Tales’. Old Russian fairy tales, unabridged. There are such texts there, **well**, exactly like those grandma does.’)
Tolja-setä toi ikivanhan kirjan. Sen nimi on Perinnesatuja. Siinä on vanhoja venäläisiä satuja, muokkaamattomia. Siellä on sellaisia tekstejä, **ihan** niin kuin mummo pudottelee. (transl. Kirsti Era)

> (‘Uncle Tolja brought a very old book. It is called Traditional tales. There are old Russian tales there, unchanged. There are such texts there, **well**, exactly like grandma gives out'.)

The explanation is simple: *nu* is a discourse word, and as such it does not even have its own meaning but is rather used to underline or emphasise certain elements of the utterance where it is used and for linking the current sentence to previous sentences. Such marker words function in different languages in very different ways, and there is no direct correspondence between them. There might be many different ways to map the message of an utterance of the source into an utterance of the target text.

When searching for Teqs for cohesion words, one often has to act by the method of exclusion, that is, to start with determining Teqs for content words—nouns, verbs, adjectives, and adverbs—and only at the next stage try to find matches for the remaining tokens (cf. automated word aligning techniques; see e.g. Tiedemann 2004). In fact, these words are not dictated by the tokens of the source text, but rather by the syntactic constructions and communicative functions of utterances. Therefore, establishing links with the source text is just a convention; the translator hardly cares about expressing the concrete lexemes like *nu* or *hotja* in translation, although he/she is likely taking pains to express the meanings of uncertainty or concession that are present in the utterance to translate.

To sum it up, although Teq and Tst searches for a function word might return some frequently reoccurring matches, as it happened in the cases above, the findings are not very helpful for practical use as opposed to Teq and Tst searches of content words: nouns, verbs, adjectives, and adverbs.

5 Conclusions

The examples given in this chapter demonstrate that findings from parallel corpora are not identical to equivalents registered in bilingual dictionaries. Parallel corpora may suggest good solutions not listed in dictionaries, and it is possible to check which equivalents are most frequently used for translating. At the same time, parallel corpora sometimes demonstrate the influence of dictionaries on translators and in this way form a vicious circle (cf. e.g. Perdek 2012, Mikhailov 2020). Despite these reservations, the community has already noticed the usefulness of these data and many lexicographical services—GlosBe, Linguee, and the like—provide in addition to dictionary entries concordances from parallel corpora.

The two reverse functions—Teq (translation equivalent) and Tst (translational stimulus)—that were introduced in this chapter give a better understanding of lexical correspondences in parallel texts. Only the former reflect real translation processes, as the other is a posteriori link leading backwards from the target to the source text. Nevertheless, it can be useful for checking out natural translation equivalents and detecting those that are ‘infected’ with source language interference.

The adequate direction of translation and the exclusion of pseudo-parallel texts play an important role in all cases. Only the correctly chosen data will provide correct results that have theoretical and practical value. One might say that this has nothing to do with specialist texts that are dealing with technical, economic, or legal issues: special terms are the same in any language. This is not quite true. Different languages have different traditions in terminological issues as well, which might result in multiple interlingual correspondences and substantial differences in frequencies depending on the direction of translation. It is probable that ignoring the direction of translation in the data used for developing MT systems might affect the quality of translation.

The examples given in the chapter show that the functions Teq and Tst work only with content words, i.e. nouns, verbs, adjectives and adverbs. For these word classes, one can get useful information on interlingual correspondence for lexemes.

Cohesion words (conjunctions, prepositions, particles) of the translation are not dictated by the source text; they appear in the target text for the purpose of joining the content words into meaningful entities, and they are adjusted at the editing stage in accordance with the language and style norms of the target language. Therefore, if we talk about translation equivalents, there would be no Teqs for specific particles, prepositions, or conjunctions, but rather for the constructions they are used in.

For example, the English preposition *with* does not have any Teq in other languages, but the construction ‘with + Noun’ does. In Russian, it would be ‘preposition *s* + noun in the Instrumental case’, in Finnish ‘noun in the Genitive case + postposition *kanssa*’ or ‘noun in the Comitative case'. In addition to these direct correspondences, other translation equivalents are possible.

When working with constructions, one would need to highlight sets of formal features of a certain construction, to collect and study examples from a corpus in the source language, and only after that look up appropriate constructions in another language. Hence, the whole procedure would be different.

Translation equivalence on the level of constructions can also be very helpful with terms and phraseological units. A construction grammar (Fried and Östman 2004) would be a useful instrument to explain relations between multiword elements.

References

Bañón, Marta, Pinzhen Chenz, Barry Haddowz, et al. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 4555–4567.

Baños, Rocío. 2013. ‘That is so cool’: Investigating the translation of adverbial intensifiers in English-Spanish dubbing through a parallel corpus of sitcoms. *Perspectives* 21 (4): 526–542. https://doi.org/10.1080/0907676X.2013.831924.

Čermák, Petr. 2019. InterCorp. A parallel corpus of 40 languages. In *Parallel Corpora: Creation and Applications*, eds. Irene Doval and Maria Teresa Sánchez Nieto, 93–102. Amsterdam/Philadelphia: John Benjamins.

Claire, Brierley, and Hanem El-Farahaty. 2019. An interdisciplinary corpus-based analysis of the translation of كرامة (karāma, 'dignity') and its collocates in Arabic-English constitutions. *JosTrans*, 32: 121–145. https://jostrans.org/issue32/art_brierley.pdf.

Čmejrek, Martin, and Jan Cuřín. 2001. Automatic extraction of terminological translation Lexicon from Czech-English parallel texts. *International Journal of Corpus Linguistics,* 6 (Special Issue): 1–12.

Comrie, Bernard. 1976. *Aspect. An Introduction to the study of verbal aspect and related problems.* Cambridge: Cambridge University Press.

Danesi, Marcel. 2000. *Encyclopedic dictionary of semiotics, media, and communication*. Toronto Studies in Semiotics. Toronto: University of Toronto Press.

Danielsson, Pernilla. 2003. Automatic extraction of meaningful units from corpora: A corpus-driven approach using the word stroke. *International Journal of Corpus Linguistics* 8 (1): 109–127.

Delpech, Estelle Maryline. 2014. *Comparable corpora and computer-assisted translation*. London, Hoboken: John Wiley & Sons.

Doval, Irene, and Maria Teresa Sánchez Nieto. 2019. Parallel corpora in focus: An account of current achievements and challenges. In *Parallel corpora: Creation and applications,* eds. Irene Doval and Maria Teresa Sánchez Nieto, 1–15. Amsterdam/Philadelphia: John Benjamins.

Dobrovol'skij, Dmitrij, and Ludmila Pöppel. 2016. Discursive constructions in the Russian-Swedish dictionary database: A case study of *v tom-to i N*. In *Proceedings of the XVII EURALEX international congress: Lexicography and linguistic diversity,* eds. Tinatin Margalitadze, George Meladze, Ivane Javakhishvili, 668–677. Tbilisi: Tbilisi University Press.

Fried, Mirjam, and Jan-Ola Östman. 2004. Construction grammar: A thumbnail sketch. In *Construction grammar in a cross-language perspective*, ed M. Fried, and J.O. Östman, 11–86. John Benjamins Publishing Company, Philadelphia.

Gamallo, Pablo. 2019. Strategies for building high quality bilingual lexicons from comparable corpora. In *Parallel corpora: Creation and applications,* ed. Irene Doval and Maria Teresa Sánchez Nieto, 251–266. Amsterdam/Philadelphia: John Benjamins.

Garabík, Radovan, and Ludmila Dimitrova. 2015. Extraction and presentation of bilingual correspondences from Slovak-Bulgarian parallel corpus. *Cognitive Studies | Études Cognitives*, 15: 327–334. https://doi.org/10.11649/cs.2015.022.

Grabowski, Łukasz. 2018. On identification of bilingual lexical bundles for translation purposes: The case of an English-Polish comparable corpus of patient information leaflets. In *Multiword units in machine translation and translation technology,* ed. Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor, Violeta Seretan. John Benjamins, 182–199. https://doi.org/10.1075/cilt.341.09gra

Hanks, Patrick. 2012. The corpus revolution in lexicography. *International Journal of Lexicography* 25 (4): 398–436. https://doi.org/10.1093/ijl/ecs026.

Härme, Juho, and Mikhail Mikhailov. 2016. From Russian to Finnish and back: Compiling Russian–Finnish–Russian parallel corpora. In *Translation from /into languages of limited diffusion*, 3rd, ed. Lubica Medvecká, 139–147. Bratislava: The Slovak society of Translators of Scientific and Technical literature.

Johansson, Stig. 2007. *Seeing through multilingual corpora on the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins.

Jurkiewicz-Rohrbacher, Edyta. 2019. *Polish verbal aspect and its Finnish statistical correlates in the light of a parallel corpus.* Ph.D. dissertation. Helsinki: University of Helsinki.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In *The Routledge Handbook of Corpus Linguistics,* ed. M. McCarthy and A. O'Keeffe. London, New York, NY : Routledge, 487–500.

Kenny, Dorothy. 2001. *Lexis and creativity in translation: A corpus based approach*. London and Manchester: St. Jerome Publishing.

Kenny, Dorothy. 2011. Translation units and corpora. In *Corpus-based translation studies: Research and applications*, ed. K. Wallmach, A. Kruger, and J. Munday, 76–102. London: Continuum.

Kilgarriff, Adam. 2013. Using corpora as data sources for dictionaries. In *The Bloomsbury companion to lexicography*, ed. Howard Jackson. Bloomsbury, London, 77–96. http://kilgarriff.co.uk/Publications/Kilg_30aug2012.doc?format=raw

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth machine translation summit* (MT Summit X), Phuket, Thailand.

Kraif, Olivier. 2002. Translation alignment and lexical correspondences. In *Lexis in contrast: Corpus-based approaches*, ed. B. Altenberg, and S. Granger, 271–289. John Benjamins Publishing Company.

Kraif, Olivier. 2003. From translational data to contrastive knowledge: Using bi-text for bilingual lexicons extraction. *International Journal of Corpus Linguistics* 8 (1): 1–29.

Krishnamurthy, Ramesh. 2008. Corpus-driven lexicography. *International Journal of Lexicography* 21 (3): 231–242. https://doi.org/10.1093/ijl/ecn028.

Kubicka, Emilia. 2019. So-called dictionary equivalents confronted with parallel corpora (and the consequences for bilingual lexicography). *Glottodidactica* XLVI/2. Adam Mickiewicz University Press, Poznań, 75–89. DOI: https://doi.org/10.14746/gl.2019.46.2.05

Marco, Josep. 2019. Living with parallel corpora: The potentials and limitations of their use in translation research. In: *Parallel corpora: Creation and applications,* ed. Irene Doval and Maria Teresa Sánchez Nieto, 39–56. Amsterdam/Philadelphia: John Benjamins.

McEnery, Tony, and Richard Xiao. 2007. Parallel and comparable corpora: What are they up to? In *Incorporating corpora: The linguist and the translator*, ed. Gunilla Anderman and Margaret Rogers. Clevedon: Channel View Publications.

McEnery, Tony, and Richard Xiao. 2010. *Corpus-based contrastive studies of English and Chinese*. London and New York: Routledge.

Mikhailov, Mikhail. 2021 (forthcoming). God, devil and christ: A corpus-based study of Russian formulaic idioms and their English and Finnish translation equivalents In *Formulaic language: Theories and methods,* ed. Aleksandar Trklja and Łukasz Grabowski. Berlin: Language Science Press.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus linguistics for translation and contrastive studies: A guide for research*. London and New York: Routledge.

Mikhailov, Mikhail, and Juho Härme. 2015. Parallelnyje korpusa hudožestvennyh tekstov v Tamperskom universitete. (=Parallel corpora of fiction texts at the University of Tampere). *Russkij jazyk za rubežom. Spetsvypusk*, 16–19.

Ooi, Vincent. 1998. *Computer corpus lexicography*. Edinburgh: Edinburgh Univ. Press.

Östling, Robert, and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics (PBML)* 106: 125–146. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf

Perdek, Magdalena. 2012. Lexicographic potential of corpus equivalents: The case of English phrasal verbs and their Polish equivalents . In *Proceedings of the 15th EURALEX international congress*, 7–11 Aug 2012, Oslo, 376–388.

Pöppel, Ludmila. 2018. The construction *возьми и + V IMP*: A corpus-based study. *Zeitschrift Für Slawistik* 63 (1): 111–119.

RG 1980. = Švedova, Natalja Ju. (ed.). *Russkaja grammatika.* [Russian grammar]. Moscow: Nauka. http://www.rusgram.narod.ru/.

Ramón, Noelia, and Belén Labrador. 2008. Translations of '-ly' adverbs of degree in an English-Spanish Parallel Corpus. *Target* 20 (2): 275–296. https://doi.org/10.1075/target.20.2.05ram.

Salkie, Rafael. 2008. How can lexicographers use a translation corpus? In *Proceedings of the international symposium on using corpora in contrastive and translation studies. xiao,* eds Richard, Lianzhen He and Ming Yue. Hangzhou: Zhejiang University. https://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf.

Salkie, Rafael. 2002. Two types of translation equivalence. In *Lexis in contrast,*, vol. 7. ed. B. Altenberg and S. Granger. Amsterdam: John Benjamins.

Sinclair, John. 2001. Data-derived multilingual lexicons. *International Journal of Corpus Linguistics* 6 (Special Issue): 79–94.

Štichauer, Pavel, and Petr Čermák. 2016. Causative constructions of the hacer / fare + verb type in Spanish and Italian, and their Czech counterparts: A parallel corpus-based study. *Linguistica Pragensia* 2: 7–20.

Tarp, Sven. 2020. A dangerous cocktail: Databases, information techniques and lack of visions. In *Studies on Multilingual Lexicography*, ed. María José Domínguez Vázquez, Mónica Mirazo Balsa and Carlos Valcárcel Riveiro, 47–66. De Gruyter.

Terryn, Ayla Rigouts, Véronique Hoste, and Els Lefever. 2020. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Lang Resources and Evaluation* 54: 385–418. https://doi.org/10.1007/s10579-019-09453-9.

Teubert, Wolfgang. 2001. Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics*, 6 (Special Issue), 125–153.

Tiedemann, Jörg. 1997. Automatical Lexicon Extraction from Aligned Bilingual Corpora. M.A. thesis. Department of Linguistics, University of Uppsala/Otto-von-Guericke Universität Magdeburg.

Tiedemann, Jörg. 1998. Extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the 11th nordic conference on computational linguistics*, Center for Sprogteknologi, Copenhagen.

Tiedemann, Jörg. 2004. Word to word alignment strategies. In *Proceedings of coling 2004*, 212–218. http://stp.lingfil.uu.se/~joerg/published/coling04.pdf

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th international conference on language resources and evaluation (LREC 2012)*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Walter, Elizebeth. 2010. Using corpora to write dictionaries. In *The Routledge handbook of corpus linguistics*, eds M. McCarthy and A. O'Keeffe. London; New York, NY: Routledge.

Zakharov, Viktor, and Svetlana Bogdanova. 2020. *Korpusnaâ lingvistika*. Sankt-Peterburg: SPbGU.

Zalizniak Anna, A., G.V. Denisova, and I.L. Mikaeljan. 2018. Russkoe kak-nibud′ po dannym parallel′nyh korpusov. In *Komp'ûternaâ lingvistika i intellektual'nye tehnologii: po materialam meždunarodnoj konferencii «Dialog 2018»*. Moskva: RGGU.

Ziemski, M., M. Junczys-Dowmunt, and B. Pouliquen. 2016. The United Nations Parallel Corpus. In *Language resources and evaluation (LREC'16)*, Portorož, Slovenia, May 2016. https://conferences.unite.un.org/UNCORPUS/Content/Doc/un.pdf.

Mikhail Mikhailov is a Professor of Translation Studies (Finnish and Russian) at the Tampere University, Finland. He is one of the authors of the book “Corpus Linguistics for Translation and Contrastive Studies” (Routledge 2016). He compiles multilingual corpora and develops web-based corpus software. His research covers corpus-based translation studies with a particular focus on parallel and comparable corpora, terminological studies and translation technologies.

An Intralingual Parallel Corpus of Translations into German Easy Language (Geasy Corpus): What Sentence Alignments Can Tell Us About Translation Strategies in Intralingual Translation

Silvia Hansen-Schirra, Jean Nitzke, and Silke Gutermuth

Abstract Parallel corpora are traditionally interlingual and contain source and target texts in different languages. However, intralingual translations into Easy Language (EL) become more and more common in various countries. First intralingual corpora have been built up and investigated in terms of linguistic and structural features, but a translation-driven corpus linguistic approach is still missing to empirically describe the strategies of Easy Language translation, the characteristics of translated texts as well as to make these parallel corpora usable for professionalising and automatising translation processes. In this paper, we introduce an intralingual parallel corpus of translations into German Easy Language (*Geasy Corpus*). It contains published professional translations from Standard German into German Easy Language, including different text types and various formulation guidelines for German Easy Language. Currently, the corpus contains 1,087,643 words of source text and 292,552 words of Easy Language translations. So far, 93 (of 276) texts have been sentence aligned. We compare descriptive values, investigate the alignments, and describe which translation strategies are revealed to give first empirical evidence on the characteristics of Easy Language translations. Finally, we will discuss the potentials of tree annotations for Easy Language corpora, summarise our findings and give on outlook on future research.

S. Hansen-Schirra · J. Nitzke (✉) · S. Gutermuth
Johannes Gutenberg University, Mainz, Germany
e-mail: jean.nitzke@uia.no

S. Hansen-Schirra
e-mail: hansenss@uni-mainz.de

S. Gutermuth
e-mail: gutermsi@uni-mainz.de

J. Nitzke
University of Adger, Kristiansand, Norway

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_11

Keywords Parallel corpora · Intralingual translation · Easy Language · Translation strategies

1 Introduction

Since the implementation of the UN Convention on the Rights of Persons with Disabilities in Germany in 2009, accessible communication, and especially intralingual translations into Easy Language become more and more common. From an applied perspective, Easy Language was developed for people with intellectual disabilities with the aim to reduce linguistic complexity and to enhance comprehensibility. From a scientific-empirical point of view, first parallel corpora have been built and investigated in terms of linguistic and structural features (e.g., Battisti et al. 2019; Klaper et al. 2013). The scientific examination of these corpora, however, still lacks a translation-driven approach to empirically contrast the characteristics of source and target texts as well as to examine the strategies in Easy Language translation. These parallel corpora can help professionalizing the field and automatizing translation processes (as suggested in Hansen-Schirra et al. 2020a). Hence, we want to start exploring the field by investigating sentence alignments and what they can tell us about potential translation strategies.

In this paper, we will first present a brief overview about Easy Language in Germany (Sect. 2). We will then introduce an intralingual parallel corpus of translations into German Easy Language (Geasy Corpus) that we currently build (Sect. 3). The corpus contains published professional translations from Standard German into German Easy Language. The corpus includes different text types and various formulation guidelines for German Easy Language (Bredel and Maaß 2016a, b; Netzwerk Leichte Sprache 2013; Inclusion Europe 2009). Both corpus creation and alignment processes are work in progress. We will compare descriptive values like lexical density, sentences length, type-token ratio, and term density, as well as briefly describe the use of pictures in the corpus to shed light on the differences in information density and structural complexity of German Easy Language. In Sect. 4, we will quantify and investigate the sentence alignment of the texts that have been aligned so far. Section 5 will present how these alignments can be used to potentially reveal translation strategies. Sentence alignment is not very straightforward in Easy Language corpora, as the translation is usually not a direct equivalent of the source text segment. There happen to be considerable changes between source and target texts which are necessary to make the texts readable and comprehensible for the target audience. Therefore, we will also describe the shortcomings of a sentence-based analysis and present the potential benefits of a tree alignment. Finally, we will summarize our findings and present some ideas for future research.

2 A Quick Introduction to German Easy Language

Access to information for people with disabilities has become an important issue in the countries that have ratified the UN Convention on the Rights of People with Disabilities. Easy Language is a linguistic approach to reduce language complexity for the sake of better readability and comprehensibility. It is, therefore, seen as one of the central pillars of communicative inclusion. Accordingly, the legislation concerning Easy Language progressed greatly in recent years. In Germany, the Convention on the Rights of Persons with Disabilities of the United Nations (UN CRPD) has significantly changed the way in which disabilities are addressed in the political and legal discourse. The implementation of the UN CRPD is on the agenda of each German Federal state, each municipality, each regional parliament, and local authority, as well as the Federal Government and its public bodies. Laws and regulations have been passed and implemented at the federal and national level (on the legal situation of Easy Language, see Lang 2019). On this basis, paragraph 11 on "Comprehensibility and Easy Language" [1]was added to the Act on Equal Opportunities of Persons with Disabilities ("Behindertengleichstellungsgesetz") in 2018. For a more detailed historical development and a comparison of different varieties of EL in Europe see Lindholm and Vanhatalo (forthcoming).

As a result, political and public institutions have to face the need to translate existing texts with domain-specific contents into Easy Language. Easy Language ("Leichte Sprache" in German) addresses recipients with cognitive disabilities, prelingual hearing loss, aphasia, dementia type illnesses and Parkinson's disease. They can be regarded as primary target groups, which are legally entitled to Easy Language (roughly 3% of the German population, cf. Maaß et al. forthcoming). In addition, there are other communication impairments which afford Easy Language texts, such as poor reading or language skills (e.g., migrants).

Translating specialized or technical content to the target group means filling the gap between expert knowledge and the knowledge of the recipient. Easy Language is used as the means to create a common ground (Pickering and Garrod 2004). It is a sort of controlled language variety of Standard German, which aims for a better readability and comprehensibility of texts. A controlled language is a subset of a natural language such as German, which is restricted according to certain rules (Lehrndorfer 1996). Controlled languages have traditionally been used for technical documentation to make them more consistent and more comprehensible. Within the context of accessible communication, other text types are relevant for EL translation, too. This especially holds true for administrative and legal texts since they are mentioned in the respective regulations and acts to ensure participation of the target groups (see above). More recently, EL translation has also been considered for multimedia text types (e.g., EL subtitling, cf. Maaß and Hernández Garrido 2020) or for literary texts (cf. Maaß et al. forthcoming). Another area of application is the medical discourse where illnesses and their treatment are explained, or informed

[1] https://www.gesetze-im-internet.de/bgg/__11.html, last accessed 08/05/21

consent documents are translated for the target groups (cf. Maaß 2020). It is hard to quantify the amount of EL texts since they are typically used as printed leaflets and translated by various stakeholders (e.g., translation agencies, empowerment associations, editors in federal ministries or publishing houses, etc.). However, the amount of EL texts is increasing rapidly.

So far, rules and formulation guidelines for Easy Language for German have been based on practical experience (Inclusion Europe 2009; Netzwerk Leichte Sprache 2013) or linguistic theory (Bredel and Maaß 2016a, b). The guidelines and rule sets for Easy Language can vary a lot and can also be rather vague. Further, not all guidelines are freely accessible. In general, they suggest, amongst others:

- limitations in the lexicon: specific terms and lexical variation should be avoided
- reduced complexity on the morphological level: compounds should be avoided, or segmentation aids (hyphen or mediopoint) should be used
- reduced complexity on the phrasal level: complex pre- and post-modifications of phrases (e.g., genitive attributes) should be avoided
- reduced complexity on the syntactic level: subordinate clauses should be avoided and rephrased in main clauses instead
- reduced complexity on the textual level: repetition of nouns should be used instead of pronouns to build cohesive ties; each sentence should include only one proposition
- typographic facilitation: negations should be printed in bold face; each sentence should start on a new line; the integration of pictures should enhance the comprehension of key concepts and terminology.

Maaß (2015) differentiates between general principles of Easy Language, rules regarding the semiotic level, word level, sentences level, text level, rules regarding typography and layout as well as rules regarding translation strategies. Some rules influence the sentence structure and, hence, can be used to interpret the alignment results. One of the general principles for Easy Language texts is an orientation towards activities, which includes amongst others a spread and repetition of information (instead of compensating information and avoiding redundancies, ibid.: 76–81). The information selection for an Easy Language text/translation depends on the topic of the text (ibid.: 129). This implies for Easy Language texts that not all information of the target text necessarily has to be integrated into the target text. Often, information is added that is not explicitly available in the source text to explain or exemplify difficult terminology or phrases. On the other side, information from the source text is reduced to make the amount of text suitable for the target group. A comprehensive comparison of the different rule sets for German can be found in Maaß (2020) and for European EL varieties in Lindholm and Vanhatalo (forthcoming).

The following example presents some of the mentioned characteristics (reduced complexity, added and reduced information, adapted typography—more examples of German Easy Language and its source text can be found in the following sections):

- **SL:** Die UN-Konvention
  Die Unterzeichnung der UN-Behindertenrechtskonvention durch Deutschland während der deutschen EU-Ratspräsidentschaft war ein politisch wichtiges Signal für die anderen Mitgliedstaaten der Europäischen Union.
  (English: The UN Convention
  The signing of the UN Convention on the Rights of Persons with Disabilities by Germany during the German EU Council Presidency was a politically important sign for the remaining member states of the European Union)
  EL: Die UN-Konvention
  Konvention ist ein anderes Wort für **Vertrag**.
  Wir sprechen:
  Kon-wen-zion
  Die **UN-Konvention** ist ein Vertrag zwischen sehr vielen Ländern auf der Welt.
  Die **Vereinten Nationen** haben diesen Vertrag geschlossen.
  Für **Vereinte Nationen** sagen wir auch **UN**.
  Deshalb nennen wir den Vertrag:
  UN-Konvention.
  In dem Vertrag steht:
  Menschen mit und ohne Behinderung haben die gleichen Rechte.
  Menschen mit und ohne Behinderung müssen gleich behandelt werden.
  Die Regierungen müssen alle Hindernisse für Menschen mit Behinderung beseitigen.
  Auch viele Länder in der EU haben die UN-Konvention unterschrieben.
  (English:
  The UN Convention
  Convention is another word for **contract**.
  We pronounce it:
  Con-wen-tion
  The **UN Convention** is a contract between a lot of countries in the world.
  The **United Nations** entered this contract.
  We call the **United Nations** also **UN**.
  Therefore, we call the contract:
  UN Convention
  The contract says:
  People with and without disabilities have the same rights.
  People with and without disabilities have to be treated equally.
  The governments have to eliminate all obstacles for people with disabilities.
  Many countries of the EU have signed the UN conventions, too.

Currently, these rules have been empirically evaluated from a user-based perspective (Bock 2019, Hansen-Schirra et al. 2020b, Deilen 2020, Schiffl 2020, Sommer 2020, Gutermuth 2020) and result in recommendations to optimise or specify the rules.

First corpus studies focus on rather computational-linguistic aspects. Battisti et al. (2019) use their corpus of simplified German texts for a natural language approach and for unsupervised machine learning to empirically investigate "whether different complexity levels exist in previous German simplification practice in the first place" (ibid.: 3). Klaper et al. (2013) built a parallel corpus with Standard German and Simple German with the aim to have training data for statistical machine translation systems. Developing machine translation systems for intralingual translation is also discussed in Hansen-Schirra et al. (2020a). However, a corpus-based quantification of existing Easy Language usage patterns, translation strategies, and contrastive text characteristics remains a research desideratum, which we try to address in this paper with the help of the Geasy Corpus.

3 The German Easy Language (Geasy) Corpus

In translation studies, we differentiate between two types of corpus designs: the parallel corpus and the monolingually comparable corpus. A parallel corpus is defined as a collection of source language texts and translations of those texts into a target language. In computational linguistics, such corpora are used in bilingual lexicography (Sahlgren and Karlgren 2005) and as training corpora for machine translation systems (Koehn 2005; Artetxe et al. 2017). In translation research, parallel corpora provide information on language-pair specific translation patterns and to investigate translation quality (Zanettin 2000).

Monolingually comparable corpora are collections of translations and original texts in the target language. Comparable corpora "should cover a similar domain, variety of language and time span, and be of comparable length" (Baker 1995: 23); they have

> the potential to reveal most about features specific to translated text, i.e., those features that occur exclusively, or with unusually low or high frequency, in translated text as opposed to other types of text production, and that cannot be traced back to the influence of any one particular source text or language (Kenny 1997).

Comparable corpora are used to test hypotheses on translation-universals or specific features of translations such as explicitation, simplification, normalization/ conservatism (Baker 1995, Hansen-Schirra et al. 2013).

The Geasy Corpus is a combination of both corpus types described. It consists of parallel but monolingual subcorpora. The corpus is parallel since it includes source and target texts, which we are currently aligning. However, both of these monolingual subcorpora are in German language. The translation direction is Standard German into Easy Language German, i.e., the Geasy Corpus is a monolingual, parallel corpus containing texts from various genre. The translations were created according to different guidelines because different guidelines are common in Germany as mentioned above. In the following, we will present some basic characteristics of both subcorpora in comparison. The alignment data will be analysed and assessed in Sects. 4 and 5.

Currently, the corpus contains 1,087,643 words of source text and 292,552 words of Easy Language translations. For now, most source and target texts are publicly available on website, usually provided by a public organisation, etc. So far, 93 texts (33.7% of 276 texts in total) from three sources have been sentence aligned. Not all texts in the corpus are suitable for alignment. For example, one source text contains 59,795 words, while the Easy Language translation only consists of 2,728 words. We can assume that the choice was deliberate to reduce the information of the Easy Language text that drastically. However, we decided that this example and other source and target texts were too different for alignment and would not fit our purposes. This also applies, e.g., for leaflets in Easy Language that give additional information to enable the reader to fill out a form.

The source texts of the aligned data contain 33,061 words and 1596 sentences, which results in a medium sentence length of 20.7 words per sentence. The Easy Language translations of the aligned data consist of 41,722 words and 4090 sentences (average sentence length: 10.2 words/sentence). Easy Language translations, therefore, present the contents in more words and more sentences (we will discuss information restructuring and sentences splitting in Sects. 4 and 5). On average, sentences in Easy Language are only half as long as sentences in Standard Language. Both characteristics (more sentences and less words per sentence) point into the direction that the Easy Language texts are less complex than the Standard Language texts, which corresponds to the goals of Easy Language. Certain rules for Easy Language texts encourage sentence splitting, e.g., there should be only one information per sentences, subclauses should be avoided, or information should be spread (Maaß 2015). To further investigate the complexity of Easy Language translations compared to Standard Language texts, we will have a closer look a lexical density and the type-token ratio.

Table 1 shows the distribution of four main lexical word classes (all content words) in the Easy Language and Standard subcorpora. The part-of-speech tagging was carried out with Sketch Engine (Kilgarriff et al. 2014). The general assumption is that the lexical density in Easy Language should be lower compared to Standard Language (Hansen-Schirra and Gutermuth 2018), but this is not the case in our data (t-test: t(5.82) = 0.41, p = 0.7). In total, the numbers for lexical density are very similar. Interestingly, we find a similar frequency for nouns in the two subcorpora. According to Maaß (2015: 76), however, Easy Language texts should rather be in verbal than in nominal style. One explanation for the high number of nouns might be that the texts often are domain-specific (e.g., explaining legal contents) and, hence, have to use and introduce important concepts and terms. We can observe a

Table 1 Distribution of lexical word classes in the Geasy Corpus

	Easy Language	Standard German
Nouns	13,128 (31.47%)	11,119 (33.63%)
Verbs	7572 (18.15%)	4163 (12.59%)
Adjectives	2403 (5.76%)	3437 (10.40%)
Adverbs	2258 (5.41%)	1122 (3.39%)
TOTAL	25,361 (60.79%)	19,841 (60.01%)

significant shift in the usage of verbs and adjectives ($X^2(1) = 867.34$, $p < 0.0001$). Fewer adjectives and more adverbs are used in Easy Language, which is in line with an increase in verbal style. Further, adjectives are often used in complex nominal phrases, which should be avoided in Easy Language. The frequency of verbs increases in the Easy Language corpus. This corroborates the results of our sentences analysis. As mentioned above, information is split—noun phrases, enumerations, etc. are dissolved—and presented in single sentences as in example 1. In example 1, the people who can get counselling and support are summarised in one sentence in the source text, while they are listed in two separate sentences in the target text. Further, an explanation for chronic diseases is added to the target text.

(1) SL: Die Beratungsstelle bietet Menschen mit körperlichen Behinderungen oder chronischen Erkrankungen und ihre Angehörigen Beratung und Unterstützung […] an.[2] (English: The counselling centre offers counselling and support to people with physical handicaps or chronic diseases and their relatives […].)
EL: Menschen mit Körper-Behinderungen und ihre Familien können Beratung und Hilfe bekommen. Die Beratung ist auch für Menschen mit chronischen Krankheiten. Chronisch heißt in Leichter Sprache: Diese Krankheiten gehen nicht mehr weg. (English: People with physical handicaps and their families can get counselling and help. The counselling is also for people with chronic diseases. Chronic means in Easy Language: The disease is permanent.)

Table 2 presents the type-token ratio of the Easy Language and Standard subcorpora. The assumption concerning this feature is that the easier the text, the lower the type-token ratio becomes, which seems to be true on first sight in the analysed section of the Geasy corpus. The figures suggest that fewer lexical items and terms are introduced in the Easy Language texts, which makes it less lexically diverse and accordingly less complex. This is in line with the findings in Hansen-Schirra and Gutermuth (2018: 16). More sophisticated analyses will shed more light on the complexity of the texts. Some studies already started to explore complexity in EL texts, e.g., Gutermuth (2020) correlated textual complexity with cognitive processing costs while reading EL texts and Battisti et al. (2019) applied unsupervised machine learning techniques to EL texts to cluster and classify complexity levels of EL corpora.

Finally, we address the quantification and characterisation of the use of images and pictures. None of the source texts contained pictures or images. According to Maaß (2015: 86) different semiotic representations help recipients of Easy Language to understand the contents of the texts. Therefore, it is valid to use pictures, figures, or to highlight important terms or phrases in the texts. Only one text of the 93 corpus texts in Easy Language did not contain pictures in the original formatting. In total, 867 pictures were counted in the target texts (mean = 9.42 SD = 8.83). The first two subcorpora used the same set of pictures (*Lebenshilfe*

[2] All examples discussed in the paper are taken from the Geasy Corpus. For further information see https://traco.uni-mainz.de/geasy-korpus/, last accessed 08/05/21.

Table 2 Type-token ratios of the two subcorpora in the Geasy Corpus

	Easy Language	Standard German
Types per 100 tokens	7.35	18.51

Bremen), while the third subcorpus used different pictures. All pictures were designed in a comic style, except for some real-life pictures in the third corpus that referred to real life objects, e.g., the town hall of Hamburg. Interestingly, some pictures recurred in the texts, usually when same or similar concepts are introduced in different texts, probably to increase the memorability of the concepts.

In summary, the analysed subcorpus already shows differences in characteristics between source and target texts. These differences confirm that the Easy Language translations seem to be less complex than the Standard Language source texts. However, more data points might be necessary to get significant result for some aspects.

4 Alignment Characteristics

In this section, we want to analyse the sentence alignment of the subcorpus that was aligned so far. Source and target texts were aligned with the help of the translation memory tool *memsource*. Source and target texts can be automatically prealigned and then be viewed and corrected manually.

The basis for the alignment process was the source text. In total, the subcorpus consists of 1816 alignments. Different alignments can be observed between source and target text. These will be introduced with a brief interpretative approach and quantified in the following:

- 1:1 alignment—one source sentence is aligned with one target sentence
- 1:0 alignment—a source sentence has no equivalent in the target sentence
- n:1 alignment—several sentences in the source texts are aligned with one sentence in the target text
- 0:n alignment—several sentences in the target text have no equivalent in the source text
- 1:n alignment—one source text sentence is represented by several sentences in the target text
- n:m alignment—several sentences are represented by several sentences in the target text
  - n > m (without n:1)—more sentences in the source sentence were aligned with less sentences in the target text
  - n = m (without 1:1)—several sentences in the source text were aligned with an equal number of sentences in the target text
  - n < m (without 0:n and 1:n)—less sentences in the source sentence were aligned with more sentences in the target text

Table 3 Quantification of alignments in absolute numbers and %

	Absolute figures	in %
1:1 alignment	664	36.56
1:0 alignment (including one 2:0 alignment)	17	0.94
n:1 alignment	21	1.16
0:n alignment = > 0:1	146 = > 6	8.04
1:n alignment = > 1:2	868 = > 308	47.80
n:m alignment = > n > m (without n:1) = > n = m = > n < m (without 0:n and 1:n)	100 = > 4 = > 29 = > 67	5.51

The following table shows the absolute and relative figure of our alignments.

Table 3 shows that most alignments are 1:n-alignments, followed by 1:1-alignments. 0:n-alignments and n:m-alignments occurred in the lower medium section, while the least alignments can be counted for n:1 and 1:0 alignments. These findings corroborate the results by Klaper et al. (2013) who report very low scores for automatically aligning an intralingual corpus of Standard and Easy Language. They attribute this low performance to specificities of the language variety, the domain and massive changes from source to target text.

Furthermore, the alignments suggest that, for those texts, Easy Language translations rather add, enrich and restructure information than reduce or delete contents. On the basis of the alignment figures, one could assume that it might be more likely that information is added or enriched than that the information and structure is identical. This seems also to be the case for the n:m-alignments, where n < m-alignments were counted more often than n = m-alignments, and by far as n > m-alignments. To shed more light on this assumption, we will examine the alignments in further detail and interpret them with respect to translation strategies in the next section.

5 Translation Strategies

Let us look at some phenomena in more detail. Different translation strategies can be applied when translating from Standard Language into Easy Language. These strategies might be similar to interlingual translation strategies, might differ, or might deviate to a certain degree. Potentially, different alignments point towards certain translation strategies. In this section, we want to concentrate on translation strategies that focus on the information presentation and structure. We will combine alignment patterns and potential translation strategy and discuss this procedure critically.

Interestingly, the majority of alignments are 1:n-alignment in the subcorpus, which is in line with most of the Easy Language rules presented above. These alignments, further, point to a form of sentence splitting. Sentence splitting is a strategy to restructure and simplify complex sentences in interlingual translation, as well: "[T]ranslating a source sentence by a sequence of independent target sentences aims at reducing informational density of the target text as opposed to the source text by increasing incrementality." (Fabricius-Hansen 1999: 188).

On the other hand, only few instances were found, in which deletion and reduction processes become obvious, although an information selection is supported by the rules to keep the texts short enough for the intended recipients. Another aspect, which can hardly be tested with the sentence alignment is the structure on the macro level. It was observed, however not quantified, that the texts in the analysed subcorpora did not show any evidence concerning a restructuring on a macro-level, meaning that the information of the source text was presented in a similar order in the target text. The reason might be that the source texts we aligned so far were rather short (mean: 355.5 words/text) and accordingly the target texts were still short enough (mean: 448.6 words/text) to be appropriate for the target group.

Further, we want to stress that sentence alignment does not shed light on the information structure within the sentences. In (2), you can see an example in which the source text is reduced. However, the alignment remains a 1:1-alignment.

(2) **ST**: Themen wie Bildung, lebenslanges Lernen und berufliche Anpassung werden in Zeiten einer globalisierten und digitalisierten Arbeitswelt immer wichtiger. (Eng.: Topics like education, life-long learning and career adaptation are becoming more important in times of a globalised and digital work environment.)

EL: Eine gute Schul-Bildung ist immer wichtiger. (English: Good school-education becomes more important.)

"Lebenslanges Lernen", "berufliche Anpassung", and "in Zeiten einer globalisierten und digitalisierten Arbeitswelt" are not represented in the target text sentence illustrating an information reduction strategy on the lexical level (cf. Hansen-Schirra et al. 2020a for a discussion of translation strategies). Further the term "Bildung" was constrained to the education that is imparted in school. These changes are not presented in the sentence alignment, which evokes the impressions that the same information are presented. Hence, it seems plausible to include word or character counts in future research in addition to sentence alignment to represent these changes.

In example (3), the mere alignment figures suggest that information was added since we have a 1:3 alignment. However, when we look closer at the contents, the information was reduced concerning the actors involved, while the information on the "discussion process" itself was exemplified with questions. This exemplification strategy is typical of Easy Language translation (cf. Hansen-Schirra et al. 2020a).

(3) **ST**: Hierzu wurde am 23. Januar 2009 ein Diskussionsprozess mit der Beauftragten für die Belange behinderter Menschen, den Verbänden behinderter Menschen, der Leistungserbringer und den kommunalen Spitzenverbänden angestoßen. (English: To this topic, a discussion process was initiated on the 23rd of January in 2009 including the representative for people with disabilities, associations of people with disabilities, providers, and the community's umbrella organisations.)
EL: Alle haben überlegt: Sind die Vorschläge gut? Was muss anders sein? (English: Everybody reflected on: Are the proposals good? What has to be different?)

Finally, the example in (4) represents another difficulty when it comes to sentence alignment.

(4) ST: Außerdem müssen Prüfungsordnungen die besonderen Belange von Studierenden mit Behinderung zur Wahrung der Chancengleichheit berücksichtigen. (English: Furthermore, exam regulations have to respect the special interests of students with disabilities to keep the opportunities equal.)
EL: Studenten und Studentinnen mit Behinderung müssen bei den Prüfungen die gleichen Chancen haben. Zum Beispiel muss ein blinder Student die Prüfungs-Aufgabe in Blinden-Schrift bekommen. Das muss auch in den Prüfungs-Ordnungen stehen. (English: Male and female students with disabilities have to have the same opportunities in exams. For example, a blind student has to get the exam task in braille. This has to be mentioned in the exam regulations.)

The source sentences do not give any reasons to include the example that is inserted in the target text. The translator probably decided to include the example to make the content more concrete. When analysing the sentence alignment, however, the question comes up whether to describe the alignment as a 1:3-alignment or as a 1:2-alignment plus a 0:1-alignment. In our analysis, we decided that examples that were added and for which there was no indication in the source text were counted as the latter. However, both counting possibilities seem reasonable.

These are only some examples why a mere sentence alignment might not be enough to analyse translation strategies in Easy Language translation. Information reorganisation is not only processed by sentences splitting, additions or deletion of whole sentences, or similar strategies. Often, information is reorganised within the sentences or split sentences are further reduced or enriched. In the next section, we want to discuss another approach.

6 Tree Alignment

As we have discussed in Sect. 4, a sentence-based analysis only gives first impressions of the information restructuring in intralingual translation. According to the rules for German Easy Language (see above), complex sentences with finite and non-finite subordinate clauses have to be split into several main clauses in the EL translation. 1:n or n:m alignments show these translation patterns but they do not reveal where which parts of the clause complex are exactly moved to, whether the information is restructured or preserved in the same order, which subordinate clauses are not translated and which information is added. Hence, we suggest a tree-based alignment in future research, i.e., building up a parallel treebank. A treebank is a corpus which is annotated for syntactic information (i.e., syntactic trees). This means that it includes information on parts-of-speech, morphology, phrase structure, syntactic functions, main and subordinate clauses and their dependency structure. Recently, the need has emerged to build up interlingual parallel treebanks: In computational linguistics, they are employed for multilingual grammar induction, as test suites and gold standards for alignment tools and multilingual taggers and parsers (Volk et al. 2011). Additionally, they are used for the development of corpus-based machine translation systems (cf. Čmejrek et al. 2004). In translation studies, interlingual parallel treebanks are needed as linguistically enriched text basis for empirical research on translations strategies and specific properties of translations (cf. Hansen-Schirra et al. 2012). Following this argumentation, we argue that we need a monolingual parallel treebank for the investigation of translation strategies from Standard into Easy Language. Such a monolingual parallel treebank sheds light on phrase and clause structures, syntactic functions and dependencies, and the alignments thereof. The annotation and alignment of the Geasy Corpus in terms of treebank structures and alignments is still work in progress. Nevertheless, example 5 taken from our corpus illustrates the advantages of such a monolingual parallel treebank for the investigation of n:m alignments.

(5) **ST**: Auf Grundlage der Globalrichtlinie Stadtteilkultur stellt die Behörde den sieben Hamburger Bezirken Fördermittel zur Verfügung. (English: On the basis of the global guidelines for urban district culture, the authority provides funding to the seven Hamburg districts.)
EL: Hamburg ist eine sehr große Stadt. Es gibt 7 große Bezirke in Hamburg. Es gibt besondere Förderung für die Kultur in den Bezirken. (English: Hamburg is a very big city. There are 7 large districts in Hamburg. There is special funding for culture in the districts.)

In Fig. 1, we can see the dependency trees, clause and phrase annotations of the source sentence and the translation, which consists of three sentences. The ParZu parser was used for dependency annotation (cf. Sennrich et al. 2009). The alignment on the sentence level suggested in Sect. 4 results in a 1:3-alignment since one source sentence is translated into three target sentences (s. Table 4). However,

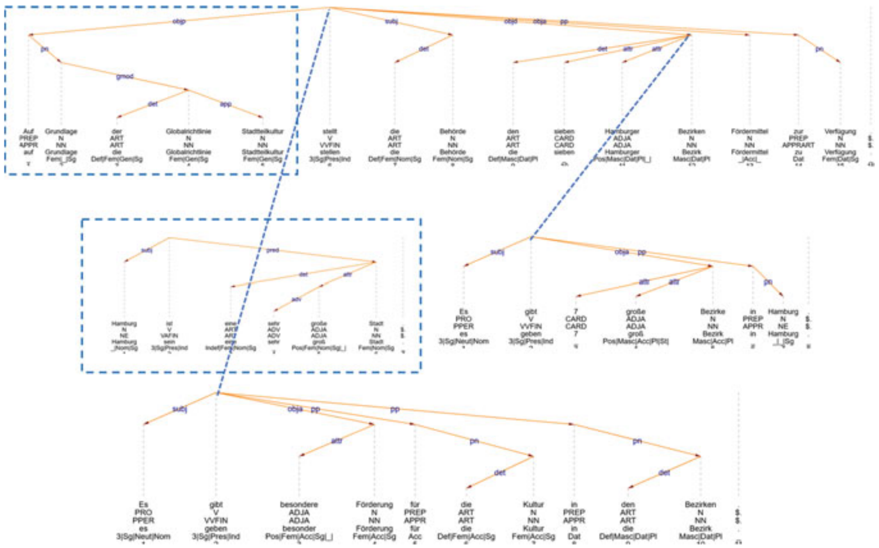

objp
pn
gmod
det
app
subj
det
objd
obja
pp
det
attr
attr
pn
Auf PREP APPR auf
Grundlage N NN Grundlage Fem|_|Sg
der ART ART die Def|Fem|Gen|Sg
Globalrichtlinie N NN Globalrichtlinie Fem|Gen|Sg
Stadtteilkultur N NN Stadtteilkultur Fem|Gen|Sg
stellt V VVFIN stellen 3|Sg|Pres|Ind
die ART ART die Def|Fem|Nom|Sg
Behörde N NN Behörde Fem|Nom|Sg
den ART ART die Def|Masc|Dat|Pl
sieben CARD CARD sieben
Hamburger ADJA ADJA Hamburger Pos|Masc|Dat|Pl|_
Bezirken N NN Bezirk Masc|Dat|Pl
Fördermittel N NN Fördermittel _|Acc|_
zur PREP APPRART zu Dat
Verfügung N NN Verfügung Fem|Dat|Sg
$. $.
subj
pred
det
attr
adv
Hamburg N NE Hamburg _|Nom|Sg
ist V VAFIN sein 3|Sg|Pres|Ind
eine ART ART eine Indef|Fem|Nom|Sg
sehr ADV ADV sehr
große ADJA ADJA groß Pos|Fem|Nom|Sg|_
Stadt N NN Stadt Fem|Nom|Sg
$. $.
subj
obja
pp
attr
attr
pn
Es PRO PPER es 3|Sg|Neut|Nom
gibt V VVFIN geben 3|Sg|Pres|Ind
7 CARD CARD 7
große ADJA ADJA groß Pos|Masc|Acc|Pl|St|
Bezirke N NN Bezirk Masc|Acc|Pl
in PREP APPR in
Hamburg N NE Hamburg _|_|Sg
$. $.
subj
obja
pp
pp
attr
pn
det
pn
det
Es PRO PPER es 3|Sg|Neut|Nom
gibt V VVFIN geben 3|Sg|Pres|Ind
besondere ADJA ADJA besonder Pos|Fem|Acc|Sg|_
Förderung N NN Förderung Fem|Acc|Sg
für PREP APPR für Acc
die ART ART die Def|Fem|Acc|Sg
Kultur N NN Kultur Fem|Acc|Sg
in PREP APPR in Dat
den ART ART die Def|Masc|Dat|Pl
Bezirken N NN Bezirk Masc|Dat|Pl
$. $.

Fig. 1 Parallel treebank of example (5)

Table 4 Linear alignment of clauses and phrases of example (5)

Source text in Standard German		Target text in Easy German
		[A] Hamburg ist eine sehr große Stadt. [clause simplex]
[a] Auf Grundlage der Globalrichtlinie Stadtteilkultur [prepositional phrase]		
[b] stellt die Behörde [predicate, subject] [d] Fördermittel zur Verfügung [direct object]		
	[c] den sieben Hamburger Bezirken [indirect object]	[B] Es gibt 7 große Bezirke in Hamburg. [clause simplex]
		[C] Es gibt besondere Förderung für die Kultur in den Bezirken. [clause simplex]

taking a closer look into this translation reveals that we have several empty links from Standard to Easy Language but also the other way around. Empty links are units in the target text which do not have matches in the source text and vice versa (cf. Hansen-Schirra et al. 2017). They may occur on all language levels. This requires alignments on several levels: sentence level, clause level, phrase level, word level, etc. In the example, the third sentence in the target text can be aligned with the source sentence (the alignments are indicated by the dotted line in the figure). However, the second sentence in the EL text is equivalent to the indirect object in the source language, which results in a phrase-sentence alignment from source to target text. This sentence further explains the word “Bezirke” (districts) in the aligned source text phrase, but it is also referring to the same word in the third sentence of the EL translation and establishes a co-reference relation between the two EL sentences. On the basis of this annotation and alignment, we will be able to automatically extract phrase-sentence or clause-sentence alignments.

In addition, we can also search for empty alignment links (see the dotted boxes in the figure). The first sentence in the EL translation is an explanation which is added to create further common ground for the target group (cf. Hansen-Schirra et al. 2020b; Pickering and Garrod 2004). It explains that Hamburg is a big city, which is subdivided into districts. This information is not necessary in the source text since it can be taken as general knowledge by the unimpaired reader. This additional explanation results in an empty link when aligning the sentence pairs. Another empty link occurs on phrase level and affects the first prepositional phrase in the source sentence “On the basis of the global guidelines for urban district culture”. This prepositional phrase specifies the administrative basis for the funding. This detail is not necessary for the target group of the EL text and therefore left out. Here, information selection is an important translation strategy in order to keep the text short and processable for the EL reader (cf. Hansen-Schirra et al. 2020a). In conclusion, this example shows that we will be able to automatically quantify and extract empty links in the source text resulting from information selection strategies as well as in the target text resulting from explanation or exemplification strategies. An integration of the word alignment into the analyses would shed light on more fine-grained translation strategies. This remains object to future research.

7 Conclusion and Future Research

In this paper, we have shown that the analysis of the sentence alignment of a parallel corpus of intralingual translations gives first indications of intralingual translation strategies. To summarise the results, we found evidences for information splitting and adding of information (mainly explanations), while there were only few evidences of reductions and deletions of information or a restructuring processes on a text level. In future research, these kinds of analyses could benefit from taking word or character counts into consideration to integrate adding, deletion, or restructuring processes within the sentence.

Like all branches of linguistics, also research on Easy Language translation profits from empirical corpus analyses. For a more fine-grained perspective, linguistically enriched corpora are still a desideratum. Especially translation phenomena resulting from translation selection and deselection strategies need—apart from a lexical analysis—a more detailed linguistic enrichment in order to be answered. This linguistic enrichment should cover syntactic functions, phrase structure and dependency trees but also semantic annotation such as coreference relations (cf. Kunz 2010) or frame semantics (cf. Czulo 2017), which requires of course more comprehensive annotation methods and more sophisticated query facilities.

While state-of-the-art corpus and treebank research in Standard and Easy Language share several problems such as multi-layer annotation and exploitation, they are also different in a number of ways. Monolingual parallel corpora including Easy Language texts raise numerous problems which are specific to this kind of research and which are caused by the specificities of this language variety—especially when it comes to the alignment of source and target texts. This includes the following problems: the lacking comparability of segments, the repetition of concepts and the differences in lexical and idea density. Future research has to address exactly these issues aiming at a comprehensive treebank including Easy Language translations.

Another research gap affects the empirical analysis of text-image integration in Easy Language translations. In our corpus analyses, we have so far not taken into account the role of images, pictures, pictograms, etc. In future research, it would be interesting to investigate in-depth the use of the pictures and images in Easy Language texts. According to Maaß (2015: 143), central concepts of a text can be illustrated by pictures. However, using pictures only to make the texts more appealing should be avoided since there is also the risk of causing distraction and cognitive overload (Bock 2018). Hence, an assessment if the pictures are solely used to make the text more understandable would be desirable. Quantifiable conclusions thereof still remain a research desideratum.

Finally, we will continue to enlarge our corpus and align the data so future analysis can draw on a larger data set. Further, the aligned data, for example, can be used to train machine translation systems for intralingual translation (Hansen-Schirra et al. 2020a).

References

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2017. Unsupervised neural machine translation. arXiv preprint http://arxiv.org/abs/1710.11041.

Battisti, A., S. Ebling, and M. Volk. 2019. An empirical analysis of linguistic, typographic, and structural features in simplified German texts. In *Proceedings of the sixth italian conference on computational linguistics (CLiC-it 2019)*. Bari.

Baker, M. 1995. Corpora in translation studies: An overview and some suggestions for future research. *TargetInternational Journal of Translation Studies* 7 (2): 223–243.

Bock, B.M. 2018. *Leichte Sprache: Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSAProjekt*. Technical report, Universität Leipzig.

Bock, B.M. 2019. *„Leichte Sprache". Kein Regelwerk.* Berlin: Frank & Timme.

Bredel, U., and C. Maaß. 2016a. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Berlin: Duden.

Bredel, U., and C. Maaß. 2016b. *Ratgeber Leichte Sprache*. Berlin: Duden.

Čmejrek, M., J. Cuřín, J. Havleka, J. Hajič, V. Kuboň. 2004. Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In *4th International conference on language resources and evaluation*. Lisbon, Portugal.

Chesterman, A. 2007. What is a unique item? In *Doubts and directions in translation studies*, ed. Y. Gambier, M. Shlesinger, and R. Stolze, 3–13. Amsterdam: Benjamins.

Czulo, O. 2017. Aspects of a primacy of frame model of translation. In *Empirical modeling of translation and interpreting*, ed. S. Hansen-Schirra, O. Czulo, Oliver and S. Hofmann, 465–490. Berlin: Language Science Press.

Deilen, S. 2020. Visual segmentation of compounds in easy language: Eye movement studies on the effects of visual, morphological and semantic factors on the processing of German Noun-Noun compounds. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 241–256. Berlin: Frank & Timme.

Gutermuth, S. 2020. *Leichte Sprache für alle? Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*. Berlin: Frank & Timme.

Hansen-Schirra, S., S. Neumann, and E. Steiner. 2012. *Cross-linguistic corpora for the study of translations: Insights from the language pair English-German*. Berlin, New York: de Gruyter.

Hansen-Schirra, S., S. Neumann, and E. Steiner. 2013. *Cross-linguistic corpora for the study of translations*.Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110260328

Hansen-Schirra, S., S. Neumann, O. Čulo, and K. Maksymski. 2017. Empty links and crossing lines: Querying multi-layer annotation and alignment in parallel corpora. In *Annotation, exploitation and evaluation of parallel corpora: TC3 I*, ed. S. Hansen-Schirra, S. Neumann, and O. Čulo, 53–87. Berlin: Language Science Press.

Hansen-Schirra, S., and S. Gutermuth. 2018. Modellierung und Messung Einfacher und Leichter Sprache. In *Barrieren abbauen, Sprache gestalten*, ed. S. Jekat, M. Kappus and K. Schubert, 7–23. Winterthur: ZHAW.

Hansen-Schirra, S., J. Nitzke, S. Gutermuth, C. Maaß, and I. Rink. 2020a. Technologies for the translation of specialised texts into easy language. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 99–127. Berlin: Frank & Timme.

Hansen-Schirra, S., W. Bisang, A. Nagels, S. Gutermuth, J. Fuchs, L. Borghardt, S. Deilen, A.K. Gros, L. Schiffl, and J. Sommer. 2020b. Intralingual translation into Easy Language—or How to reduce cognitive processing costs. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 197–225. Berlin: Frank & Timme.

Inclusion Europe. 2009. *Informationen für alle. Europäische Regeln, wie man Informationen leicht lesbar und verständlich macht.* http://easy-to-read.eu/wp-content/uploads/2014/-12/DE_Information_for_all.pdf

Kenny, D. 1997. (Ab)normal Translations: a German-English Parallel Corpus for Investigating Normalization in Translation. In *Practical applications in language corpora. PALC '97 proceedings*, ed B. Lewandowska-Tomaszczyk, and P.J. Melia, 387–392.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel. 2014. The sketch engine: Ten years on. *Lexicography* 1: 7–36.

Klaper, D., S. Ebling, and M. Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *Proceedings of the ACL workshop on predicting and improving text readability for target reader populations*, 11–19. Madison/USA: Omnipress.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit* 5: 79–86.

Kunz, K. 2010. *English and German nominal co-reference: A study of political essays*. Frankfurt/Main: Peter Lang.

Lehrndorfer, A. 1996. *Kontrolliertes Deutsch: Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation.* Tübingen: Narr.

Lindholm, C., and U. Vanhatalo (eds.) Forthcoming. *Easy language in Europe*. Berlin: Frank & Timme.

Maaß, C. 2015. *Leichte Sprache. Das Regelbuch*. Münster: Lit.

Maaß, C. 2020. *Easy language—plain language–easy language plus*. Berlin: Frank & Timme.

Maaß, C., and S. Hernández Garrido. 2020. Easy and plain language in audiovisual translation. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra, and C. Maaß, 131–161. Berlin: Frank & Timme.

Maaß, C., I. Rink, and S. Hansen-Schirra. Forthcoming. Easy language in Germany. In *Easy language in Europe*, ed. C. Lindholm and U. Vanhatalo. Berlin: Frank & Timme.

Netzwerk Leichte Sprache. 2013. *Leichte Sprache. Ein Ratgeber, Bundesministerium für Arbeit und Soziales.* Paderborn: Bonifatius GmbH.

Pickering, M.J., and S.C. Garrod. 2004. Toward a mechanist psychology of dialogue. *Behavioral and Brain Sciences* 27: 169–226.

Sahlgren, M., and J. Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering* 11 (3): 327–341.

Schiffl, L. 2020. Hierarchies in lexical complexity: Do effects of word frequency, word length and repetition exist for the visual word processing of people with cognitive impairments? In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 227–239. Berlin: Frank & Timme.

Sennrich, R., G. Schneider, M. Volk, and M. Warin. 2009. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology* 115: 124.

Sommer, J. 2020. A study of negation in German easy language—does typographic marking of negation words cause differences in processing negation? In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 257–272. Berlin: Frank & Timme.

Volk, M., T. Marek, and Y. Samuelsson. 2011. Building and querying parallel treebanks. *Translation: Computation, Corpora, Cognition (Special Issue on Parallel Corpora: Annotation, Exploitation and Evaluation)* 1 (1): 7–28.

Zanettin, F. 2000. Parallel corpora in translation studies: Issues in corpus design and analysis. *Intercultural Faultlines*, 105–118.

Silvia Hansen-Schirra Professor for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany. Director of the Translation & Cognition (TRA&CO) Center, Head of the Research Group "Simply complex—Easy Language".

Jean Nitzke Associate Professor for Translation with a focus on Translation Technology at the University of Agder, Norway. Former Lecturer and Post-Doc for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany.

Silke Gutermuth Lecturer for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany. Research Assistant at the Translation & Cognition (TRA&CO) Center, Manager of the on-site eyetracking lab.

Making Sense of the Prefix de- with an English–Chinese Parallel Corpus

Vincent Xian Wang

Abstract This study uses a large-scale English–Chinese parallel corpus to examine the senses of the prefix de- in English. Our investigation was designed in three main steps. The most commonly used de- verbs were first identified in the British National Corpus, their corresponding Chinese lexical items were manually collected from the English–Chinese parallel corpus, and using these Chinese items, we were able to identify the recurring Chinese characters and words that correspond to de- verbs in English, i.e. their translation candidates. The Chinese characters and words served as effective means to tease out five major sense groups entailed by the prefix de-. The results underscore the value of parallel corpora not only for sense disambiguation and translation candidate retrieval but also for contrasting typologically distant languages in terms of their morphological as well as alternative means for conveying meaning.

Keywords Prefix de · Sense disambiguation · Parallel corpus · Translation candidates · Lexicography

1 Introduction

This study uses a large scale English–Chinese parallel corpus to disambiguate the sense of the prefix de-. We attempt to extend the scope of sense disambiguation from words to morphemes. Our interest in studying morphemes also relates to contrastive language studies and translation studies.

First, derivational morphology marks a sharp difference between English and Chinese. The English language has a large repertoire of prefixes and suffixes of Latin, Greek and Germanic origins, which can be systematically and effectively used to derive new words from the existing ones and convey sophisticated mean-

V. X. Wang (✉)
Department of English, University of Macau, Avenida de Universidade, Taipa, MACAU, Macau SAR
e-mail: vxwang@um.edu.mo

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_12

ings by individual words (cf. Lightner 1983; Taylor 2014). By contrast, modern Chinese has a humble repertoire of (quasi-) derivational morphemes (cf. Arcodia 2012: 93ff), which entered into the Chinese language largely under the influence of other languages. Even there are some (quasi-)affixes corresponding to the derivational morphemes in English, e.g. 非- *fēi* ‘non-, a- ‘(ibid: 187), -化 *huà* ‘-ize, -ify’ (ibid: 110), they are much less productive compared with the ones in English. The Chinese language uses other means such as word compounding—which can undergo the process of clipping and truncation though—to convey complex lexical meanings. In short, English has ample flexibility to articulate complex meaning with derivational morphology, while Chinese needs to borrow and develop (quasi-) affixes or combine words to form compounds to convey the meaning. At this juncture, English–Chinese parallel corpora provide valuable language materials that showcase the ways in which the meaning expressed by derivational morphology in English is conveyable by similar and alternative means in Chinese.

Second, English–Chinese corpora can be considered a “goldmine” for harvesting translation candidates that are useful to translation practice and translation studies. Insomuch as English affixes are concerned, the conventional resources—i.e. monolingual and bilingual dictionaries and glossaries—provide rather concise classifications of the senses of the affixes and a small number of examples to show their usage (cf. Sect. 2.1). By contrast, sizable parallel corpora offer a broad range of examples used in authentic communication contexts, which translators and translation scholars can draw on as tangible resources to tackle their interested problems. For example, one can address the issue of (un)translatability at the morpheme level between English and Chinese, two typologically distant languages.

The present study focuses on a highly productive prefix de- in English, whose core meanings require various Chinese lexical items to convey, because there is not a single item in Chinese that corresponds to de- in English. Following previous studies that effectively captured word meanings using the translation data (see Sect. 2.1), we draw on evidence from a large English–Chinese parallel corpus (“UM-Education”: see Sect. 3.2) to describe the meanings that de- conveys in terms of sense groups. Our rationale is that the meaning of the prefix de- manifests in the commonly used de- words, which correspond to various Chinese words and expressions in the parallel corpus that can eventually be used to capture and depict the senses of the prefix de-. The present study, therefore, attempts to answer the following research questions:

(a) To what extent the English–Chinese parallel corpus is effective for identifying the major senses of the prefix de-?
(b) To what extent the parallel corpus can serve as a resource for harvesting translation candidates in Chinese for the prefix de-?
(c) According to the translation candidates in Chinese, does the prefix de- tend to be rendered into Chinese as affixes or words?

2 Relevant Studies

In this concise literature survey, we survey the most relevant studies on (a) using parallel corpora for sense disambiguation and (b) parallel corpus-based contrastive studies on affixation.

2.1 *Sense Disambiguation Using Translated Texts*

Parallel corpora have been used in a growing body of (contrastive) language studies and translation studies. One of the areas closely related to the present study is sense disambiguation using translated text. Johansson (2007: 28) noted that (word) meaning tends to be more accessible in translated texts—i.e. in parallel/multilingual corpora—than in monolingual corpora, because the former "make meanings visible through translation". Johansson cited the example of Norwegian adverbs that exhibit notable differences from each other in terms of the meaning/s conveyed by their translation correspondents (i.e. their translated forms) in English (ibid: 29–30). In the same vein, Mauranen (2002) used a parallel corpus that consists of Finnish translations of English fiction to disambiguate the lemma "think". For Mauranen, the translation equivalents of "think" in Finnish presented materials far richer than any bilingual dictionaries, with which she identified a broad range of senses conveyed by "think", e.g. mental processing, belief, hedges, see as, think up, remember, consider, suspect (ibid: 169). Apart from the studies using parallel corpora involving Indo-European languages, Lim (2018) investigated the English lemma "braise" in a parallel corpus of dish names in both English and Chinese. Lim gathered a large range of Chinese words and expressions that correspond to the lemma "braise". The corresponding Chinese lexical items indicate that the lemma "braise" can mean a wide range of cooking methods and cooked foods in Chinese cuisines, an illustrative example of sense disambiguation using languages of two distance language families—i.e. English of Indo-European and Chinese of Sino-Tibetan—in the culinary domain. In recent years, corpus-based or corpus-assisted sense disambiguation at the word level using Chinese–English parallel or comparable corpora can be found in a growing body of research (e.g. Huang and Wang 2020; Li et al. 2020a, b; Lim 2019; Lu et al. 2019; Wang et al. 2017; 王宪 2018).

Corpus-based studies on sense disambiguation primarily look at words and their meanings (e.g. Kilgarriff 1997, 2007). Some other studies set out to examine certain *syntactic* structures using the source and the translated texts, which are eventually narrowed down by gathering and analysing a (manageable) number of representative lexical items that exemplify the syntactic structure in question. Such studies, which are not uncommon to see, turn out to be the investigation of numerous selected lexical items and their meanings and linguistic properties. For example, Xu and Li's (2014) study on Chinese splittable compounds (SCs: 离合词 *líhé cí*) used

22 SCs as the representative items, which were searched in their Chinese–English parallel corpus and scrutinised to determine their semantic meanings and grammatical properties. Apart from the corpus-based studies on word senses and syntactic structures, the investigator was unable to find previous studies that disambiguate the senses of *morphemes* using translated texts other than Lim's (2021 in this volume) study on the suffix -ism.

2.2 Research on Morphemes with Translation Corpora

Of the studies on the use of morphemes across languages using large-scale translation data (cf. Lefer and Grabar 2015; Quah 1999), Cartoni and Lefer's (2011) study can be a good exemplar. Cartoni and Lefer (ibid) attempted to gain a bird's eye view of negation morphemes in three Indo-European languages—i.e. English, French and Italian—using large parallel corpora of millions of sentence pairs to examine the translation of the words that contain the negation morphemes. They discovered that the Romance prefixes in English—i.e. "de-", "dis-", "non-", "in-"—are mainly rendered into negation morphemes in French and Italian, while the Germanic affixes—the suffix "-less" and the prefix "un-"—are largely paraphrased in the translations. The results strongly suggest that language family plays an important role in the translatability of the morphemes, given that French and Italian are both Romance languages. There is still a real need to conduct studies that look at translatability of the affixes of Indo-European languages into languages of other language families, e.g. Sino-Tibetan languages.

Apart from the parallel corpus-based morphological studies on Indo-European languages, Quah (1999) is an exceptional study that investigates Malay, an Austronesian language into which is a large repertoire of English words containing Greco-Latin affixes needed to be translated. Following the independence of Malaysia in 1957, the Malay language needed to "absorb" a large number of scientific and technical words of English so that the language for teaching can shift from English to Malay. Quah built a small parallel corpus of academic texts, with which she identified several major translation methods—e.g. borrowing and adaptation, coining with an orthographic spelling resembling that of the English word, using additional words for rendering suffixes because Malay is mainly a prefixal language (ibid: 612). Malay uses alphabets and is rather ready to borrow or adopt English words and affixes with small changes. However, translating affixes from English into *Chinese* evinces a drastically different scenario, in which much different translation methods are expected. This is the area the present study aims at gathering more empirical evidence.

3 Method

We carried out our investigation in three main steps (cf. Sects. 3.1, 3.2 and 3.3), which are explained in this section. We focus on *verbs* that contain the prefix de-, because de- verbs present a major part of speech from which various other parts of speech can be derived, for example, nouns (e.g. derailment, depression, defamer) and adjectives (e.g. depressive, depressed, defamatory). In addition, we did not extend our investigation to the related prefixes such as dis- and des-. The attempt in this study is, therefore, to capture some major trends of de- words in terms of the senses they entail and their Chinese correspondents, rather than covering a comprehensive range of the parts of speech.

3.1 List of de- Words

We first identified the commonly used de- words using the British National Corpus (BNC), since it is a balanced corpus that covers a variety of text types and genres, consisting of both written and spoken texts. We accessed BNC at Sketch Engine (SkE) and ran a wordlist search of verbs starting with “de”. This led to a wordlist of 744 verbs, in which we removed noises and the ones that do not contain the bound morpheme de-, e.g. “deal”, “deem”, “deepen”. We are most interested in the de-words in which the stem is a free morpheme, e.g. the stem “code” in “decode”, because the meaning given by the prefix de- to the word tends to be clearer, compared with the meaning contributed by de- to the words in which the stem is a bound morpheme, e.g. “develop”, “desire”. The list of de- verbs are presented in Table 1 (cf. Sect. 4.1), ordered by the relative frequency. Based on the 37 top de-verbs identified in Table 1, we proceeded to the next step.

3.2 Collecting Corresponding Chinese Items for the Top de- Verbs

We searched all the de- verbs listed in Table 1 one at a time in the parallel corpus “UM-Education” to collect their corresponding words and expressions in Chinese. The UM-Education corpus contains 450,000 pairs of sentences in both English and Chinese, which has been used in similar studies for retrieving translation candidates (Tian et al. 2014, also see Lim 2021 in this volume). Table 2 (cf. Sect. 4.2) presents the top de- verbs with Chinese correspondents added. The Chinese correspondent items were manually collected from the Chinese concordance lines, while we made use of the function of Frequency of keyword in context (KWIC) at SkE and also the Find-on-page feature of the web browser from time to time to gain more efficiency

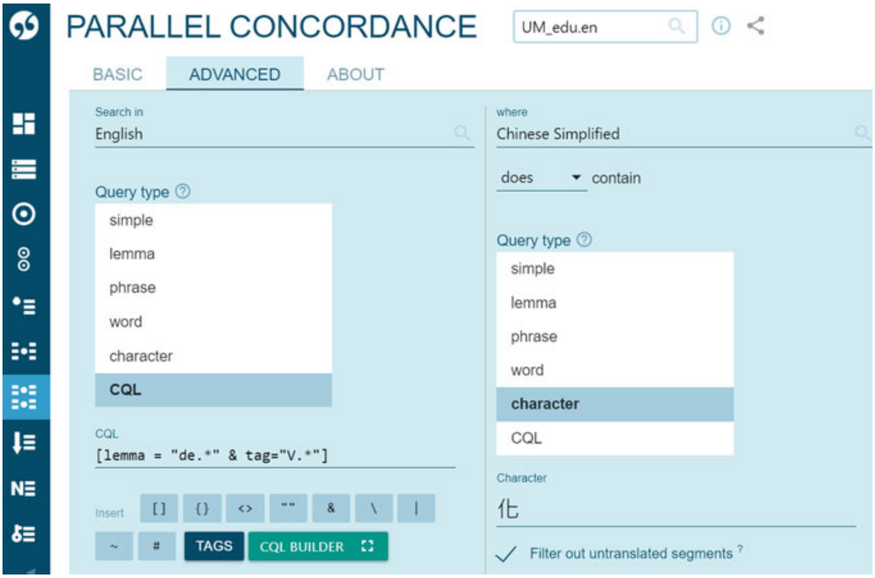
Table 1 List of de- verbs that contain the stems as free morphemes: top 37 items ordered by the relative frequency

de- verbs	Frequency	Relative frequency
detail	789	7.02296
devalue	394	3.50703
defuse	272	2.4211
decode	256	2.27868
degenerate	237	2.10956
decompose	208	1.85143
decentralise	192	1.70901
deform	177	1.57549
demean	146	1.29956
decommission	112	0.99692
deregulate	108	0.96132
dehydrate	106	0.94352
demoralize	97	0.86341
debase	94	0.8367
delimit	94	0.8367
debug	94	0.8367
deface	91	0.81
defrost	91	0.81
destabilize	76	0.67648
debrief	73	0.64978
decentralize	71	0.63198
derail	68	0.60527
defile	60	0.53407
demystify	55	0.48956
demobilize	55	0.48956
destabilise	50	0.44505
debunk	40	0.35604
deconstruct	39	0.34714
dethrone	37	0.32934
dehumanise	36	0.32044
defame	36	0.32044
depolarize	35	0.31154
detoxify	35	0.31154
demilitarize	29	0.25813
decaffeinate	26	0.23143
dehumanize	26	0.23143
depopulate	25	0.22253

in this laborious task. The Chinese items collected not only assist to reveal the meaning conveyed by the de- verbs but also serve as potential translation candidates for the de- verbs.

3.3 *Chinese Characters and Words Mapped to the Senses of the Prefix de-*

We constructed a specific corpus that consists of all the Chinese correspondents of the top de-verbs, intending to identify the most frequently occurring Chinese characters used for translating the de-words. Based on the Chinese character list ordered by frequency, we were able to identify the most frequently used Chinese characters that potentially correspond to de- words, e.g. 化 huà “dissolve, melt”, 解 jiě “undo, release”, 分 fēn “split, divide”, 毁 huǐ “destroy, damage”, 除 chú “remove, wipe out”. We referred to the usage of these Chinese characters in the words in Table 3, and searched each of them in combined queries in the parallel corpus UM-Education. For example, in the Parallel Concordance search interface at SkE, we simultaneously made a CQL (corpus query language) search of [lemma = “de.*” & tag = “V.*”] (to retrieve verbs beginning with “de”) in the English corpus and a character search of 化 in the Chinese corpus (see Fig. 1). The combined search returned 1,799 pairs of concordance lines, from which we manually collected Chinese words that serve as useful translation candidates for the corresponding de-words in English, e.g. 化解 huàjiě “dissolve, settle” for translating “defuse”, and 化开 huàkāi “melt away” for rendering “defrost” (cf. Table 3, Sense 2). A careful and systematic search of the Chinese characters in the parallel corpus UM-Education is time-consuming. However, we were able to identify a rich and extensive repertoire



Search in
English

Query type
CQL

CQL
[lemma = "de.*" & tag="V.*"]

where
Chinese Simplified

does contain

Query type
character

Character
化

✓ Filter out untranslated segments

Fig. 1 A combined search in both the English and the Chinese sub-corpora of the UM-Education corpus in the Parallel Concordance interface of SkE

of Chinese words useful to translate the de- verbs in English. The results are presented in Table 3 (in Sect. 4.3), which is particularly lengthy.

In summary, the three steps enabled us to identify the top de- verbs in English, retrieving their corresponding Chinese words and expressions in the parallel corpus and using these Chinese correspondents as "seeds" in the subsequent searches. From the analysis of these Chinese items that correspond to the de- verbs, we identified a list of Chinese characters that tend to translate de- words in English. These Chinese characters further allowed us to systematically retrieve the Chinese words that contain these characters and correspond to the de- verbs in the parallel corpus. The results depict the senses of de- words in the light of their corresponding Chinese characters and words.

4 Results and Discussion

The main results of our investigation obtained in the three steps are presented and briefly discussed in this section.

4.1 List of de- words

Of all the de- verbs collect from BNC, we identified the ones that contain the stem as a free morpheme and tabulate them with the information of frequency and relative frequency. Table 1 lists the top 37 items ordered by their frequencies of occurrence: the frequencies range from 789 for "detail" to 25 for "depopulate". We can observe that a small number of the most frequently occurring de- verbs—e.g. from "detail" (number 1) to "dehydrate" (number 12)—account for a notably large portion of the total use of de- verbs. The top 37 de- verbs were further examined in the parallel corpus "UM-Education" (cf. Sects. 2 and 3).

4.2 Chinese Lexical Items Corresponding to the Top de- Verbs

We queried the top de- verbs of Table 1 in the parallel corpus UM-Education (cf. Sect. 3.2) and gathered translation candidates in Chinese for each de- verb. The Chinese translation candidates are presented in Table 2. In the context of this study, we place more emphasis on the translation candidates that literally convey the meaning of the prefix de- than those that tend to convey the meaning figuratively, implicitly, or metaphorically. Also, we are more interested to figure out the ways in which the Chinese lexical items express the meaning of the prefix de- than the ways in which the stems are rendered into Chinese. For the conciseness of Table 2, the Chinese words and expressions are only given pinyin to illustrate their pronunciation, while their gloss translation in English is not added. A large portion of these Chinese words reoccurs in Table 3, in which both pinyin and gloss translations are provided. From the Chinese words and expressions in Table 2, we can spot certain tendencies for rendering de- words into Chinese, e.g. 贬 biǎn "belittle, devalue", 破 pò "break, crack (a code)". This enables investigators' to formulate hypotheses, and paves the way for more systematic studies of the Chinese characters and words that tend to be used to render de- verbs (see Sect. 4.3). In addition, Table 2 demonstrates some translation candidates used in previous translation tasks for a range of commonly used de- verbs. The information can be useful for translators in their translation assignments. The parallel corpus UM-Education, therefore, presents a rich resource for retrieving a reasonably large variety of translation candidates for de- words.

In terms of the use of translation strategies when there is no readily useable word-for-word translation, two main strategies are observed at work in our data. The first strategy pertains to paraphrasing with related words (cf. Baker 2018: 38). For example, "debug" can be rendered into Chinese as 查错 chacuo (cf. Table 2), meaning check for, detect, or spot something that goes wrong. The translated expression 查错 paraphrases the meaning of "debug" in the context in which to check for what goes wrong also implies to remove the wrong, which is the purpose of checking. 查错 is a related expression to "debug" because 错 cuo meanssomething that goes wrong, a part of the meaning in "debug".

By contrast, the second strategy is about paraphrasing using unrelated words (Baker 2018: 40). For example, "debug" can be rendered as 调试 tiaoshi "(fine-) tune/adjust and test" (cf. Table 2), a paraphrasing translation that explicitly indicates what debugging (a machine) is about in the (equipment) industry, although it appears that the translation 调试 is not a related word to "debug". The paraphrasing translations of both strategies were collected in our investigation.

4.3 Chinese Characters and Words Mapped to Different Senses of the Prefix de-

The major finding of the present study is displayed in Table 3. Using the most frequently used Chinese characters and words in Table 2 as "seeds", we were able to systematically query each of the Chinese characters in the Chinese sentences of the UM-Education corpus, while requiring that the corresponding English sentences contain a de– verb (cf. Fig. 1 in Sect. 3.3). We, therefore, identified the de- verbs that correspond to the Chinese character in question and retrieved the Chinese words or expressions that potentially translate the de- verbs. By closely examining the Chinese characters and words and the corresponding de- verbs in English in

Table 2 The top de- verbs and their Chinese correspondents in the UM-Education corpus

de- verb	Translation candidates in Chinese and *pinyin*
detail	详述, 详细阐述 /指出 /列出 / 描述 /介绍, 细节 *xiáng shù, xiángxì chǎnshù/zhǐchū/lièchū/miáoshù/jièshào, xìjié*
devalue	贬值, 贬低 *biǎnzhí, biǎndī*
defuse	化解, 拆除, 避免, 解决, 缓和, 消除 *huàjiě, chāichú, bìmiǎn, jiějué, huǎnhé, xiāochú*
decode	破译,破解,译码 *pòyì, pòjiě, yì mǎ*
degenerate	退化, 衰退,陷入, 堕落, 变成 *tuìhuà, shuāituì, xiànrù, duòluò, biàn chéng*
decompose	分解, 毁灭, 降解 *fēnjiě, huǐmiè, jiàngjiě*
decentralise/ze	(权力) 分散(式),分权,分级 *(quánlì) fēnsàn (shì), fēnquán, fēnjí*
deform	变形, 畸形, 变丑, 变化, 残废, 残缺, 形变 *biànxíng, jīxíng, biàn chǒu, biànhuà, cánfèi, cánquē, xíngbiàn,*
demean	侮辱/有辱人格, 辱没, 贬低, 刻薄的, 丢/太没面子 *wǔrǔ/yǒu rǔ réngé, rǔmò, biǎndī, kèbó de, diū/ài méi miànzi*
decommission	退役, 废弃 *tuìyì, fèiqì*
deregulate	解除管制/对… 的监管, 放松管制 *jiěchú guǎnzhì/duì… de jiānguǎn, fàngsōng guǎnzhì*
dehydrate	脱水*tuōshuǐ*
demoralise/ze	挫折感, 挫伤士气, 士气低落, 沮丧, 灰心丧气, 丧失斗志, 腐蚀, *cuòzhé gǎn, cuòshāng shìqì, shìqì dīluò, jǔsàng, huīxīn sàngqì, sàngshī dòuzhì, fǔshí,* 泄气,精神上的不利, 意志消沉, 蜕化成道德败坏 *xièqì, jīngshén shàng de búlì, yìzhì xiāochén, tuìhuà chéng dàodé bàihuài*
debase	降低, 低俗, 贬损, 忽略, (被)败坏, 腐化 *jiàngdī, dīsú, biǎnsǔn, hūlüè, (bèi) bàihuài, fǔhuà*
delimit	分隔, 定界, 界, 限定, 划(入) *fēngé, dìng jiè, jiè, xiàndìng, huà (rù)*
debug	调试, 测试, 查错, 修正(臭虫) *tiáoshì, cèshì, chá cuò, xiūzhèng (chòuchóng)*
deface	毁损, 销毁, 损毁, 涂污 *huǐsǔn, xiāohuǐ, sǔnhuǐ, tú wū*
defrost	解冻, 除霜,化开 *jiědòng, chú shuāng, huà kāi*
destabilise/ze	破坏(其)稳定, 打破, 颠覆, 动摇, 摧毁 *pòhuài (qí) wěndìng, dǎpò, diānfù, dòngyáo, cuīhuǐ*
	非建设, (使…)不稳定,不安定 *fēi jiànshè, (shǐ…) bù wěndìng, bù āndìng*
debrief	(情况) 汇报(任务), 听取 … 汇报, 报告, 审讯 *(qíngkuàng) huìbào (rènwù), tīngqǔ… huìbào, bàogào, shěnxùn*
derail	出轨, 脱轨, 脱离, 破坏, 搁浅, 阻挠 *chūguǐ, tuōguǐ, tuōlí, pòhuài, gēqiǎn, zǔnáo*
defile	玷污, 玷污 污损, 污秽, 污染 *diànwū, diànwū wū sǔn, wūhuì, wūrǎn*

(continued)

Table 2 (continued)

de- verb	Translation candidates in Chinese and *pinyin*
demystify	排除神话, 让 … 浅显化 *páichú shénhuà, ràng… qiǎnxiǎn huà*
demobilize/se	复员, 离职 *fùyuán, lízhí*
debunk	揭穿, 揭露, 言明 *jiēchuān, jiēlù, yánmíng*
deconstruct	解构, 分解 *jiěgòu, fēnjiě*
dethrone	废黜, 拉下…宝座, 赶/撵下台, 摘掉…冠, 反驳, 打败, 落魄 *fèichù, lā xià…bǎozuò, gǎn/niǎn xiàtái, zhāi diào…guān, fǎnbó, dǎbài, luòpò*
dehumanise/ dehumanize	非人道, 去人性化, 毫无尊严, 丧失, 人性, 诋毁 *fēi réndào, qù rénxìng huà, háo wú zūnyán, sàngshī, rénxìng, dǐhuǐ*
defame	诽谤, 破坏…名誉, 名誉毁损, 诬告 *fěibàng, pòhuài…míngyù, míngyù huǐsǔn, wūgào*
depolarise/ze	去极化, 消退极化 *qù jí huà, xiāotuì jíhuà*
detoxify	解毒, 排毒 *jiědú, páidú*
demilitarise/ze	非军事化, 无戒备区 *fēi jūnshì huà, wú jièbèi qū*
decaffeinate	除去咖啡 *chúqù kāfēi*
depopulate	人口衰退, 人口稀少/减少 *rénkǒu shuāituì, rénkǒu xīshǎo/jiǎnshǎo*

terms of their semantic meaning, we classified the senses of the prefix de- into five sense groups, as they emerged from our data.

The first sense group of the prefix de- entails the meaning of removing, wiping out something, and also the dissolving or breaking apart of an entity. The typical example is 除去 *chúqù* "take away, remove", while the thing being removed can be caffeine as in "decaffeinate", frost in "defrost", ice in "deice" (see Sense 1a). If the unwanted thing cannot be totally removed, it can be damaged, destroyed or devalued, and this pertains to Sense 1b. Sense 1c is related to Sense 1b and denotes the breaking or cracking of a code, and, as a consequence, the code loses its original value or due function. Apart from Senses 1a to 1c, something can fall apart or dissolves by itself, and this pertains to Sense 1d. In this section, de- verbs that contain either bound or free stems were collected when retrieving them from the Chinese characters back to English, waiving the requirement that the de- verbs should be composed of a free stem imposed in Step 2 (cf. Sect. 3.2).

In Sense 2, the prefix de- denotes the reverse of a process that has taken place in the first place. For example, "defrost" is the reverse of the freezing process, and "decrypt" is the reverse process of encrypting. The Chinese characters 解 *jiě* "undo, untie", 化 *huà* "melt, resolve", 分 *fēn* "split" and 退 *tuì* "retreat, retire" capture the meaning of reversing a process rather clearly and vividly.

Sense group 3 denotes moving (down), growing (weak), or changing (form). The de- verbs such as "decay", "decline" and "deflate" and the corresponding Chinese characters 下 *xià* "(go) down", and 降 *jiàng* "decrease, descend" express the

Table 3 The major senses of the prefix de- denoted by Chinese characters and words

Senses	Chinese characters, pinyin, gloss trans	Chinese words, pinyin, gloss translation	Corresponding de- verbs in English
1 remove, destroy, dissolve			
1a remove, cleanse, or eliminate			
	除 *chú* “remove”	除去 *chúqú* “take away”	decaffeinate
		除霜 *chúshuāng* “defrost”	defrost
		除鳞*chúlín* “descale”	descale
		除垢 *chúgòu* “descale”	descale
		除盐 *chúyuan* “desalt”	desalt
		除冰 *chúbīng* “de-ice”	deice
		除臭 *chúchòu* “de-odour”	deodorise
		除错 *chúcuò* “remove error”	debug
	脱 *tuō* “remove”	脱氧 *tuōyǎng* “de-oxygen”	deoxy
		脱水 *tuōshuǐ* “dehydrate”	dehydrate
		脱气*tuōqì* “de-gas”	degas
		脱碳 *tuōtàn* “de-carbon”	decarbonise
		脱去盐分 *tuōqù yánfèn* “de-salt”	desalinate
		脱色 *tuōsè* “de-colour”	decolour
		脱钙 *tuōgài* “de-calcium”	demineralise
	去 *qù* “(take) away”	去中心化 *qù zhōngxīn huà* “de-centralise”	decentralise
		去人性化 *qù rénxìng huà* “de-humanise”	dehumanise
		去离子 *qù lízǐ* “de-ion”	deionise
		去骨 *qùgǔ* “de-bone”	debone
		去除 *qùchú* “take away”	detangle
		去毒 *qùdú* “de-tox”	detoxicate
		去活性 *qù huóxìng* “remove vitality”	devitalise
		去氧 *qùyǎng* “deoxygen”	deoxy
	排 *pái* “exclude, eject, discharge”	排毒 *páidú* “de-tox”	detoxify
		排除 *páichú* “exclude, remove”	debug, demystify, detoxify
	拆 *chāi* “dismantle”	拆除*chāichú* “demolish”	deactivate, defuse, demolish
		拆焊 *chāihàn* “de-solder”	desolder

(continued)

Table 3 (continued)

Senses	Chinese characters, pinyin, gloss trans	Chinese words, pinyin, gloss translation	Corresponding de- verbs in English
	消 *xiāo* “cancel”	消除 *xiāochú* “eliminate, cleanse”	destigmatise, deflate, defuse, delete, decelerate
		消毒 *xiāodú* “disinfect”	decontaminate
		消气 *xiāoqì* “de-gas, de-air”	degas
		消磁 *xiāocí* “demagnetise”	degauss
		消灭 *xiāomiè* “wipe out”	destroy
	废 *fèi* “waste, stop”	废黜 *fèichù* “dethrone”	dethrone
	罢 *bà* “oust, stop”	罢黜 *bàchù* “depose”	depose
1b damage, destroy, abase			
	毁 *huǐ* “destroy”	毁损 *huǐsǔn* “damage, impair”	deface, defame, destroy
		毁灭 *huǐmiè* “destroy”	defoliate
	诋 *dǐ* “defame”	诋毁 *dǐhuǐ* “defame, slander”	dehumanise
	败 *bài* “undermine”	败坏 *bàihuài* “ruin”	debase, demoralise
	破 *pò* “destruct”	破坏 *pòhuài* “wreck”	destabilise, derail, defame, devastate
	损 *sǔn* “damage”	损毁 *sǔnhuǐ* “damage”	defame
	贬 *biǎn* “belittle”	贬低 *biǎndī* “belittle”	decry, demean, degrade
		贬损 *biǎnsǔn* “disparage”	debase
		贬值 *biǎnzhí* “devalue”	devalue
1c resolve a problem, interpret hidden information			
	解 *jiě* “interpret”	解译 *jiěyì* “decode, interpret”	decipher
		解密 *jiěmì* “interpret (the secret)”	decipher
		解读 *jiědú* “interpret”	decipher
		解密 *jiěmì* “un-classify”	declassify
	破 *pò* “break, crack”	破解 *pòjiě* “break (a code)”	decode, decipher
		破译 *pòyì* “break and interpret”	decode, decipher
1d fall/break apart, diminish, dissolve			
	分 *fēn* “split, separate”	分解 *fēnjiě* “decompose”	decompose
	裂 *liè* “crack, split open”	裂解 *lièjiě* “degrade and dissolve”	decompose
	消 *xiāo* “disappear, wear out”	消耗 *xiāohào* “deplete”	deplete

(continued)

Table 3 (continued)

Senses	Chinese characters, pinyin, gloss trans	Chinese words, pinyin, gloss translation	Corresponding de- verbs in English
		消解 *xiāojiĕ* “fade, degrade”	degrade
		消退 *xiāotuì* “recede”	decay, depolarise
	毁 *huĭ* “damage”	毁坏 *huĭhuài* “damage”	decay
	丧 *sang* “lose”	丧失 *sàngshī* “lose”	demoralise, dehumanise
	失 *shī* “lose”	失效 *shīxiào* “lose power”	deactivate
2 a reverse process			
	解 *jiĕ* “undo, untie”	解冻 *jiĕdòng* “de-freeze”	defrost
		解码 *jiĕmă* “decode”	decode
		解密 *jiĕmì* “decrypt”	decrypt
		解毒 *jiĕdú* “de-tox”	detoxify
		解决 *jiĕjué* “resolve, settle”	defuse
		解构 *jiĕgòu* “deconstruct”	deconstruct
		解除 *jiĕchú* “remove, release”	deregulate
		解压(缩) *jiĕyā (suō)* “de-compress”	decompress
		解耦 *jiĕŏu* “de-couple”	decouple
		解吸附*jiĕ xīfù* “desorb”	desorb
		解调 *jiĕtiáo* “demodulate”	demodulate
		解聚 *jiĕjù* “depolymerise”	depolymerise
	化 *huà* “melt, resolve”	化解 *huàjiĕ* “dissolve, settle”	defuse
		化开 *huàkāi* “melt away”	defrost
	分 *fēn “*split”	分权 *fēnquán* “decentralise power”	decentralise
		分散 *fēnsàn* “disperse”	decentralise
		分拆 *fēnchāi* “split, partition”	demerge
	退 *tuì* “retreat, retire”	退役 *tuìyì* “retire from service”	decommission

(continued)

Table 3 (continued)

Senses	Chinese characters, pinyin, gloss trans	Chinese words, pinyin, gloss translation	Corresponding de- verbs in English
3 move (downwards), grow weaker, change (form)			
3a move downwards			
	下 *xià* "(go) down"	下降 *xiàjiàng* "go down, decrease"	decay, decrease, decline, degrade, depressurise, depress, descend, deteriorate
		下跌 *xiàdiē* "decrease, decline"	deflate, decline
	降 *jiàng* "decrease, descend"	降低 *jiàngdī* "go down, decrease"	debase, decivilise, decrease
		降解 *jiàngjiě* "degrade, decompose"	degrade, decompose
		降级 *jiàngjí* "demote, degrade"	demote, degrade
3b move backwards			
	退 *tuì* "go back"	退化 *tuìhuà* "degenerate"	degenerate, deteriorate, degrade
3c move out of track			
	出 *chū* "go out"	出轨 *chūguǐ* "derail"	derail
		出列 *chūliè* "de-queue"	dequeue
3d grow weaker			
	衰 *shuāi* "decline"	衰退 *shuāituì* "degenerate, decline"	degenerate, deteriorate, depress, decline
		衰减 *shuāijiǎn* "decline, grow weaker"	decay, decrease, decline
		衰落 *shuāiluò* "decline, go down"	decay, decline
		衰老 *shuāilǎo* "grow old"	decelerate, defer, delay
		衰变 *shuāibiàn* "decay"	decay
		衰竭 *shuāijié* "deplete, exhaust"	deplete, decline
3e change (form)			
	变 *biàn* "change'	变形 *biànxíng* "change form"	deform
		变质 *biànzhí* "degenerate"	degenerate, degrade
		变性 *biànxìng* "change nature"	denature

(continued)

Table 3 (continued)

Senses	Chinese characters, pinyin, gloss trans	Chinese words, pinyin, gloss translation	Corresponding de- verbs in English
		变化 *biànhuà* “change”	deviate, deform
		变成 *biànchéng* “become”	degenerate
		变得 *biàndé* “become”	debilitate, devolve
		变丑 *biànchǒu* “become ugly”	deform
		变小 *biànxiǎo* “become smaller”	decrease, deflate
		变干 *biàngān* “become dryer”	desiccate
4 negation (of a verb)			
	不 *bù* “non-”	不安定 *bù āndìng* “not settle”	destabilise
		不得 *bùdé* “not be able to”	desist
		不再*búzài* “no more”	desist
	非 *fēi* “non-”	非军事化 *fēi jūnshì huà* “not militarise”	demilitarise
		非建设性*fēi jiànshè xìng* “not construct”	destabilise
		非人道 *fēi réndào* “not humanise”	dehumanise
	无 *wú* “non-”	无戒备 *wú jièbèi* “not guarded”	demilitarise
5 intensifying a verb			
	定 *dìng* “determine”	定界 *dìngjiè* “determine (the) boundary”	delimit
		定名 *dìngmíng* “determine (the) name”	denominate
	计 *jì* “calculate”	计价 *jìjià* “determine (the) price” 计值 *jìzhí* “determine (the) value”	denominate denominate

Note Only UK English spelling is used in this table, while US spelling is not specified. So the word form like “devitalise” stands for both “devitalise” and “devitalize” downwards movement. The tendencies of moving toward a decline or in terms of change (of form or of nature) are depicted by Chinese characters 衰 *shuāi* "decline" and 变 *biàn* "change" (see Senses 3d and 3e), with the corresponding de- verbs such as "degenerate", "deplete", "deform", "debilitate" and "denature".

In Sense 4, de- serves as a device for denoting negation—e.g. "dehumanise" is the negated form of "humanise" and "destabilise" the negated form of "stabilise". The negation is rather explicitly expressed by the corresponding Chinese characters 不 *bù* "non-", 非 *fēi* "non-", and 无 *wú* "non-", which are the commonly used negation devices in Chinese. However, the prefix de- appears to be far more productive than the three Chinese characters do—i.e. 不, 非, 无—in terms of the frequency of occurrence and the power to associate with other lexical items and create new words. This is probably one of the main reasons explaining why de-verbs in English are mainly translated into Chinese by words and expressions other than those negators—i.e. 不, 非 and 无.

Finally, de- in Sense 5 is used to form a verb that intensifies the meaning of an existing verb. For example, the verb "delimit" is formed based on the existing verb "limit", and the meaning of "delimit" is well captured by its corresponding Chinese item 定界 *dìngjiè* determine (the) boundary. Similarly, "denominate" is created out of the verb "nominate", while its translation candidate in Chinese 计价 *jìjià* "determine (the) price" reflects the meaning of "denominate" rather closely.

The five sense groups are laid out to capture the major categories of the semantic meaning entailed by the prefix de-. Since the sense groups are not entirely exclusive one from another, some overlaps are possible between them. For example, the sense of "abase" in Sense 1b may also be interpreted as "moving downwards" in Sense 3a, while the sense of "breaking a code" in Sense 1c can be taken as "a reverse process" as well, i.e. Sense 2. However, possible overlaps of (sub-)senses certainly do not override the necessity to distinguish major sense groups of the prefix de-, which can serve as valuable information for translators, language learners and lexicographers.

Compared to the prefix de-, its corresponding Chinese lexical items in Table 3 are mostly words rather than prefixes. Most of the Chinese items can be used as separate verbs by themselves, e.g. 除 *chú* "remove", 脱 *tuō* "remove", 除去 *chúqù* "remove/wipe away", 化 *huà* "melt, resolve", whereas only the ones in Sense 4 exhibit clearer properties of bound morphemes. Of the items in Sense 4, 非 *fēi* "non-" in particular can be considered a representative "prefix-like formative" in Chinese (Arcodia 2012: 188), a negator that contributes to word formation and "has lost its free status in Modern Mandarin Chinese" (ibid: 190). Apart from those items in Sense 4, the Chinese lexical items that correspond to de- verbs are predominantly (compound) words and expressions composed of two (sometimes more) Chinese characters that are free morphemes. Our results on the prefix de- reveal the tendency that morpheme-to-morpheme rendition from English to Chinese accounts only for a low proportion, while the predominant instances evidence morpheme-to-word translation. At this juncture, the words gathered in Table 3 not only serve as resources for practising translators but may also be informative to researchers working on contrastive language studies.

5 Concluding Remarks

The present study attests to the value of using a large-scale parallel corpus to tease out the senses of the prefix de-. Five main sense groups of the prefix de- emerged in our data. It can entail the meaning of removal (Sense 1), a reverse process (Sense 2) and making (downward) movement or changes (Sense 3), while it can also function as a negation device (Sense 4) and a means to create a new verb from an existing verb (Sense 5 on intensification). Our study casts light on the situation of translating from a morphologically rich language (English) into a morphologically impoverished language (Chinese). This area has not been much researched, in contrast with the extensive body of contrastive morphological studies on Indo-European languages and a few studies on non-Indo-European languages such as Quah's (1999: cf. Sect. 2.1) investigation of Malay. However, Malay has been under the great influence of English and uses alphabets, exhibiting ample flexibility to borrow English words and affixes or to adapt them with minimal changes. This is in stark contrast with the situation of translating morphemes from English into Chinese, a primarily pictographic language, as examined in the present study. There is still much room in future studies to draw on parallel corpora to investigate two typologically distant languages for the interests of contrastive language studies and translation studies.

In terms of the idea of disambiguating word senses using parallel corpora (cf. Sect. 2.1), the present study not only underscores the feasibility of word sense disambiguation, but, more importantly, also extends, together with similar studies (e.g. Lim 2021 in this volume), the scope to sense disambiguation from words to *morphemes*, evidencing the power of parallel corpora in capturing the senses entailed by commonly used and highly productive derivational affixes. We have observed that different senses of the prefix de- tend to be rendered into different Chinese characters and words, and this significantly facilitates sense disambiguation and further enables the classification of the senses into major groups. Our results on two typologically different languages lend support to Johansson's (2007) observation that sense disambiguation tends to be less effective with monolingual corpora (cf. Sect. 2.1), while the task can be markedly facilitated using parallel/multilingual corpora. The merit of parallel corpora can be further tapped on in subsequent studies. The results of this study point to the implications on the training of translators, who can benefit from effective applications of corpus tools, while the empirical studies in this area are certainly much needed. In addition, our findings can be informative for second language learning as well as bilingual dictionary-making.

Methodologically, the present study demonstrates that a large-scale parallel corpus enables one to disambiguate the senses of de- words by their translated lexical items. The two subcorpora of the UM-Education corpus are aligned only at the sentence level, rather than at the word or the phrase level. The manual screening for the corresponding lexical items from aligned sentence pairs is time-consuming, although still possible and fruitful. This points to the need for technological advancement on alignment at a smaller-than-sentence level, which will tangibly benefit the investigations on parallel corpora. Finally, the three steps of investigation used in this study can be a reference for designing parallel corpus-based studies in future.

Acknowledgements The present author is especially grateful to the reviewers of the chapter for their valuable comments and suggestions and acknowledges the research grant (MYRG2019-00162-FAH) of the University of Macau.

References

Arcodia, G.F. 2012. *Lexical derivation in mandarin Chinese*. Crane Publishing Company.

Baker, M. 2018. *In other words: A coursebook on translation*, 3rd ed. Routledge.

Cartoni, B., and M.A. Lefer. 2011. Negation and lexical morphology across languages: insights from a trilingual translation corpus. *Poznan Studies in Contemporary Linguistics* 47 (4): 795–843. https://doi.org/10.2478/psicl-2011-0039 .

Huang, C.-R., and X. Wang. 2020. From faithfulness to information quality: On 信 in translation studies. In *Key issues in translation studies in China*, ed. L. Lim and D. Li, 111–142. Springer.

Johansson, S. 2007. *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. John Benjamins.

Kilgarriff, A. 1997. I don't believe in word senses. *Computers and The Humanities* 31 (2): 91–113.

Kilgarriff, A. 2007. Word senses. In *Word sense disambiguation*, ed. E. Agirre and P. Edmonds, 29–46. Springer.

Lefer, M.A., and N. Grabar. 2015. Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures* 16 (2): 187–206. https://doi.org/10.1556/084.2015.16.2.3

Li, L., S. Dong, and V.X. Wang. 2020a. Gaige and reform: A Chinese-English comparative keywords study. In *From minimal contrast to meaning construct: Corpus-based, near synonym driven approaches to Chinese lexical semantics*, ed. Q. Su and W. Zhan, 321–332. Springer/Peking University Press.

Li, L., C.R. Huang, and V.X. Wang. 2020b. Lexical variations and human behavior changes: A corpus-assisted investigation of gambling and gaming in the past centuries. *SAGE Open* 10 (3): 1–14. https://doi.org/10.1177/2158244020951272

Lightner, T.M. 1983. *Intrduction to English derivational morphology*. John Benjamins.

Lim, L. 2018. A corpus-based study of braised dishes in Chinese-English menus. In *Proceedings of the 32nd pacific asia conference on language, information and computation: 25th joint workshop on linguistics and language processing*, ed. S. Politzer-Ahles, Y.Y. Hsu, C.R. Huang, and Y. Yao, 887–892. Association for Computational Linguistics.

Lim, L. 2019. Are TERRORISM and *kongbu zhuyi* translation equivalents? A corpus-based investigation of meaning, structure and alternative translations. In *Proceedings of the 33rd pacific asia conference on language, information and computation*, ed. R. Otoguro, M. Komachi, and T. Ohkuma, 516–523.

Lim, L. 2021. A Corpus-based examination of the translation of the suffix -ism into Chinese. In *New perspectives on corpus translation studies*, ed. V.X. Wang, L. Lim, and D. Li, 29–55. Springer. https://doi.org/10.1007/978-981-16-4918-9_2

Lu, W., F. Meng, S. Wang, G. Zhang, X. Zhang, A. Ouyang, and X. Zhang. 2019. Graph-based Chinese word sense disambiguation with multi-knowledge integration. *Computers, Materials and Continua* 61 (1): 197–212.

Mauranen, A. 2002. Will ‘translationese’ ruin a contrastive study? *Languages in Contrast* 2 (2): 161–185.

Quah, C. 1999. Issues in the translation of English affixes into Malay. *Meta: Translators' Journal*, 44 (4): 604–616. https://doi.org/10.7202/003881ar

Taylor, J.R. 2014. *The Oxford handbook of the word*. Oxford University Press.

Tian, L., D.F. Wong, L.S. Chao, P. Quaresma, F. Oliveira, and L. Yi. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, ed. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, 1837–1842. European Language Resources Association (ELRA).

Wang, X., X. Tang, W. Qu, and M. Gu. 2017. Word sense disambiguation by semantic inference. In *2017 international conference on behavioral, economic, socio-cultural computing (BESC)*, 1–6. IEEE. https://doi.org/10.1109/BESC.2017.8256391

Xu, J., and X. Li. 2014. Structural and semantic non-correspondences between Chinese splittable compounds and their English translations: A Chinese-English parallel corpus-based study. *Corpus Linguistics and Linguistic Theory* 10 (1): 79–101. https://doi.org/10.1515/cllt-2013-0019

Reference in Chinese

Wang, V.X. 王宪 2018. 用语料库考察‘demonize’和‘妖魔化’的使用 (‘Demonize’ and the rise and use of *yaomohua* in Modern Chinese). In *“语言和现代化”学术研讨会论文集 (Proceedings of the symposium on language and modernisation) 2017* ed. 李向玉 Heong Iok Lei, 114–122. 澳门理工学院 (Macao Polytechnic Institute).

Vincent X. Wang an associate professor of the University of Macau and a NAATI-certified translator, received his MA and PhD in Applied Linguistics from the University of Queensland (2006). His research interests are in interlanguage pragmatics, corpus-based contrastive language studies, and discourse and pragmatics in translation. He published journal articles in *Sage Open*, *Target*, *Journal of Language, Literature and Culture* and TESOL-related periodicals, book chapters with Springer, Routledge and Brill, conference papers with PACLIC and CLSW, and a monograph *Making Requests by Chinese EFL Learners* (John Benjamins). His recent research draws on big data and corpus linguistics methodologies to investigate language properties, discourse and the use of conceptual metaphors in social events such as COVID-19.

Correction to: Probing a Two-Way Parallel T&I Corpus for the Lexical Choicesof Translators and Interpreters

Oi Yee Kwong

Correction to:
Chapter "Probing a Two-Way Parallel T&I Corpus
for the Lexical Choices of Translators and Interpreters"
in: V. X. Wang et al. (eds.),
***New Perspectives on Corpus Translation Studies*,**
New Frontiers in Translation Studies,
https://doi.org/10.1007/978-981-16-4918-9_5

The original version of the book was published with incorrect spell error in the abstract section, now corrections has been incorporated. The chapter and book have been updated with the changes.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-981-16-4918-9_5

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_13