Machine Translation: Technologies and Applications
Series Editor: Andy Way

Michael Carl *Editor*

Explorations in Empirical Translation Process Research

Springer

Machine Translation: Technologies and Applications

Volume 3

This book series tackles prominent issues in Machine Translation (MT) at a depth which will allow these books to reflect the current state-of-the-art, while simultaneously standing the test of time. Each book topic will be introduced so it can be understood by a wide audience, yet will also include the cutting-edge research and development being conducted today.

MT is being deployed for a range of use-cases by millions of people on a daily basis. Google Translate and Facebook provide billions of translations daily across many language pairs. Almost 1 billion users see these translations each month. With MT being embedded in platforms like this which are available to everyone with an internet connection, one no longer has to explain what MT is on a general level. However, the number of people who really understand its inner workings is much smaller.

The series includes investigations of different MT paradigms (Syntax-based Statistical MT, Neural MT, Rule-Based MT), Quality issues (MT evaluation, Quality Estimation), modalities (Spoken Language MT) and MT in Practice (Post-Editing and Controlled Language in MT), topics which cut right across the spectrum of MT as it is used today in real translation workflows.

More information about this series at http://www.springer.com/series/15798

Michael Carl
Editor

Explorations in Empirical Translation Process Research

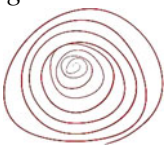
Springer

Editor
Michael Carl
Kent State University
Kent, OH, USA

ISSN 2522-8021 ISSN 2522-803X (electronic)
Machine Translation: Technologies and Applications
ISBN 978-3-030-69776-1 ISBN 978-3-030-69777-8 (eBook)
https://doi.org/10.1007/978-3-030-69777-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to Rose—for her sweet fragrance and her critical thorns.



Foreword

Over the past four decades, translation process research has gone a long way from the first think-aloud protocol studies carried out in the mid-1980s to present-day empirical studies that draw on large data sets and use computational tools to generate robust evidence that accounts for how the act of translation unfolds as a cognitive activity. Spearheaded through the triangulation paradigm (Alves 2003), translation process research has striven to build methodological approaches that allow empirical studies to be critically assessed and potentially replicated. Explorations of the translation unit and segmentation patterns, inquiries into user activity data (UAD) and alignment units, modeling translation entropy, among several other topics, have built a strong research agenda for translation process research with worldwide impact. The development of the database CRITT TPR-DB (Carl et al. 2016) allowed the storage and integration of translation process data in a large repository, enabling researchers to use a data pool to compare and extend empirical studies of translation process data. All these achievements have created a strongly motivated international research community of which I am proud to be a member.

In parallel to these developments in empirical research, the study of translation as a cognitive activity has also seen the emergence of a new line of research that considers human cognition, and indirectly the act of translating and interpreting, to be situated, embodied, distributed, embedded, and extended (Risku and Rogl 2020). This has created a different focus of inquiry and a separate research agenda. Authors affiliated to this alternative paradigm have challenged the standard computational-oriented and information processing views of translation process research and claimed that studies need to be placed in context and consider translation as embodied, embedded, and affective action even at the risk of lacking empirical validation.

At the same time, human-computer interaction (HCI) has gained prominence in translation, and improvements in machine translation (MT) systems have given an impetus to post-editing tasks as one of the most demanded translation-related activities. The impact of HCI in translation has led researchers to expand the agenda of empirical translation process research. As O'Brien (2020) states, the

merging of translation memories and MT, as well as the advent of adaptive and interactive neural MT (NMT) systems and the use of multimodal input, has brought about significant changes in translation task execution and created an impact on the process of translation. Further research is thus required to understand the implications of these new forms of translational activity.

Considering these disparate yet potentially complementary approaches, Alves and Jakobsen (2020, 548) have insisted that only by integrating these different trends into a coherent whole cognitive translation studies can lay the epistemological, paradigmatic, and interdisciplinary foundations for its development. It should ground itself "in theories of semiosis (meaning making) and linguistics (language use) and on cognitive science (neurocognition and situated-action cognition)." Therefore, for Alves and Jakobsen, cognitive translation studies must incorporate to its research agenda not only features of MT and aspects of HCI but also enlarge the scope of its theoretical formulations to include situated, distributed, and extended aspects of human cognition.

The current volume responds exactly to what Alves and Jakobsen have suggested and extends the scope of explorations in empirical translation process research to shed new light into the intricacies of translation as a cognitive activity. In 15 chapters, the book brings the discussion to a new level which moves away from a dichotomous separation of computational and noncomputational approaches in cognitive translation studies and offers an integrated alternative to clarify the relation between the translation process, the translation product, and machine-related activities in translation. The volume reports on translation experiments carried out in several language pairs and in different translation modes, using eye-tracking and keylogged data to inquire into the translation of cognates, neologisms, metaphors, idioms, disparate text types as well as figurative and culturally specific expressions. It also introduces innovative measurements and sound advances in data treatment and analyses.

In my understanding, the volume excels, above all, by attempting to bridge the gap between representational and nonrepresentational views of human cognition and by offering a probabilistic dimension to the study of translation as a cognitive activity. Thus, the focus falls on a framework of dynamic cognitive theories of the mind and fosters the interactive nature of horizontal and vertical cognitive processes in translation.

Nearly 40 years on, as this line of research comes of age, *Explorations in Empirical Translation Process Research* plays a fundamental role in repositioning the research agenda and definitely constitutes a leap forward toward the consolidation of cognitive translation studies as a subdiscipline within translation studies.

Professor of Translation Studies
UFMG, Brazil

Fabio Alves, PhD

References

Alves F (ed) (2003) Triangulating translation: Perspective in process oriented research. John Benjamins, Amsterdam

Alves F, Jakobsen A (2020) Grounding cognitive translation studies: goals, commitments and challenges. In: Alves F, Jakobsen AL (eds) The Routledge handbook of translation and cognition. Routledge, New York, NY, pp 545–554

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research – exploring the CRITT TPR-DB. Springer, London, pp 13–56

O'Brien S (2020) Translation, human-computer interaction and cognition. In: Alves F, Jakobsen AL (eds) The Routledge handbook of translation and cognition. Routledge, New York, NY, pp 376–388

Risku H, Rogl R (2020) Translation and situated, embodied, distributed, embedded and extended cognition. In: Alves F, Jakobsen AL (eds) The Routledge handbook of translation and cognition. Routledge, New York, NY, pp 478–499

Series Editor Foreword

When I persuaded Springer to enter into an agreement to publish six books in the area of Machine Translation (MT), one of the areas we wanted to focus on was MT in use. Far too often, research is conducted without *ever* contemplating who the end user might be, or indeed, whether there might be *any* end user at all who might ultimately find that research of some practical use.

While it is indubitably the case that MT is being used more and more, and that MT quality on the whole is getting better and better, the main professional users of MT will always be human translators who are using MT as a tool in their armory to output translations which are of better quality, or which can be produced more quickly, or with less cognitive effort.

While we can all argue about how MT has been introduced into the pipeline—largely as a result of people who do not understand translation or MT, but who are fixated on lowering the price per word—there have always been researchers in our field who have studied the translation process *per se*, with the vast majority interested in how the translation process with MT included can be improved for human translators. As I say, not everyone has had this as their goal, with (1) the continual pressure on price, and (2) where the metrics clearly demonstrate that human translators are more effective when post-editing MT output, but they do not *feel* as if they were more productive.

All these issues are addressed in this volume, *Explorations in Empirical Translation Process Research*, diligently put together by its editor Michael Carl. It is broken down into four separate but related parts: (1) Translation Technology, Quality and Effort; (2) Translation and Entropy; (3) Translation Segmentation and Translation Difficulty; and (4) Translation Process Research and Post-cognitivism. Altogether, there are 22 contributors, all of whom are household names in this field, so there will be plenty of interest in this volume for sure, whether the reader is interested in how translation activity is captured and measured or what cognitive activity underpins translation. As an encapsulation of the state of the art, or as pointers to future work in the area, this volume will prove invaluable to researchers in these and related

disciplines, and I am very pleased that the book is appearing in the series that I agreed to edit.

ADAPT Centre, Dublin City University, Dublin, Ireland Andy Way

Acknowledgments

The CRITT TPR-DB[1] and the work presented in this volume would not have been possible without the support of many researchers who collected and analyzed the TPR-DB data, and the countless translators and translation students who produced more than 5000 translation sessions with traces of their own translation activities, from (mostly) English into 14 different languages. During more than 10 years these professionals, students, and researchers have contributed their data and experiences to the CRITT TPR-DB.

I am indebted to my former colleague, Arnt Lykke Jacobsen, who first developed the idea to record keystroke (and later gaze) data during translation production to better investigate and understand the hidden mental translation processes. I am also indebted to Matthias Buch-Kromann who invited me to CRITT in 2008 and who initiated the Danish Dependency Treebanks Project[2] at the Copenhagen Business School (CBS) with the idea to integrate linguistic analysis (i.e., dependency annotations), machine learning (i.e., Machine Translation), and translation process analysis, a project that was taken up and pursued to some extent by the CRITT TPR-DB.

I am grateful to collaborators at CBS for supporting the creation of the CRITT TPR-DB: Kristian Hvelplund, Barbara Dragsted, Laura Balling-Winther, and Annette Sjørup for providing their research data as initial seed, and the translation process data that were collected within the Eye-to-IT project (2006–2009). Subsequently, the CASMACAT[3] project (2011–2014) brought together project staff: Mercedes García Martínez, Bartolomé Mesa-Lao, Nancy Underwood, and Ragnar Bonk and collaborators from Universitat Politècnica de València, Spain, and from

[1] The Center for Research and Innovation in Translation and Translation Technology (CRITT) hosts the Translation Process Research Database (TPR-DB) which can be downloaded free of charge from this website: https://sites.google.com/site/centretranslationinnovation/tpr-db

[2] The Copenhagen Dependency Treebanks can be accessed from: https://mbkromann.github.io/copenhagen-dependency-treebank/

[3] Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation project webpage: http://www.casmacat.eu/

the University of Edinburgh, who extended the database with new data and new features. My special appreciation goes to Moritz Jonas Schaeffer who joined the CASMACAT team in 2013 and added thought-provoking research perspectives. Our long collaboration resulted in numerous research publications and several international workshops which advanced the TPR-DB and translation process research in unanticipated ways.

From 2012, the CRITT TPR-DB expanded to non-European languages (Chinese, Hindi, Japanese, among others), first with the energetic determination of late Márcia Schmaltz and well-connected, late RMK Sinha then within several international research projects. These project collaborations included: Vijendra Shukla from CDAC-Noida/India, Fabio Alves at the Federal University of Minas Gerais/Brazil, Eichiiro Sumita from NICT/Japan, and Srinivas Bangalore from AT&T/USA. I am thankful for the productive brainstorming sessions and innovative forms of collaboration which led, among other things, to the idea to conduct international TPR-DB research bootcamps. Many thanks go to Enno Hofeldt for his creative administrative support that enabled CRITT to successfully organize these international events.

The TPR-DB was enriched with Japanese translation data in 2015 during my fellowship at NII, Tokyo, thanks to the generous support of Akiko Aizawa and the enduring collaboration with Masaru Yamada. I am also thankful to Feng Jia who introduced me in 2016 to join Renmin University of China as faculty. The unconditional help from Frank (ZHU YUAN) and the TPR workshops he organized made the TPR-DB known to many scholars in China. Defeng Li and Victoria Lei from the University of Macau made an immense effort to organize the first MEMENTO bootcamp in 2018, where many of the authors in this volume participated. Many thanks to Silvia Hansen-Schirra and the members of her TRA&CO center at the Gutenberg University of Mainz in Germany for hosting the TPR-DB since 2016.

In 2018 CRITT moved to Kent State University (KSU) thanks to the sustained engagement of my colleagues Isabel Lacruz and Françoise Massardier-Kenney, director of the Institute of Applied Linguistics. I would like to thank Keiran Dunne, Chair of Modern and Classical Language Studies, for his enthusiastic support for CRITT, the IT department for finding novel solutions for installing the TPR-DB on the server cluster in Kent, and the other members of the CRITT program advisory committee: Arvind Bansal, Javed Khan, and Manfred van Dulmen, for their unconditional support to establish CRITT@Kent.

I would also like to thank the 22 authors and coauthors of this volume for their patience, flexibility, and hard work during the entire process of this book. Special thanks to more than 30 reviewers who reviewed the individual chapters and helped sharpen the ideas:

Fabio Alves, Bergljot Behrens, Sheila Castilho, Igor Antônio Lourenço Da Silva, Stephen Doherty, Mikel Forcada, Ana Guerberof Arenas, Kristian Hvelplund, Arnt Lykke Jakobsen, Maarit Koponen, Jan-Louis Kruger, Yves Lepage, Elliott Macklovitch, Kirsten Malmkjaer, Chris Mellinger, Joss Moorkens, Ricardo Muñoz Martín, Jean Nitzke, Constantin Orasan, David Orrego-Carmona, Katharina Oster, Adriana Pagano, Reinhard Rapp, Ana María Rojo López, Moritz

Jonas Schaeffer, Katarzyna Stachowiak, Carlos Teixeira, Antonio Toral, Vincent Vandeghinste, Boguslawa Whyatt, Kairong Xiao, Binghan Zheng, and Michael Zock.

Finally, I would like to thank the Springer Series editor, Andy Way, with whom I share a 20-year work relationship, for helping to bring this body of research to the scientific community.

Introduction

1 Empirical Translation Process Research

Translation process research (TPR) may perhaps best be described as a research tradition[1] within cognitive translation studies (CTS)—which is a subdiscipline within translation studies (TS)—exploring factors that determine human translation behavior. Empirical TPR studies human translation (HT) behavior using a range of technologies to record and analyze user activity data. Even though the first attempts to study translation as a cognitive activity date back to the 1960s and 1970s (e.g., Albir et al. 2015; Muñoz 2017), TPR is often said to begin in the 1980s with the analysis of *thinking aloud protocols* (TAP[2]). Since the 1980s and early 1990s, TPR has evolved in several phases with the increasing availability and usage of new sensor and tracking technologies, suitable for recording and analyzing the translation process, with the objective to find out "by what observable and presumed mental processes do translators arrive at their translations?" (Jakobsen 2017, 21) In the last decade, the technological repository of data collection methods has dramatically increased and includes EEG, fMRI, and fNIRS technologies, besides keylogging and eye-tracking.[3] In the development of TPR, Jakobsen distinguishes three phases: the TAP phase, a keylogging and eye-tracking phase, and more

[1] A research tradition has the following characteristics: (1) it defines the aspects of quantification which are viewed as problematic, (2) it defines the methods which can be used to address these problems, and finally, (3) through the definition of measurement problems and methods, a research tradition has a significant impact on how social science research is conducted (https://www.rasch.org/rmt/rmt44d.htm).

[2] "Thinking aloud protocol," a data acquisition method in which the translator is asked to concurrently comment her translation activities.

[3] Some researchers have suggested to widen the scope and definition of TPR and include qualitative and ethnographic research (Risku 2013), or moving into the reality of professional workplaces (Ehrensberger-Dow et al. 2017), while others (e.g., Muñoz 2017) make out different branches within cognitive translation studies all together see also Chap. 13 in this volume.

recently the integration and deployment of methods originating in data analytics and data sciences. Jakobsen says:

> TPR has been dominated methodologically, first by the use of introspective methods, primarily TAPs and retrospection, then by (micro-) behavioral methods, keylogging and eye-tracking, sometimes in combination with cued retrospection, and more recently by the application of computational methods in the analysis of very large amounts of process data. (Jakobson 2017: 39)

Three factors are most notable in the context of recent development:

1.1 Size of Data Collection

With the availability of advanced sensor technologies, it is possible to collect far larger amounts of data, which require automatized data preparation tools for successive statistical analysis. As it is often not possible for one researcher to collect, process, and analyze enough data in one experiment, several initiatives have been started (e.g., LITTERAE, CRITT TPR-DB) to store and integrate the data in large repositories that allow to compare and integrate new recordings with a large body of legacy data. The CRITT TPR-DB started from a small pool of data that included a couple of translation studies conducted at the Copenhagen Business School (CBS) around 2008, all of which made use of the keylogging software, Translog. The idea of the TPR-DB was to integrate translation process data into a coherent open-source database—which has by now grown into the largest set of publicly available translation process data—and to develop analysis tools that would allow for seamless investigation of heterogeneous sets of data across different languages and translation modes. A community effort would make it possible to collect a much bigger data set than one researcher would be able to gather by her/himself. Most chapters in this volume make use in some way of this rich database or its analysis tools.

1.2 Higher Sampling Rates, New Measures, and Theories

While with TAP events on the level of narration (>3 s) would be collected and analyzed, keystroke and eye-tracking technology would allow to go below the integrative scale (0.5–3 s). As a consequence of higher sampling rates and a more precise recording of the behavioral data, new questions could be asked that were not possible to be addressed with smaller data collections. Those new questions would require new measures to uncover latent relations in the data, but also call for new explanatory models and the "introduction of theories, perspectives and concepts from other disciplines" (Sun and Wen 2018) including models of psycholinguistics, cognitive psychology, cognitive science, and bilingualism studies (see also Albir

et al. 2015). These trends are pervasive throughout the volume. Several chapters in this volume introduce new metrics and develop novel, empirically grounded models of the translation process. Using larger datasets with different languages makes it possible to uncover invariances—or to reveal subtle differences—across different translators, languages, and translation modes with high(er) statistical accuracy. Higher precision equipment and more accurate data recordings also make it possible to dig deeper into single examples and to conduct qualitative analysis on a more fine-grained level than would be possible otherwise. Accordingly, the contributions in this volume take their starting point for research in the relations inherent in the data, and several authors explain the observed patterns in novel theoretical framework.

1.3 Translation Technology

TPR has, from its beginnings, not only used technology for the analysis of the translation process but also investigated the interaction of translators with technology during the translation process[4]. With the unprecedented availability and increased quality of machine translation (MT), MT post-editing has become a common practice in recent years. Interaction with translation technology and the "constitutive role" it plays in the translation process as well as in the organization of the translation practice (e.g., outsourcing, interactive MT, Crowd Translation, etc.) also calls for new explanatory models (see, e.g., Muñoz 2017). TPR increasingly stretches into MT quality and evaluation, and some of the work borders on MT development and the assessment of human-machine interfaces. Several chapters in this volume directly address the usage of translation technology and its impact on translation behavior and translation effort.

Instrumental for the collection of large amounts of translation process data and subsequent deployment of data analytics methods in empirical TPR was the development of the CRITT TPR-DB. Translog, as a data acquisition tool, was already available since the mid-1990s and experienced several upgrades. Within the Eye-to-IT project (2005–2009) and several related Ph.D. projects, numerous data sets were collected and little later integrated into an experimental database including tools for data acquisition, data processing, and data analysis. The CRITT TPR-DB has thus evolved into an open-access framework with possibilities for further extension into various directions, to accommodate diverse data acquisition tools (e.g., CASMACAT), to prototype new features, and to explore different explanatory models. Between 2011 and 2015, CRITT has organized yearly summer schools and workshops at the CBS. Participants in these five 1-week intensive summer

[4]The request to investigate the "overall translation process" and the "production of adequate reference works for the translator" was already specified in the ALPAC report (1969: 34, https://www.nap.edu/read/9547/chapter/1) and spelled out in more concrete terms in Kay's (1980) Translator's Amanuensis.

schools were acquainted with quantitative, empirical research methods and the newest versions and features of the CRITT TPR-DB. Given the complexity and interaction of the components in the TPR-DB, there was a need for more intensive hands-on sessions and to offer in-depth introductions for students who would like to dig deeper into the matter than would be possible in a one multi-week summer course. CRITT, therefore, organized multi-week-long summer "bootcamps" in 2013[5] and 2014,[6] which resulted in several publications, among others the volume *New Directions in Empirical Translation Process Research* (Carl et al. 2015) in the Springer series "New Frontiers in Translation Studies."

In 2017, the translation program at CBS was completely abolished, the department was dissolved, and CRITT was temporarily without institutional affiliation. Luckily, in 2018, CRITT found its new home at Kent State University/USA, under the label CRITT@Kent where it is integrated into the Ph.D. translation program and has its own lab space. The idea of multi-week summer bootcamps was continued in the context of the MEMENTO[7] project in which international collaborates would meet to discuss and elaborate their translation research projects. The first MEMENTO bootcamp took place in July 2018 in Macau/China with nearly 30 researchers from 11 countries, followed by a public MEMENTO workshop in Beijing (November 2018) in the context of the fifth ICCTI conference. The second MEMENTO bootcamp took place in July/August 2019 at Kent State University with around 20 international participants and was sponsored by a start-up fund of the newly established CRITT@Kent. The results of this bootcamp were publicly presented at the second MEMENTO workshop in 2019 in the context of the MT Summit XVII in Dublin. The current volume is—among others—a result of these bootcamps and workshops where each chapter has at least one coauthor who attended at least one MEMENTO event.

2 Structure of the Volume

The current volume provides a snapshot of recent developments in TPR. It addresses some of the "classical" TPR topics, including translation competence and expertise, translation difficulty and translation effort, translation units, and translation universals. Some chapters introduce sophisticated measures to assess the translation process and translation product in novel ways, including universal dependency parsing, the human edit rate—a variation of the translation edit rate

[5] SEECAT: https://sites.google.com/site/centretranslationinnovation/projects/seecat

[6] TDA: https://sites.google.com/site/centretranslationinnovation/projects/tda-project

[7] Modelling Parameters of Cognitive Effort in Translation Production (Memento) was a three-year research project (2018–2021) headed by Centre for Studies of Translation, Interpreting and Cognition (CSTIC) at the University of Macao (https://sites.google.com/site/centretranslationinnovation/projects/myrg).

(TER, Snover 2006)—information-theoretic approaches to measure translation literality, and artificial neural networks. Several chapters investigate the translation of metaphors, neologisms, cognates, and culturally specific expressions. Others deal with theoretical constructs including translation universals (shining through, first translational response, the literal translation hypothesis) and the definition of units of translation.

The 15 chapters in the volume are structured in four parts. The first part starts with chapters that have an applied orientation, investigating the (psychological) reality of TER in post-editing. The second part presents four contributions that address various aspects of word translation entropy. The third part deploys qualitative and quantitative methods to address topics in translation segmentation and translation difficulty. Part four provides conceptual, methodological, and theoretical support for a post-cognitivist perspective in TPR.

2.1 Translation Technology, Quality, and Effort

Since its beginnings, TPR has been concerned with investigating the impact of translation technology on the human translation process. The aim of using computer assistance in translation has been to support translators at work, to offer possibilities to lay off memory and cognitive load onto the environment, to provide them with customized collocation and retrieval tools, and to suggest targeted translation solutions at the right time. Specialized editing interfaces are being produced and tested under many different conditions to facilitate the MT post-editing process. Two factors are of crucial importance in this endeavor: (1) the possibilities to assess, in an objective manner, the translation quality and (2) to model, measure, and explain the hoped-for reduction of translation effort.

1. With the widespread deployment of data-driven MT systems at the beginning of this century, automatic evaluation metrics were needed to compare and fine-tune the systems toward a reference or "gold standard." The TER is such a measure which assumes four edit operation in MT post-editing: deleting, inserting, replacing, and moving words and groups of words. It computes the minimum amount of assumed edit operations to match the MT output to a reference translation, e.g., the post-edited version. The number of assumed edit operations is then taken as an indicator for MT quality, where fewer edit operations indicate better MT output.
2. Numerous models have been proposed to explain and understand the reported and observed translation behavior and to relate translation behavior with translation quality. Countless publications refer to Krings (1986) categories of temporal, technical, and cognitive effort: while temporal effort is often used as a proxy for cognitive effort, gaze data provides a more direct insight into the mental activities—though, often not without much noise. However, more fine-grained

models are being increasingly used to explain these findings, and the experimental validation of the models themselves becomes a matter of research.

The first three chapters in this section relate TER scores to human post-editing activities and assess to what extent TER is suited to describe the actual post-editing process. Do Carmo proposes in chapter "Editing Actions: A Missing Link Between Translation Process Research and Machine Translation Research" to reverse the view: instead of looking at the TER, he suggests taking a view on the Human Edit Rate (HER), which may close the gap between MT research and translation studies. Based on do Carmo's proposal, Huang and Carl introduce, in chapter "Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort", a new measure, the word-based HER (WHER) score which they show correlates well with measures of translation effort. In chapter "What Do You Say? Comparison of Metrics for Post-editing Effort", Cumbreño and Aranberri also correlate various measures of cognitive effort with TER scores, but they come to different conclusions. The last chapter in this part reports a study using translation technologies with different goals. Similar to Huang and Carl, also Tardel (chapter "Measuring Effort in Subprocesses of Subtitling: The Case of Post-editing via Pivot Language") studies aspects of cognitive effort in computer-assisted subtitling. However, she compares different settings to find a best information environment for computer-assisted subtitling.

2.1.1 Chapter "Editing Actions: A Missing Link Between Translation Process Research and Machine Translation Research" by Félix do Carmo

Do Carmo proposes a new view on edit operations. Given the increasing deployment of MT for post-editing he suggests reversing the view, and—instead of error correction—to focus on a Human Edit Rate (HER), which relates to editing effort. This, he suggests, might bridge the gap between TPR and MT research, which he illustrates with a number of recommendations.

2.1.2 Chapter "Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort" by Jie Huang and Michael Carl

Huang and Carl introduce a word-based HER (WHER) score as a predictor to measure post-editors' cognitive effort. Their WHER measure computes the minimum number of expected edit operations for each ST word given the MT output. In an experiment with 21 student translators, they find that WHER is a reliable predictor to estimate the post-editing effort.

2.1.3 Chapter "What Do You Say? Comparison of Metrics for Post-editing Effort" by Cristina Cumbreño and Nora Aranberri

Cumbreño and Aranberri present a study to measure and predict post-editing effort for different error types with various measures. They find that indicators of temporal, technical, and cognitive effort do not correlate well with HTER—indicating that a single effort measure might lead to biased measurements.

2.1.4 Chapter "Measuring Effort in Subprocesses of Subtitling: The Case of Post-editing via Pivot Language" by Anke Tardel

Tardel investigates whether and how neural MT (NMT) systems can best be used to support the process of audiovisual translation via a pivot language. She compares several settings for movie subtitle post-editing from Swedish into German and the effect of an English intermediate reference language on temporal, technical, and cognitive effort.

2.2 Translation and Entropy

Entropy is a basic physical measure that quantifies the interaction between two entities. The entropy of an entity (e.g., an ST expression) with regard to another entity (e.g., the set of its possible translations) counts the number of indistinguishable configurations (i.e., different translation for an ST expression). It is the only physical measure that is irreversible and directional (i.e., non-symmetrical), and thus tightly linked to the notion of time, which is also directional and irreversible. The *word translation entropy* (HTra) represents one of three criteria to measure transition literality (e.g., Carl and Schaeffer 2016). HTra has since then been used as a powerful predictor for several translation measures. It has been shown to correlate with various behavioral observations of the translation process, such as translation production duration, gaze time, the number of revisions, but also with properties of the translation product, including translation errors of HT and MT systems. HTra is an information-theoretic measure that quantifies whether there are strong translation preferences. Stronger entrenched translation solutions are less translation-ambiguous, they carry less translation information, they are easier to retrieve, and their production requires less cognitive effort as compared to more ambiguous and less entrenched translations. More entrenched translations are also thought to be semantically closer to their ST equivalent.

In this section, Carl (chapter "Information and Entropy Measures of Rendered Literal Translation") describes the implementation of a "rendered literality" measure, which extends the three previous literality criteria: monotonicity, compositionality, and entrenchment, with the additional constraint of TL compliance. In chapter "redBird: Rendering Entropy Data and ST-Based Information into a

Rich Discourse on Translation", Ogawa et al. observe HTra correlations for HT and MT, across different languages and investigate in detail to what extent this is also the case for different word classes and sets of collocations. Wei investigates, in chapter "Entropy and Eye Movement: A Micro-analysis of Information Processing in Activity Units During the Translation Process", in detail scan paths that are triggered through a high HTra word and discusses patterns of visual search to find disambiguating clues in low-entropy context. Heilmann and Llorca-Bofí detect, in chapter "Analyzing the Effects of Lexical Cognates on Translation Properties: A Multi-variate Product- and Process-Based Approach", an effect of cognateness, as computed with a Levenshtein (1966) distance, on HTra and suggest adding a formal similarity criterion to the list of literality criteria.

2.2.1 Chapter "Information and Entropy Measures of Rendered Literal Translation" by Michael Carl

Carl extends the notion of *translation literality* to include constraints of TL grammar. This "rendered translation literality" is formulated in an information-theoretic framework and is based on the assumption that the most frequent translations are also the most literal ones that cohere with the TL grammar. He shows that rendered literality measures are predictive of translation duration on a word level and a segment level.

2.2.2 Chapter "redBird: Rendering Entropy Data and ST-Based Information into a Rich Discourse on Translation" by Haruka Ogawa, Devin Gilbert, and Samar Almazroei

Ogawa et al. find a significant correlation of HTra values across HT and MT and three very different languages: Arabic, Japanese, and Spanish. In order to explain this correlation, they look deeper into various word classes, figurative expressions, voice, and anaphora and find a similar intra- and inter-language correlation.

2.2.3 Chapter "Entropy and Eye Movement: A Micro-analysis of Information Processing in Activity Units During the Translation Process" by Yuxiang Wei

Wei takes a view from the processing perspective investigating in detail assumed mental processes in a difficult translation related to the metaphorical use of "cough up," a word with a very high HTra value. He assesses gaze patterns from various translators in their effort to find disambiguating clues and explains the observed scan path in terms of information integration, surprisal, and information gain.

2.2.4 Chapter "Analyzing the Effects of Lexical Cognates on Translation Properties: A Multi-variate Product- and Process-Based Approach" by Arndt Heilmann and Carme Llorca-Bofí

Heilmann and Llorca-Bofí investigate the cognate facilitation effect which predicts that, due to more direct inter-lingual connections, the translation of cognates results in lower HTra values and decreased processing time. They test this hypothesis for translations from English into German, Danish, and Spanish and find that higher cognateness leads indeed to lower entropy values, but processing time is moderated by the translator's experience in reading, translation duration, and revision.

2.3 Translation Segmentation and Translation Difficulty

Segmentation during translation production has been a topic of research for many years and has been in some ways at the very core of TPR. Since its beginning, TPR has produced many models to describe, explain, and conceptualize the basic units of translation, but has—until now—not reached a generally accepted conclusion about its nature. Two fundamentally different approaches can be distinguished to describe translation units: by looking into the translation product, one can try to find linguistic and/or cross-linguistic clues that indicate coherent translation segments, such as sequences of monotonous or isomorphic translational correspondence. Incoherent and smaller segments of translational correspondence, larger amount of translation reordering, and less isomorphic ST-TT representations are taken to engender more translation difficulties and potential increased translation effort.

Another approach investigates behavioral data directly—mainly logs of fixations and/or keystroke data—to determine the assumed mental processes of text segmentation and integration during translation production. Less fluent typing, longer keystroke pauses, and more dispersed visual attention and search are taken as indicators of translation difficulty and extended effort. The assumption is that both approaches—the view from the process or the product—would converge and allow us to come to the same conclusions about translation difficulty and translation effort.

This part of the volume addresses translation segmentation and translation difficulty from those different angles. In chapter "Micro Units and the First Translational Response Universal", Carl provides a definition of micro (translation) unit that integrates properties of the translation process and product and allows to investigate the relation between the first translational response and the final translation product. Vanroy et al. examine in chapter "Metrics of Syntactic Equivalence to Assess Translation Difficulty" the translation product from a computational linguistics view and base their notion of translation difficulty on various definitions of cross-lingual syntactic equivalence. In chapter "Using a Product Metric to Identify Differential Cognitive Effort in Translation from Japanese to English and Spanish", Lacruz et al. explore novel types of segmentation, based on the Japanese "bunsetsu," to assess translation difficulties of culturally and contextually dependent expressions, based

on the variation of HTra values. Chen takes a process view on translation difficulty in chapter "Translating Chinese Neologisms Without Knowledge of Context: An Exploratory Analysis of an Eye-Tracking and Key-Logging Experiment" when assessing the success of several translation strategies depending on whether the meaning and background knowledge of neologisms were available.

2.3.1 Chapter "Micro Units and the First Translational Response Universal" by Michael Carl

In this chapter, Carl looks backward from the translation product into the translation processes to uncover the "First Translational Response Universal." He is able to relate gaze patterns preceding the first translation draft with properties of the final product and maps indicators of translation difficulty, such as the number of revisions, on properties of product segments, such as its translation entropy.

2.3.2 Chapter "Metrics of Syntactic Equivalence to Assess Translation Difficulty" by Bram Vanroy, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken

Vanroy et al. take a computational linguistics view on the translation product to assess segmentation and translation difficulty. They develop three different approaches to measure syntactic equivalence in translation, based on the degree of word group reordering and the isomorphy of derivation trees. They show that each of the three different measures covers a different aspect of cognitive effort, as measured by total reading time.

2.3.3 Chapter "Using a Product Metric to Identify Differential Cognitive Effort in Translation from Japanese to English and Spanish" by Isabel Lacruz, Haruka Ogawa, Rika Yoshida, Masaru Yamada, and Daniel Ruiz Martinez

Lacruz et al. analyze translations from Japanese to English and Spanish based on a segmentation into bunsetsus, which is the smallest coherent linguistic units of Japanese sentences. They compare word translation entropy of bunsetsus as indicators of translation difficulty for L1 and L2 translations and show that differences between the two languages and L1/L2 directions are due to cultural and linguistic divergencies.

2.3.4 Chapter "Translating Chinese Neologisms Without Knowledge of Context: An Exploratory Analysis of an Eye-Tracking and Key-Logging Experiment" by Jinjin Chen

Chen asks what the possible strategies are to compensate for the lack of background knowledge when translating neologisms from Chinese to English and when do translators switch from a horizontal to a vertical processing model. Chen makes a distinction between several categories of neologisms and finds that different translation strategies come into play for different categories of Chinese neologisms.

2.4 Translation Process Research and Post-cognitivism[8]

TPR has primarily been concerned with technologically heavy methodologies to collect and analyze translation process data that help elucidate the human translation and post-editing processes. Various explanatory models have been deployed that were borrowed—among others—from cognitive science, psycholinguistics, and bilingualism research so as to interpret the TPR findings in a coherent theoretical framework. With the development of those disciplines in the past 20 years utilizing more sophisticated data acquisition tools, new translation devices, and their technological possibilities combined with the collection of big data sets and more rigorous analysis methods, the explanatory models in TPR have also changed and adapted to the new situation. As pointed out by several scholars in the field (e.g., Sun and Wen 2018, Shreve and Angelone 2010), new process models have to be developed that are able to accommodate those novel developments and research findings. A trend toward post-cognitivist theories can be noticed in recent TPR, and also in this volume where several chapters refer to connectionist models as the explanatory framework.

The last part of this volume underpins this post-cognitivist perspective of TPR. Chapter "Computation and Representation in Cognitive Translation Studies" by Michael Carl postulates that TPR has mainly developed and been a methodology that suggests a mechanistic view on computation. It postulates that "representation" and "computation" are independent concepts, that computational devices are useful for developing and verifying theories of the mind, and that the status of a statement as methodological and ontological has perhaps not always been clearly marked. Chapter "Translation Norms, Translation Behavior, and Continuous Vector Space Models" by Michael Carl introduces a new triangulation method based on artificial neural networks. It integrates findings from bilingualism and translation research and assesses to what extent results from single-word translations may carry over to

[8]Post-cognitivists reject the assumption that the mind performs computations on objects that are faithful representations of an outside world, which is usually associated with the *computational theory of mind* and their protagonists.

translation in context. Chapter "A Radical Embodied Perspective on the Translation Process" by Michael Carl develops a radical embodied post-cognitivist perspective on the translation process. The chapter extends translation affordances with a probabilistic recursive layer and maps this framework onto a dynamic systems approach, capable of explaining "representation hungry" cognition as covariation between the model and the world.

2.4.1 Chapter "Computation and Representation in Cognitive Translation Studies" by Michael Carl

This chapter addresses a recent terminological confusion with respect to the role of the *computational theory of mind* within TPR. It underpins the importance to separate methodological from ontological claims and concludes that TPR is compatible with an extended and embodied view on cognition.

2.4.2 Chapter "Translation Norms, Translation Behavior, and Continuous Vector Space Models" by Michael Carl

This chapter integrates findings from single-word translation experiments in bilingualism research with contextual translation behavioral data from the TPR-DB within a neural network implementation. Observed latencies are explained as vector similarities in an English-to-Spanish word2vec space. The isometry of the ST and TT vector spaces seems to play a decisive role in the translation difficulty. Results from single-word experiments may be adopted with caution to translation in context.

2.4.3 Chapter "A Radical Embodied Perspective on the Translation Process" by Michael Carl

The final chapter in this volume proposes an anti-representational perspective on the translation process that is compatible with research in bilingualism and Schaeffer and Carl's (2013) recursive model of shared representations. The chapter introduces the notion of probabilistic translation affordances, which can be factorized to optimize translation abilities or environmental configurations, apt to describe translation units in a new light.

References

Albir AH, Alves F, Dimitrova BD, Lacruz I (2015) A retrospective and prospective view of translation research from an empirical, experimental, and cognitive perspective: the TREC network. Transl Interpreting 7(1):5–25

Carl M, Bangalore S, Schaeffer M (2016) New directions in empirical translation process research. Springer, Cham. ISBN 978-3-319-20357-7

Ehrensberger-Dow M, Hunziker Heeb A, Jud P, Angelone E (2017) Insights from translation process research in the workplace. In: De Gruyter (ed) Doing applied linguistics. pp 116–123. https://doi.org/10.1515/9783110496604-014

Jakobsen AL (2017) Translation process research. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Wiley, Hoboken, NJ, pp 19–49. https://doi.org/10.1002/9781119241485.ch2

Kay M (1980) The proper place of men and machines in language translation. Xerox Palo Alto Research Center, Palo Alto, CA (Reprinted in: Machine Translation (1997) 12(1–2):3–23

Krings H-P (1986) Was in den Köpfen von Übersetzern vorgeht: Eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern. Tübingen, Narr

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Doklady 10(8):707. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf

Muñoz Martín R (2017) Looking toward the future of cognitive translation studies. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Blackwell, pp 555–573

Risku H (2013) Translation process research as interaction research: from mental to socio-cognitive processes. MonTi Monografías de Traducción e Interpretación (ne1):331–353. https://doi.org/10.6035/MonTI.2014.ne1.11

Schaeffer M, Carl M (2013) Shared representations and the translation process: a recursive model. Transl Interpreting Stud 8(2):169–190. reprint in Describing cognitive processes in translation: acts and events. Ehrensberger-Dow M, Dimitrova BE, Hubscher-Davidson S, Norberg U (eds) Benjamins Curr Top 77

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation of the Americas (AMTA 2006). Visions for the future of machine translation, 8–12 Aug 2006, Cambridge, MA, pp 223–231

Shreve GM, Angelone E (eds) (2010) Translation and cognition. John Benjamins, Amsterdam

Sun S, Wen J (2018) Translation process research: an overview. In: Shei C, Gao Z-M (eds) The Routledge handbook of Chinese translation. Routledge, London, pp 275–290

About the Book

The book assembles 15 original, interdisciplinary research chapters that explore methodological and conceptual considerations as well as user and usage studies to elucidate the relation between the translation product and translation/post-editing processes. It introduces numerous innovative data-driven measures as well as novel classification schemes and taxonomies to investigate and quantify the relation between translation quality and translation effort in from-scratch translation, machine translation post-editing, and computer-assisted audiovisual translation. Translation experiments are conducted for several language pairs in different translation modes using eye-tracking and/or keylogging technology, to compare different types of translator expertise, different types of texts, and various types of linguistic expressions. The research addresses questions in the translation of cognates, neologism, metaphors, idioms, figurative and cultural-specific expressions, re-assesses the notion of translation universals and translation literality, elaborates on the definition of translation units and syntactic equivalence, investigates the impact of translation ambiguity and translation entropy, suggests alternative interpretations of the human translation edit rate, and explores the possibilities of computer-assisted translation via pivot languages. The findings are interpreted in the context of psycholinguistic models of bilingualism and re-frame empirical translation process research within the context of modern dynamic cognitive theories of the mind. The book aims at bridging the gap between translation process research and machine translation research.

Contents

About the Contributors

Samar Almazroei is a Ph.D. candidate at Kent State University. Her research interests are in TPR, quality assessment, and the impact of technological advances on translation studies. Her current research is concerned with revision processes in translation, post-editing, and sight translation and their implications for translation pedagogy and translation profession in general.

Nora Aranberri is a Lecturer in the Faculty of Education of Bilbao at the University of the Basque Country. She holds a Ph.D. in Translation Studies from Dublin City University. Since 2012, she has been a member of the Ixa NLP research group, and more recently, also part of HiTZ, the Basque Center for Language Technology. Her research focuses on machine translation from the user's perspective, professional and nonprofessional, including topics such as quality valuation and post-editing.

Michael Carl is a Professor at Kent State University/USA and Director of the Center for Research and Innovation in Translation and Translation Technology (CRITT). He has published widely in the fields on machine translation, natural language processing, and cognitive translation studies. His current research interest is related to the investigation of human translation processes and interactive machine translation.

Félix do Carmo is a Senior Lecturer in Translation and Natural Language Processing at the University of Surrey. After having finished his Ph.D. at the University of Porto, where he was a guest lecturer, he worked as a postdoctoral researcher in Dublin City University, under a 2-year EDGE-MSCA fellowship. For more than 20 years, he was a translator and a translation company owner in Portugal.

Jinjin Chen is currently a Ph.D. candidate at Centre for Studies of Translation, Interpreting and Cognition (CSTIC), University of Macau. Previously she pursued a master's degree at Guangdong University of Foreign Studies. She takes a keen interest in cognitive translation studies and pragmatic translation studies.

Orphée De Clercq is a postdoctoral researcher in the Department of Translation, Interpreting and Communication at Ghent University (Belgium) with extensive experience in deep semantic processing of natural language. Her main research interests are readability prediction, automated writing evaluation, sentiment analysis, and text mining of user-generated content. She teaches courses on digital communication and Computer-Assisted Language Learning.

Cristina Cumbreño is a Computational Linguist at the Spanish telecommunications company Telefonica. In 2019 she obtained a master's degree in NLP from the University of the Basque Country, collaborating with the university's Ixa NLP research group to complete her master thesis. Her interests revolve around machine translation and post-editing.

Joke Daems is a postdoctoral research assistant at Ghent University, where she conducts research as a member of the Language and Translation Technology Team (LT³). Her research interests include (machine) translation, post-editing, translatability, translation quality assessment, human-computer interaction, and translation stylometry.

Devin Gilbert is currently a doctoral fellow at Kent State University's Translation Studies program. Aspiring researcher/educator by day, freelance translator and interpreter by night. His research interests include translation technology user experience, quality estimation of machine translation, authentic project-based learning, student-directed learning, and how to go mountain biking more often.

Arndt Heilmann is a researcher at RWTH Aachen University and has been working in the DFG research project TRICKLET. He relies on computer linguistic and psycholinguistic approaches to help uncovering the mental processes and states during bilingual language processing. Within the scope of the TRICKLET project, he finished his Ph.D. on the effects of syntactic source text complexity on translation.

Jie Huang is a Ph.D. student in the School of Foreign Languages at Renmin University of China and was a visiting scholar at Kent State University, USA. Her research interests include audiovisual translation, Machine Translation post-editing, and cognitive translation studies.

Isabel Lacruz is Associate Professor of Spanish Translation and Translation Studies at Kent State University. She holds a Ph.D. in Experimental Psychology (Cognitive) from Kent State University and has professional experience as a translator and interpreter. The main focus of her research is to understand and measure cognitive processes involved in translation and post-editing.

Carme Llorca-Bofí is a recent graduate from the Universitat de València. During her B.A. she focused her research on authorship studies. She later expanded her

own research on empirical translation and the effects of cognates at RWTH Aachen University in collaboration with the DFG research project TRICKLET. Her recent M.A. thesis tests both language and music perception skills in bilinguals.

Lieve Macken is Assistant Professor in the Department of Translation, Interpreting and Communication at Ghent University (Belgium). She has strong expertise in multilingual natural language processing. Her research focuses on the impact of translation technology and machine translation on the process and product of translation. She is the operational head of the language technology section of the department, where she also teaches Translation Technology, Machine Translation, and Localization.

Daniel Ruiz Martínez is a researcher and adjunct professor of Japanese-Spanish translation in the Department of Translation and Interpreting of the University of Salamanca, Spain; Japanese language teacher at the Spanish-Japanese Cultural Center (CCHJ); professional translator; and lexicographer.

Haruka Ogawa is a Ph.D. fellow at Kent State University. Having acquired an MA in theoretical linguistics, she is currently writing her dissertation on translation difficulties in English-Japanese translation. Her ultimate research goal is to better understand how people utilize their knowledge and skills when engaging in bilingual operations, such as translation.

Anke Tardel is a Ph.D. student, research assistant, and lecturer at the Faculty of Translation Studies, Linguistics and Cultural Studies in Germersheim, Germany. She has been involved in various eye-tracking studies and projects at the TRA&CO Center and is a junior member of the Gutenberg Academy at JGU Mainz. Her research interests include translation process research with a focus on revision and post-editing, audiovisual translation, and translation technologies.

Arda Tezcan has a background in mathematics and artificial intelligence and holds a Ph.D. in translation studies. He currently works at the Language and Translation Technology Team (LT3) at Ghent University as a postdoctoral researcher. His research interests include natural language processing, machine translation, and human-machine interaction in the context of translation studies. He also teaches introductory courses on natural language processing and on translation technology.

Bram Vanroy is a doctoral researcher working on the PreDict project (Predicting Difficulty in Translation) in the Department of Translation, Interpreting and Communication at Ghent University (Belgium). As such, his main interests lie in readability, translatability, and quality estimation. He obtained a master's degree in Computational and Formal Linguistics, followed by an advanced master's degree in Artificial Intelligence, both obtained from KU Leuven.

Yuxiang Wei is a doctoral candidate at the Centre for Translation and Textual Studies, Dublin City University (DCU), Ireland. He has a cross-disciplinary background and holds a Master of Philosophy (M.Phil.) from the Chinese University of Hong Kong. His doctoral research, which focuses on the post-editing effort of Machine Translation in respect of lexical and structural ambiguity, is funded by the School of Applied Language and Intercultural Studies, DCU.

Masaru Yamada is Professor in the Faculty of Foreign Language Studies at Kansai University. He specializes in translation process research (TPR), including human-computer interaction (HCI) and machine translation plus post-editing (MTPE), and translation in language teaching (TILT). His publication includes "The impact of Google Neural Machine Translation on Post-editing by student translators," JoSTrans 31.

Rika Yoshida is a part-time lecturer at Rikkyo University, Sophia University, and Juntendo University and a visiting researcher at the Institute of Interpreting and Translation of the Aichi Prefectural University. She is a JP-to-SP, EN conference interpreter. Her research interests include interpreter-mediated court discourse analysis, linguistic anthropology, sociolinguistics, and interpreting and translation studies.

Abbreviations

AG	Alignment group
AU	Activity unit
AVT	Audiovisual translation
CTM	Computational theory of mind
HCI	Human-computer interaction
HER	Human edit rate
HT	Human (from-scratch) translation
HTER	Human-targeted translation edit (error) rate
HTra	Word translation entropy
L1	First (native) language
L2	Second language
LMM	Linear mixed model
MT	Machine translation
NMT	Neuro machine translation
PE	Post-editing of machine translation
SL	Source language
SMT	Statistical machine translation
ST	Source text
TER	Translation edit (error) rate
TL	Target language
TM	Translation memories
TPR	Translation process research
TPR-DB	Translation process research database
TrtS	Total reading time, source word
TrtT	Total reading time, target word
TT	Target text
TU	Translation unit

Part I
Translation Technology, Quality and Effort

Editing Actions: A Missing Link Between Translation Process Research and Machine Translation Research

Félix do Carmo

Abstract This chapter presents a discussion, on theoretical and methodological grounds, of the subject of editing as a bridge between Translation Process Research (TPR) and Machine Translation (MT). Editing is described in the chapter as a technical dimension of the writing task performed by translators when they apply four actions to previous text: deleting, inserting, replacing, and moving words and groups of words. The chapter shows some of the difficulties in interpreting the results obtained by TPR data collection methods, and it discusses how edit distances can help improve this interpretation. In the next step, the discussion and demonstrations focus on the need to invert the perspective of edit distances, from the errors of MT to the edits made by translators, for metrics to be more accurate in describing editing work. A new form of using Translation Edit Rate (TER) is suggested, which may be called Human Edit Rate (HER). The chapter also analyzes the limitations of perspectives on editing from MT research, suggesting ways to complement these perspectives with process data. The last part of the chapter extends the discussion, to approach questions related to the advantages of common research between TPR and MT, and its implications for the common understanding of the complexity of translation.

Keywords Editing · Translation process research · Machine translation · Translation edit rate · Translation data · Translation product and translation process · Editing actions

1 Introduction

TPR is one of the branches of Translation Studies, a discipline rooted in Arts and Humanities. TPR brings together empirical approaches to translation that imply an intense engagement with technology. On a different disciplinary setting, that of

F. do Carmo (✉)
Centre for Translation Studies, University of Surrey, Guildford, UK
e-mail: f.docarmo@surrey.ac.uk

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_1

Computer Science, MT research has gained increasing relevance in recent years, often achieving a higher visibility than that of Translation Studies. Researchers have built many bridges across the divide between Humanities and Computer Science, but somehow the connection between TPR and MT has not been effectively explored. One of the reasons for this may be the fact that each of these approaches is looking at the same process using a different language and different methods, which yield outcomes that cannot be reused by the researchers from the other field. Editing, one of the components of the translation process, is seen in this chapter as an element that could help bridge the gap between TPR and MT.

The aim of this chapter is to present a broad analysis of different methods to study editing, looking for a common ground between TPR and MT. From this analysis arises the notion that edit distances provide the best set of elements for the description of editing, although they cannot be said to describe the process itself. The chapter has several objectives. We start with the description of editing and of the difference in perspective between an error rate and an edit rate. This is followed by a series of experiments that show the limitations of edit distances in describing editing. The purpose of these tests is to identify the best methods to reveal what was edited and how it was edited. The final objective is presented in a discussion about current uses of edit distances, related to the usefulness of edit rates, and the chapter concludes with a few advantages of more cooperation between TPR and MT.

After this brief introduction, the chapter clarifies the terminology used and the motivation for this research. This is followed by a discussion on process-based and product-based perspectives of editing, before the focal point moves to the application of edit distances to translation. Most of this introductory content comes from our previous work and is presented here for contextualization of the issues discussed in the chapter. In Sect. 3, the chapter includes an analysis of different tools that can be used to study editing. It covers word processors and tools that log key actions used in TPR research and contrasts these to results presented by edit distances used in MT. The purpose is to show the gaps in each of these approaches, so as to see how they can complement each other. The discussion in the final section grows from the points presented in the chapter, as a demonstration of the relevance of this study both for research and for practical applications.

2 Editing as a Research Subject

The next subsections explain what is meant by “editing” and other terms that are central to the chapter and the context of research on the translation process.

2.1 Processes, Tasks and Actions

Translation, revision, and post-editing (PE) are usually considered the three major processes performed by translators (Lacruz 2017). We can think of reading and writing as the tasks that distinguish these processes at the technical and observable level: translation implies mostly writing, revision implies mostly reading, and PE implies a combination of both tasks, in a proportion that is still not clear (Carmo 2017).

Writing and text production have been studied side by side and in contrast to translation, as Dam-Jensen et al. (2019) show. Writing can be studied as a cognitive process in research that tries to understand the decisions that create observable technical actions. It can also be part of models of technical competence, associated with the study of the technical strategies applied by text producers. Writing is only one of the dimensions of translation, the one that we can call technical, as in Krings (2001).

In all forms of translation, there are at least two main writing tasks that translators perform: (i) translating, when translators need to think and generate a translation from scratch, even when they have text already in the target language, but which they decide to dismiss and (ii) editing, which, as defined in more detail in the next paragraphs, allows translators to modify the target content by applying basic actions (Carmo 2017). We propose that the use of MT output by translators does not necessarily create a new process (known as PE). The use of MT output introduces changes to the reading and writing tasks that are part of a translation process. In other words, PE is a form of translation, in which translators use different sources of content, including MT output, to support their translation decisions, and they produce the target text by using different writing forms.

In this chapter, the term "editing" is not used to mean the same as "revision." We define editing as a writing task by which the translator applies one of four editing actions (or simply "edits") to words or groups of words. The four actions are deleting, inserting, replacing, and moving. Editing is performed in a translation process, for example, when the translator modifies a fuzzy match presented by a translation memory; editing also occurs in a revision process, when the reviser changes a sentence by the translator; and editing is also part of PE, when the translator improves or corrects a suggested translation produced by an MT system.

Editing is often confused with PE, because it plays a major role in this process. However, only when the MT output is of a very high quality can a PE project be completed with only editing. In reality, very often translators need to translate from scratch in PE projects. We have advocated that a threshold should be set between editing and translating, in terms of the density of editing required by each of these forms of writing (Carmo 2017); see also Sect. 2.2.

Several authors in Translation Studies allude to the four editing actions, especially when they describe what translators "do." Toury (1995), for example, describes matricial norms as "omissions, additions, changes of location, and manipulations of segmentation" (1995, 59). As we will see below (Sect. 2.4), these

four actions are widely assumed in MT research as good descriptors of the changes that are required to describe a transformation of a string into another, the same type of transformation that goes on when a source language sentence is translated into a target language sentence, or when an MT hypothesis becomes a proper translation by the process of PE.

2.2 Reasons to Study Editing

There are several reasons why editing, seen as a technical task composed of four actions, is an interesting and useful object of study. In MT research, for example, editing has been studied as a process that can be automated, in an application known as automatic post-editing (APE). In APE, researchers build models that learn edit patterns from aligned corpora that include three elements per segment: the source segments, the MT output, and the post-edited version of the MT output. The aim is to train translation or editing models that help correct new MT output (Carmo et al. 2020).

Since editing occurs in a translation, a revision, and a PE process, we can gain a further level of analysis by using the density of the four editing actions to test and measure these three main processes. The suggested threshold to separate editing from translating can be a specific focus of research. Considering its resemblance to the fuzzy match bands used in computer-aided translation (CAT), one may position it at 25%, i.e. editing a quarter of the words in a segment (Carmo 2017). However, Moorkens and Way (2016) warn about the implications of random positioning of this threshold, because in the industry this is associated with specific expectations in terms of required effort and quality produced. Besides this, the position at which this threshold is placed may vary according to text types, technical domains, translation tools, or user behaviors.

In this chapter, we discuss one of the questions raised by research on editing: how to accurately grasp the changes that happen in an editing process. In other words, we want to know *what* is edited and *how*. We are looking for a snapshot of editing, as well focused as possible. The focus mechanism is given by this description of the process. The resolution of the snapshot depends on the tools that we use, and this is the theme of the next sections.

Elementary properties of editing actions – The word "editing" (as a nominalization of a verb of action) conveys the notion of a process, not of a product. The product of editing (and of PE) is a translation. To study processes in a structuralist approach, we start by identifying their constitutive elements.

Editing actions happen in a space dimension: they are applied to units, which are words or phrases, that make up a static sequence known as a string or a segment. Editing changes two properties of these units: form and position. Length is the other property that requires study, as this defines the string that is edited (the length of a

segment) and the edited units (words that may lose or gain characters or they can be fully replaced by other words, and phrases, which are groups of words that may be edited at the same time).

The time dimension is present in editing by the chronological sequence in which the changes occur. These may happen linearly, from the first to the last unit, but mostly they are non-linear, being partially or cumulatively applied within units, scattered or used recursively in different points of the edited strings, all of this in a dynamic and usually unconstrained process (Krings 2001; Carl and Jakobsen 2009; Alves and Hurtado Albir 2010).

We can divide the four actions into sets of two, based on their roles as primary or secondary actions and on manipulation of form and position.

Insertion and deletion can be considered primary actions, because they cannot be decomposed. Replacement and movement can be decomposed into sequences of deletions and insertions, and so they can be considered secondary: a replacement is a deletion of one word and an insertion of another word in the same position; a movement is a deletion of one word in one position and its insertion in another position (see examples of this in Fig. 1 and Table 1).

Movement could occupy a level of its own, because it can also be decomposed into replacements. Movement is a very efficient way to modify strings, but its results are very difficult to analyze and estimate.[1] One of the simplest movements, moving one word one position forward, creates what was originally called a "shift":[2] a swap of position between two words, but this can be interpreted as two replacements in two positions, as illustrated in example 1, below. Example 2 shows the effect of moving one word (big) two positions backward: in this case, we may see three replacements, or the movement of two words (blue trailer) one position forward. Finally, in example 3, we moved one word to the end of the sentence. In this case, this simple movement could be interpreted as the replacement of all the words in the sentence.

– Example 1:

  1. A word compound is a complex thing.
  2. A compound word is a complex thing.

– Example 2:

  1. A blue trailer big truck.
  2. A big blue trailer truck.

[1] As we will see in Sect. 2.4, movement is often excluded from edit distances, and it is the last to be calculated. This has to do with its complexity—it has been demonstrated that search for all possible movements is NP-complete (Lopresti and Tomkins 1997; Shapira and Storer 2007).

[2] See Sect. 2.4: Damerau (1964) called movements of contiguous characters transposition, but he uses the term "shift" to describe the correction of this error. "Shift" was adopted by most researchers to describe any forms of movement, but we consider that it should be used only for swapping contiguous units, because some edit distances may only estimate these.

– Example 3:

1. Tomorrow I will do that.
2. I will do that tomorrow.

Primary actions are also determined by the direction by which we compare two strings. Imagine two strings (A and B), with the same words, but A is 5 words long and B is 4 words long. If we start comparing from string A, we will see a deletion in string B; but if we compare B to A, we will see an insertion in string A. In fact, deletion and insertion are inverse operations, just like sum and subtraction. The same does not happen with replacement and movement: the inverse of a replacement is a replacement, and the inverse of a movement is a movement. If strings A and B had the same length, but in one of them a word was replaced, this replacement would be identified, whether we compare A to B or B to A. Likewise with movement, a word occupying two different positions in strings A and B will be identified as a movement, with the difference that in one direction of comparison, if the movement is forward, in the other direction it will be backward.

Another way to organize editing actions is based on form and position: deletion and movement only affect position, not form: in a one-to-one alignment between strings A and B, deletion produces what we should mark as an empty position in string B, while movement is the relocation of a unit to another position. On the other hand, insertion and replacement imply a new unit (a change of form) that the user types in an empty position or in the same position of a previous unit. In other words, insertion and replacement imply typing alphanumeric characters, whereas deletion and movement can be performed without touching the keyboard, by using the mouse and a right-click option to delete or by dragging and dropping to move a word (Carmo 2017).

2.3 Data Collection and Processing in TPR and MT

The data collection stage in TPR projects that study writing behavior and the learning stage of MT projects result in different perspectives over what goes on when translators edit text.

Process data – TPR projects like CRITT TPR-DB[3] collect data from translation processes by logging every keyboard and mouse input into what is known as User Activity Data (UAD). These reports may contain translation units (TU) and alignment units (AU) (Carl 2009). In this context, TUs are neither the same as the units of translation discussed in Translation Studies nor the TUs that CAT tools produce (these are formed by bilingual pairs of aligned segments). In these TPR projects, TUs are sets of process data, as they describe all details that refer to

[3] Currently available at https://sites.google.com/site/centretranslationinnovation/tpr-db.

every keystroke in an uninterrupted target text production sequence, separated by pauses. When a word is edited more than once, all "production segments (revisions, deletions, substitutions, etc.)" that affect that unit are aggregated as macro-units (Alves and Vale 2009, 257). AUs are simpler to define, as they are "translation correspondences in the static product data" (Carl 2009, 227).

Product data – MT systems learn from product data. MT engines are trained on parallel corpora: bilingual, segment-aligned big data that is collected from reliable sources. In statistical MT (SMT), the result of this training was a target language model and a translation model, which consisted of phrase tables, with several levels of alignment between different-sized units in the two languages. With NMT, these models are not explicit, as the neural network is not transparent to what it contains. Furthermore, the theory of neural network training is not based on units and sub-units, because the power of this technology is in how it considers the source segment as a whole (Forcada 2017b).

There is nothing in this training data that describes the translation process. The translation process, in which target text is generated, is emulated by MT during decoding. This process is determined by mathematical methods for processing big data, which do not arise from theories or descriptions of the human translation process itself. Since the process of translation in NMT is "hidden" in its deep layers, it seems to be impossible to tweak it with process data.

At the evaluation stage, MT research again compares products, namely the MT output against a reference. Evaluation metrics, such as BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), and others, do not consider the process of going from source to target or the process of going from MT output to its post-edited versions; they simply compare one segment in one language against a segment in another language and estimate the differences or similarities, according to different methods.

2.4 Edit Distances

An overview of edit distances – Edit distances are metrics that estimate the shortest distance between two versions of a string, which may be a word, a sentence, or a genome sequence. The units of these metrics are the operations that are necessary to edit one string until it becomes the other.

Edit distances first appeared in the 1960s as instruments to correct computer code and spelling mistakes. Levenshtein (1966) devised a program to check bugs in computer code by estimating whether a character had been deleted, inserted, or if one character had been wrongly placed in the position of another. Damerau (1964), on the other hand, checked that most spelling mistakes belonged to one of the four categories: a wrong character was in the place of another, there was one character missing or an extra one was inserted, or two adjacent characters had been "transposed." Damerau's edit distance method started by estimating replacement of

one character (when this was unmatched in a sequence of matching characters), then a shift between two contiguous characters, then if one string was longer, it looked for the inserted character, and finally, if the string was shorter, it looked for the deleted character (Damerau 1964).[4]

Different edit distances use different operations, but the most complete ones use the four editing actions (deletion, insertion, replacement, and movement). *Diff* utilities, commonly used in several programming languages to compare two sets of data, are based on the estimation of the "longest common subsequence," only considering insertions and deletions. There are also edit distances that consider every edit as a replacement, like the Hamming distance, and others do not consider movement, like Word Error Rate (WER) (Tillmann et al. 1997). The different purposes or applications of these distances determine the actions that they model and how they consider the strings (for example, a method that only estimates replacements can only work with similar length strings).

For our analysis, it is important to stress that the estimation process of edit distances is sustained by the assumption that it is possible to devise a process of transformation by comparing two strings.

Edit distances in translation – The main purpose of using edit distances in translation was to evaluate the quality of MT output using automated methods. Edit distances allow for MT output, called the "hypothesis," to be compared to other strings which are called the "reference." These references should present the quality that the MT system aims at, and they can include gold-standard human translations, produced in a process that is independent from MT, or human post-edits of MT output.

Snover et al. (2006) presented Translation Edit Rate (TER) as a metric for the evaluation of MT output. In this metric, insertions, deletions, and substitutions are calculated by an alignment function using dynamic programming. Then, a greedy search method is used to estimate the shifts that reduce the number of insertions, deletions, and substitutions: a movement converts deletions, insertions, and replacements that affect the same units in different positions into single edits. The greedy search applied by TER is constrained by several rules, namely the length of the segment and a maximum number of positions shifted, and it privileges phrase-level edits: the algorithm starts by looking for sequences of more than one word in new positions.

For the identification of editing actions, the algorithm only needs to worry about the form and position of units (see Sect. 2.2). Length is minimally considered: the end-of-string marker identifies the length of the two strings, and units are orthographic words; time is not considered. These calculations aim at finding the "minimum distance" between the two strings, "minimum" meaning the least

[4]Edit distances are generically referred to as "Levenshtein distance." There is an imprecision in this that may result in methodological and interpretation issues. The distance known as "Damerau–Levenshtein," although it could simply be attributed to Damerau, is the one that includes the four types of edits.

number of operations required to transform one string into the other. This efficiency requirement does not necessarily have a correspondence in real PE processes.

HTER is not TER with PE – In the same paper, the authors present HTER—Human-targeted (or human-mediated) Translation Edit Rate. There has been a lot of confusion about how HTER differs from TER, so it is worth taking some time to clarify these two terms.

There are two roots for this confusion, both in the chapter "From TER to HTER," from (Dorr 2011, 837–838). In this chapter, it is stated that HTER is created in a process called PE and that the difference between TER and HTER is the fact that the first uses gold references, translated in an independent process in which MT was not involved, and the latter uses post-edited versions of the MT output. It is not very clear why it would be relevant to have a new metric, when the estimates are the same, just because the source of the reference changes, but this interpretation has been accepted by many users of these metrics. This is a comment from the final report of a shared task on APE: "Since edit distance is computed between each machine-translated sentence and its human-revised version, the actual evaluation metric is the human-targeted TER (HTER). For the sake of clarity, since TER and HTER compute edit distance in the same way (the only difference is in the origin of correct sentence used for comparison), henceforth we will use TER to refer to both metrics." (Bojar et al. 2015, 29).

However, the description of the process in Snover et al. (2006) shows that it is not accurate to differentiate TER and HTER based on the references used. The two metrics were created in an effort to find a good method to assess the quality of MT output without the need for expensive human annotators. Another point in common is that both metrics are used in evaluation tasks that involve several references for the same MT output. TER is an automated metric (which completely dispenses with the need for human annotators), but HTER is semi-automated, since human annotators are involved. However, what these annotators do is not PE.

When used with several references, to estimate the final score, TER selects the one with the shortest distance to the MT output. HTER was created to optimize this distance, by including annotators that would check the references, choose the one that would imply the lesser number of edits, and then reduce that distance even further, editing either the MT output or the chosen reference. The purpose was to analyze the semantic variation in the alternative translations and to "craft" an improved one, which maintained a close semantic relationship to the content of the references, and, most importantly, the shortest distance to the MT output of all. Note that this annotation process was performed by monolingual users, who only knew the target language (of the MT output and the references), not the language of the source text, so the semantic connection to the source might get lost. This is not what happens in an actual PE process, which is always done while looking at the segment in the source language. The reason why common PE, done while checking the source language, was avoided in this evaluation is the fact that it could increase the edit distance to the MT output (moving it closer to the source language). Bilingual PE would defeat the purpose of the annotation task, since the human in the process would insert their own interpretation of the source and expand the variation

in the references. "Human-targeted" has been interpreted as meaning something like "targeted at a human process," giving access to the process of creation of a PE version, when in fact it means "targeted [by human annotators] for this system output" (Snover et al. 2006, 2).

When researchers use an edit distance to compare an MT output and its PE versions, the metric used is still TER: not only nothing is gained by calling it HTER, but also a confusion arises between what we know from the PE process behind it and the process that was performed to create a targeted reference in HTER.

TER and TERp – In TER, all edits count as one, which means that there are no weights that consider higher difficulty of some edits over others. The sum of these costs is then normalized, i.e. divided by the average number of words in the reference. A few years later, an improved variation of TER was presented, this time called TERplus, or simply TERp. (Snover et al. 2009a,b).[5]There are four main differences between TER and TERp.

- First, TERp discriminates sub-types of substitution: replacing whole words (normal substitution), replacing words by one of its variants (stem matching), replacing with a synonym (synonym matching), and replacing phrases (phrase substitution or paraphrase).
- The second difference is cost. In TERp, the cost of substitution by paraphrase is not fixed, while all the others maintain the cost of 1. This cost is estimated by a combination of the probabilities of that paraphrase and the amount of edits needed to align the two phrases.
- The third difference is the relaxation of the criteria to identify a shift.
- The final difference is the fact that TERp is capped at the cost of 1, and TER is not. This means that when a one-word string is compared to a string with three words, TERp presents an edit score of 100.00 (which means "all words in the string [1.00 or 100%] were edited"), whereas TER presents an edit score of 300.00 (which means that the number of words tripled).

TER as an error rate or as an edit rate – Edit distances are also commonly known as "error rates." WER is employed, for example, by Tillmann et al. (1997), Niessen et al. (2000), and Popović (2011). The word "error" is naturally related to the purpose of evaluating MT output, but a footnote in Snover et al. (2006) explains that the "E" in TER should be understood as an "edit," not as an "error," the idea being that there may be different edits for the same error.

Referring to edits or errors is not just a terminological difference, but one of perspective, purpose, and method. An error rate looks for what went wrong in the MT output, when compared to the reference, so its focus is on the hypothesis. An edit rate looks at the editing process necessary to correct that error, so its focus is on the reference. Therefore, when we talk about a "deletion" in TER, we may be talking about different things (see Sect. 3.2). The perspective that TER or HTER describes

[5]TERcom is a free tool to estimate TER and HTER, available at http://www.cs.umd.edu/~snover/tercom/. A tool to use TERp is available at https://github.com/snover/terp.

or simulates edits has led to equivocations that may affect how we interpret the data these metrics provide.

In the report of the APE shared task in 2016, TER values measured between the output of several APE systems and the PE version of the original MT output are interpreted as meaning "the edit operations needed to transform the output of each system into the human post-edits available for each test sentence" (Bojar et al. 2016, 182). In fact, TER does not describe edits needed to correct MT errors; it only describes those MT errors. At the beginning of Sect. 3.2, we present a few experiments with different settings of TERcom, which allow us to discuss the methodological implications of using TER as an error rate or an edit rate.

3 Tools for the Analysis of Editing

This section shows how the details of the four editing actions are recorded by different tools. This demonstration is broken into two parts: in the first one (described in Sect. 3.1), we used different tools to record our actions while we edited a few simple sentences, one edit action per sentence. First, we used a common word processor and then two keyloggers frequently used in TPR research. The purpose was to see how accurately did the records of these tools identify the four editing actions. In the second part (Sect. 3.2), we analyze how edit distances report different types of editing. We start by showing simple examples, to check whether they register the errors in the hypothesis or the edits in the reference. After we detect the setting that allows us to see the edits, we check their accuracy, by applying more complex patterns of editing.

3.1 Process Data from Common Tools and Keyloggers

Word processors – Microsoft Word™ gives us an opportunity to show two methods of tracking editing, one process-based and the other product-based.

When we activate "Track changes" in MS Word, all edits are registered with different graphical markers, which identify their type. As illustrated in lines 2 and 3 of Fig. 1, deletions and insertions are clearly identified. When we select and overtype a word, it shows as a full word replacement (line 4). If we enable the settings for character-level tracking, when a word is only partially modified, the markers show the characters that are affected (line 5). Finally, when we select a word, then drag and drop it in a new position (performing a movement), this is registered as one word deleted in one position and inserted in another position (line 6).[6]

[6]These illustrations were created with the MS Office 365 Business™ version available in August 2019. The version available in March 2020 shows the differences at the character-level replacement

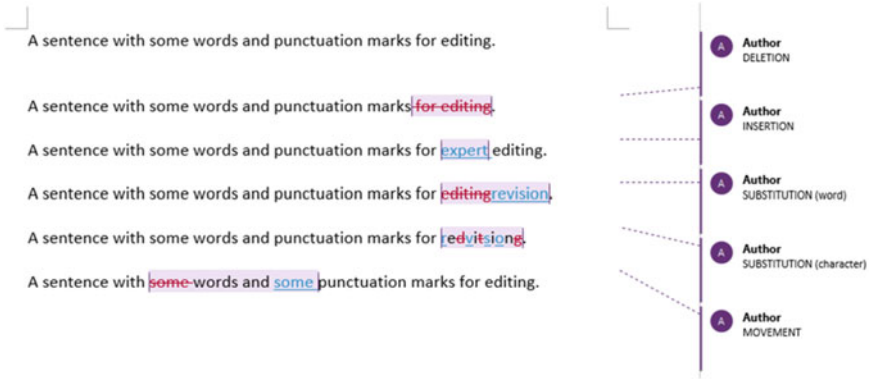


Fig. 1 Edits with track changes enabled in MS Word

Figure 2 illustrates another use of this feature, but as a tool of comparison between two versions. In this case, we produced the same edits with "Track changes" disabled. After having saved the second version, we reopened the program and we chose the "Compare" function. The compared document shows the same visual markers, with different results. In this case, even with the character-level setting enabled, all changes are recorded as affecting complete words (see lines 4 and 5). The most notorious of changes is in the moved word. The actual edit performed was moving the word "some," as illustrated in Fig. 1. Nevertheless, the Compare function, instead of identifying the word that was moved two positions forward, shows the deletion and insertion of the two words that were between those two positions.

Although word processors are not designed as process data loggers, the data they record allows for interesting observations. To start, we can see with this simple demonstration that *a posteriori* analyses of editing, no matter how simple this editing may be, are fallible. This has also shown that estimating movement is a difficult task. Still, word processors have a strong argument in their favor, which is ecological validity, an important requirement in user studies. These tools feature smart editing aids, like suggestions to correct spelling, grammar, and trailing spaces, aids to complete words and apply formatting, and resources like synonym dictionaries, thesauri, and even MT. This means that any tool that is used to collect editing behavior without any of these elements needs to take into account the effect of the absence of these features. It would be good, though, if we could observe in the logs differences in the use of the mouse or the keyboard. Drag and drop is seen as the most natural way to move words, and that is true for mouse users, but users

only in a side panel. This program tracks movement with a different type of marker, but only for movement of full sentences. It is important to note that MS Word saves this information in an XML format file, which we did not investigate, because the purpose here was simply to start illustrating a comparison between the description of a process and an analysis of the product of this process.

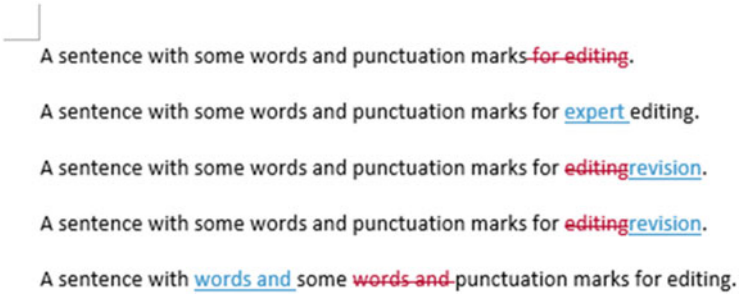


Fig. 2 Two documents compared with the same edits

who favor the keyboard navigate and move words by selecting, deleting, moving over other words, and dropping them in new positions, using different keyboard shortcuts with the Ctrl key to speed up this navigation.

More than word processors, translators regularly use software applications known as CAT tools. In previous work (Carmo 2017), there is a demonstration of similar experiments using Post-Edit Compare (PEC), an add-on to SDL Trados Studio™. That experiment showed how PEC did not record the process but applied an implementation of TER, with similar effects to the ones analyzed in this chapter.[7]

TPR keyloggers – There are many tools to log keyboard and mouse actions while writing, translating, and editing. A few examples are PET (Aziz et al. 2012), iOmegaT (Moran et al. 2014), and HandyCAT (Hokamp and Liu 2015). Some of these tools also provide a replay of the actions performed by translators, which is relevant for retrospection analyses of the process. TransType (Macklovitch et al. 2005) was one of the first tracking logs used in an MT project. In this software, the replay function serves presentation and research purposes, as a way to understand translator's behavior and develop systems that effectively support the work of translators. The wealth of information collected by such tools needs to be worked, analyzed, and interpreted according to process and action models, before we can understand all its dimensions and we are able to act on that knowledge. As the name indicates, keyloggers work at the character level, reporting each keyboard and mouse action. The replay function enables the visualization of a stream of the translation process. One has to admit, however, that the decision process that guides the translation and editing process lies somewhere between the sequence of characters that is collected from the keyboard and the reproduction of a video of the full process.

[7]PEC was further developed to integrate time management features and is now known as "Qualitivity," being amply used in the translation industry to collect data on translators' effort.

This section focuses on two of the most used keyloggers in TPR: Translog II (Carl 2012) and Inputlog (Leijten and Waes 2013).[8] We will look at the outputs of both programs and see how the data in their logs fits into the description of the four editing actions.

Translog II – This software was expressly created as "a program for recording user activity data for empirical translation process research" (Carl 2012). It has been used in experiments on behavior analysis in translation, revision, PE, and writing tasks. The software contains its own editing environment, and all data is collected from within that environment. The interface is limited in terms of language and writing aids, but its strength is in the data analysis tools that complement it.

The raw output of an experiment in Translog is a detailed XML file, presenting with millisecond precision every keyboard and mouse action performed. The XML log is interpreted into tables that organize this data into different categories, according to research objectives. In this interpretation process, an alignment method creates tables with AUs, in which source and target words are aligned, including the keystrokes that produce them. The detail and breadth of information in these tables are presented in Carl et al. (2016). This data enables advanced calculations such as typing inefficiencies, cross-lingual distortions, word translation entropy, and perplexity. Figure 3 illustrates a part of the Translog II log which shows the sequence of edits presented in the previous demonstration.

- In the first highlighted block, we see the selection of two words with Ctrl+Shift+right, the use of the Delete key, and the transformation of "for editing" into an empty space.
- The next block, after a sequence of navigating character by character to the right, shows the insertion of the word "expert," one character in each line, in a time close to two seconds (from 00:28:906 to 00:30:797).
- The third edit was replacing "editing" with "revision," first by selecting the word with the keyboard (from right to left, seven characters, from position 235 to 229) and then inserting eight characters, starting in position 229 and finishing in 236.
- Movement is identified as mouse actions, in the last block: first, selecting the word "some" in position 320 (at 00:55:375) and then dragging and dropping it in position 318 (at 00:57:359).

As illustrated, a log file from Translog contains all the information we need to retrieve the information on the editing process with the four editing actions: units, positions, length, and time. However, to the best of our knowledge, there has been no work in converting these character-based logs into the four editing actions, at the level of the word or phrase. The fact that most real editing work does not correspond to the very simple and clean editing examples that we simulate here is one of the reasons for this. In real editing, users edit only a few characters in

[8] These tools are available online, respectively, at https://sites.google.com/site/centretranslationinnovation/translog-ii and https://www.inputlog.net/.

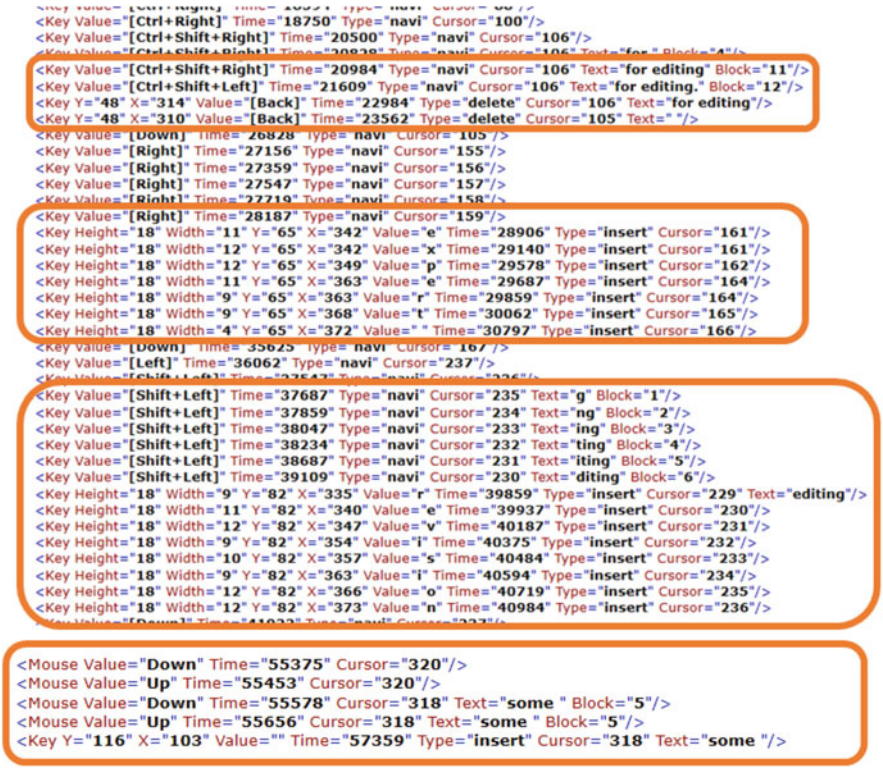


Fig. 3 A log from Translog II showing the four editing actions

one word, come back to the same word to replace it, select and delete a sequence of several words, and perform other scattered and complex editing sequences that complicate the process of mapping actions to words.

Inputlog – This software was developed to collect data and analyze writing behavior. Contrary to Translog, Inputlog does not have a specific editing interface, recording typing actions from other editors, including MS Word and SDL Trados Studio. Besides, it also records actions in web browsers, which makes it useful for adding information about what goes on during typing pauses.[9]

For this experiment, we used Inputlog to record our work in MS Word. After we installed Inputlog, the behavior of MS Word changed, in a way that interfered with our experiment. Inputlog disables drag and drop in MS Word, affecting users who

[9]Leijten and Waes (2013) describe a process to convert characters to words, using a specific type of notation in which all edits are recorded as deletions or insertions. This method is called "s-notation" (Kollberg and Eklundh 2002), but we could not get a working report in this format in any of our experiments with Inputlog.

resort to the mouse for moving words and phrases around. Besides, it disables the selection of more than one word with the Ctrl key, affecting users who depend on the use of the keyboard to speed up editing.

Although it was initially focused on writing (over an empty page), the program is commonly used in research into the translation, revision, and PE processes. The terminology used by the program documentation refers to any change made to a word as a "revision." (Edits are keystrokes that compose revisions.) Data analysis is very complete, with the program outputting several XML files. The current analysis is illustrated by the revision matrix log.

Figure 4 represents a good example of the difficulty in interpreting the data produced by a keylogger in terms of word-based editing actions. The report is sorted chronologically, with each revision numbered. The first block highlighted in the table (row 0) shows the six sentences that were already on the page, prepared for editing, as if this was a normal production unit.

The next three blocks of the table illustrate edits at the word level: first a deletion of two words (row 5), then an insertion (row 7), and then a replacement, by deleting a word and writing another in the same place (rows 8 and 9). The table shows all times and positions in which these actions occurred. The selection of rows 10 to 15 shows the replacement of a few characters in the word "editing" that result in it becoming "revision." We would need to reconstruct the process, from the initial word through all the changes to individual characters, to be able to identify this change in these rows. The order of these replacements could be:

#Revision	Type	Content	Edits	Start	End	Duration	Length	BeginPos	EndPos
0	Normal Production	A·sentence·with·some·words·and·punctuation·marks·for·editing.·A·sentence·with·some·words·and·punctuation·marks·for·editing.·A·sentence·with·some·words·and·punctuation·marks·for·editing.·A·sentence·with·some·words·and·punctuation·marks·for·editing.·A·sentence·with·some·words·and·punctuation·marks·for·editing.·A·sentence·with·some·words·and·punctuation·marks·for·editing.·	39	00:09.176	00:09.976	00:00.800	372	0	372
1	Deletion	·	2	00:25.272	00:25.352	00:00.080	1	372	371
2	Insertion	s	3	00:25.272	00:25.352	00:00.080	1	47	48
3	Deletion	·for·editing.·	7	00:29.824	00:31.753	00:01.929	14	110	124
4	Insertion	·for·editing.·	1	00:29.824	00:31.753	00:01.929	14	110	124
5	Deletion	·for·editing	6	00:31.824	00:32.328	00:00.504	12	110	122
6	Insertion	.	1	00:31.824	00:32.328	00:00.504	1	110	111
7	Insertion	expert·	16	00:32.856	00:40.208	00:07.352	7	165	172
8	Deletion	editing	9	00:41.848	00:42.984	00:01.136	7	234	241
9	Insertion	revision	10	00:44.072	00:45.208	00:01.136	8	234	242
10	Insertion	r	8	00:45.768	00:46.960	00:01.192	1	297	298
11	Deletion	t	4	00:47.240	00:47.609	00:00.369	1	299	300
12	Insertion	v	2	00:48.456	00:48.520	00:00.064	1	299	300
13	Deletion	n	4	00:48.688	00:49.032	00:00.344	1	301	302
14	Insertion	s	2	00:49.136	00:49.235	00:00.099	1	301	302
15	Insertion	o	3	00:49.304	00:50.296	00:00.992	1	303	304
16	Deletion	·	4	00:50.352	00:51.288	00:00.936	1	305	306
17	Deletion	ome·w	8	00:51.520	00:51.624	00:00.104	5	323	328
18	Insertion	some·	1	00:51.520	00:51.624	00:00.104	5	323	328
19	Deletion	some·r	15	00:51.520	00:51.624	00:00.104	6	323	329
20	Insertion	some·w	1	00:51.520	00:51.624	00:00.104	6	323	329
21	Deletion	some·	36	00:51.520	00:51.624	00:00.104	5	323	328

Fig. 4 A log from Inputlog showing the four editing actions

– insert “r” at the beginning of the word “editing” (row 10);
– replace “d” with “v” in (what is now) position 3 (“v” is inserted in row 12, but there is no record of deleting “d”);
– replace “t” with “s” in position 5 (“t” is deleted in row 11, “s” is inserted in row 14);
– insert “o” in position 7 (row 15);
– delete “g” in position 8 (there is no record of this deletion).

The number of edits (keystrokes) this operation implied and the times and the positions in which these actions take place are not very clear, as these should decrease after a deletion and increase after an insertion.

Movement is the most difficult action to be expressed simply by deletions and insertions. The example of moving “some” two positions behind is described in rows 17 to 21 of the table. Row 18 refers to the insertion of “some” and row 21 to the deletion of “some.” Around these two actions, there are other deletions and insertions, but, according to Inputlog, these actions implied a total of 61 keystrokes, which somehow were all performed in less than 104 milliseconds.

This shows that, although the information about writing and translating provided by these tools is very rich and detailed, if we want to study editing in terms of editing actions, both Translog and Inputlog data require extra work of interpretation. We have illustrated this with very simple editing work, but it should be clear that the more complex the editing is, the more difficult this conversion will be.

3.2 From Simple to Complex Editing

If collecting process data and then interpreting it are complex endeavors, maybe edit distances allow us to get a realistic picture of what editing is. However, as stressed in Sect. 2.4, edit distances look at products and try to work out the process behind the transformation of those products.

The purpose of this section is to check whether edit distances show a realistic description of the editing process. We want to know which words were edited and how.

Simulating simple editing – Let us simulate a set of simple translation edits, by applying each edit action only once to one word in four sentences. Table 1 presents the edits as we performed them. In these simulations, we are always replicating a PE scenario, in which MT output is edited by a translator. In the following table, we adopt and expand the notation used in s-notation (Kollberg and Eklundh 2002): {insertion}, [deletion], _replacement_, and **movement**.

Table 1 Four sentences with one edit each

	Unedited (MT)	Edited (PE)
1	This sentence has a redundant [superfluous] word	This sentence has a redundant [ ] word
2	In this sentence, a { } is missing	In this sentence, a {word} is missing
3	This sentence has a _incorrect_ word	This sentence has a _corrected_ word
4	In this sentence, all are correct **words**, but one is in the wrong position	In this sentence, all **words** are correct, but one is in the wrong position

Table 2 TER(mt,pe) scores for edits in Table 1

Sent Id	Ins	Del	Sub	Shft	WdSh	NumEr	NumWd	TER
Sentence1	**1**	0	0	0	0	1	7	**14.286**
Sentence2	0	**1**	0	0	0	1	9	**11.111**
Sentence3	0	0	**1**	0	0	1	7	14.286
Sentence4	0	0	0	**1**	1	1	17	5.882
TOTAL	1	1	1	1	1	4	40	10.000

So, in the first sentence, “superfluous” was deleted; in the second, the missing “word” was inserted; the “incorrect” word was replaced by “corrected” in the third sentence; and the word “words” was moved back two positions in the fourth sentence. How does TER identify these actions?

TERcom requires one basic setting: defining which version is the hypothesis and which version is the reference.[10] We first set up TERcom according to the default TER(mt,pe), i.e. we use the MT output as the hypothesis and the PE version as the reference.

Table 2 shows an insertion in sentence 1, while we actually deleted a word, and a deletion in sentence 2, while we inserted a word.[11] There are no apparent issues with the simulations of replacement and movement. In this setting, we are comparing the MT output with the reference and considering that everything that is different in the MT output is errors produced by the MT system. Therefore, the error of the first sentence is a word that the MT system wrongly inserted, since it did not exist in the reference.

TER, set up like this, is an error rate, focused on the hypothesis, respecting its intended use.

[10] The command that we used in these experiments was `java-jartercom.7.25.jar-N-s-rC:/.../tercom/files/Experiment1-Ref.txt-hC:/.../tercom/files/Experiment1-Hyp.txt-nC:/...tercom/outputs/Exp1.Out.txt`.

[11] The data in the following tables is based on the summary reports (.sum files) produced by TERcom. The titles of the tables stand for, in order: Insertion, Deletion, Substitution, Shift, Words shifted, Number of errors, Number of Words (in the reference), and TER score. The TER score is estimated by dividing the number of “errors” by the number of words. The edit scores are presented as percentages, for example, 14.286 stands for 14.286%.

Table 3 HER(pe,mt) scores for edits in Table 1

Sent Id	Ins	Del	Sub	Shft	WdSh	NumEr	NumWd	HER
Sentence1	0	**1**	0	0	0	1	8	**12.500**
Sentence2	**1**	0	0	0	0	1	8	**12.500**
Sentence3	0	0	**1**	0	0	1	7	14.286
Sentence4	0	0	0	**1**	1	1	17	5.882
TOTAL	1	1	1	1	1	4	40	10.000

Introducing Human Edit Rate – To try and test TER to identify edits, for example, to identify the correct words that were deleted and inserted during PE, we must invert the hypothesis and the reference. That way, the PE becomes the first element of comparison, i.e. the hypothesis, and the MT output becomes the reference. Let us call this metric "Human Edit Rate" and annotate this as HER(pe,mt). We ran a second experiment in TERcom, with this inverted setting.[12]

In Table 3, we can see the actual deletion in sentence 1 and the insertion in sentence 2, correctly estimated by HER. If we now look at the edit scores at the sentence level (at the end of each row in the two tables), we see that the scores of these two sentences are different for TER and HER, because of the change in the denominator (which is always the reference), according to whether this is the version in which an extra word was inserted or the version in which it was deleted. However, the total score of this set of sentences in TER and HER is the same, as this is estimated at the total row, dividing 4 edits by 40 words. It might seem that the inversion of TER did not present surprising results, but, as we will see in the rest of the chapter, there are other factors that affect global edit scores and interesting details that are worth studying.

HER, the inversion of hypothesis and reference in TER, is an edit rate, since it estimates the edits that were done to the MT output.

The wrongful identification of deletions and insertions by TER is corrected by HER, but replacement and movement seem not to be affected by the inversion of reference and hypothesis. This is at least what we would conclude if we only looked at the edit count reports. As we said above (see Sect. 2.2), primary actions, deletion and insertion, are determined by the direction of comparison. Replacements and movements, on the contrary, always count the same, no matter the direction of comparison. But TERcom shows more effects of this inversion in other reports, which include word alignments and the respective editing actions.

Figure 5 shows two extracts of two XML reports by TERcom, one created using the TER(mt,pe) setting and the other created by using the HER(pe,mt) setting. These extracts only show the result of the action we applied to the last sentence in Table 1.

[12]For this experiment, we used the same command but swapped the content of the Hyp and Ref files. Maja Popović kindly reproduced these two experiments with WER, and the results were similar, the only difference being in the fact that movement is estimated as two edits: a deletion and an insertion.

Fig. 5 Two extracts of the word alignment report from TERcom, comparing the TER and HER scores for the last sentence in Table 1

```
Sentence ID: Sentence1:1
Original Ref: This sentence has a redundant {superfluous} word .
Original Hyp: This sentence has a redundant word .
Hyp After Shift: This sentence has a redundant word .
Alignment: (     D  )
NumShifts: 0
Score: 0.125 (1.0/8.0)
Sentence ID: Sentence2:1
Original Ref: In this sentence , a [] is missing .
Original Hyp: In this sentence , a word is missing .
Hyp After Shift: In this sentence , a word is missing .
Alignment: (     I   )
NumShifts: 0
Score: 0.125 (1.0/8.0)
Sentence ID: Sentence3:1
Original Ref: This sentence has a incorrect word .
Original Hyp: This sentence has a _corrected_ word .
Hyp After Shift: This sentence has a corrected word .
Alignment: (    S  )
NumShifts: 0
Score: 0.14285714285714285 (1.0/7.0)
Sentence ID: Sentence4:1
Original Ref: In this sentence , all are correct words , but one is in the wrong position .
Original Hyp: In this sentence , all words are correct , but one is in the wrong position .
Hyp After Shift: In this sentence , all are correct words , but one is in the wrong position .
Alignment: (                    )
NumShifts: 1
  [5, 5, 7/7] ([words])
Score: 0.058823529411764705 (1.0/17.0
```

Fig. 6 A .pra report from TERCom, depicting the calculation of HER(pe,mt) for the sentences in Table 1

Both reports show all words aligned and signal the number of words and direction of the movement. On the left, from the TER setup, we can see the annotation "C,-2" in front of the aligned "words": this means that this word was correctly aligned after it was moved two positions back. The HER setup, on the right, shows "C,2," meaning HER identified a movement two positions forward. Actually, the movement we performed was the backward movement, and the resulting sentence from our process appears in the correct order in the TER report. To explain this, we can resort to another TERcom report.

The .pra report from TERcom, presented in Fig. 6, shows the whole sequence of four actions applied to the four sentences, with the HER(pe,mt) setting. In this report, we can see the correct deletion in sentence 1 and the correct insertion in

sentence 2. In sentence 3, we can see the replacement of the word "incorrect" by the "corrected" word. Since the reference (the MT output) is compared to the hypothesis (the PE), these describe the real processes that we performed, with the edited sentences appearing in the second line, under "Original Hyp." The description of the edits in sentence 4 is more detailed and requires more explanation. We see in "Hyp after shifts" that the movement is simulated over the hypothesis, which, in the case of the HER setting, creates the opposite effect: the simulated shifts invert the order of the movement we produced. (In the similar report in the TER setting, the "Hyp after shifts" line shows the sentence with the order that results from our editing.) We can also see that after "Numshifts," the movement is reported as being from position 5 to 7, when it was in the opposite direction, from 7 to 5 (TERcom counts the first position as "0").

These simple experiments allowed us to identify issues in using an edit distance to approximate the editing process and to show that the inversion of the terms of comparison brings us one step closer to that. However, this may not be sufficient to have an accurate method to estimate editing. Furthermore, care should be taken when interpreting the results of an edit distance as mere edit counting and even when we explore the reports that an edit distance tool as TERcom presents.

We have commented on how misinterpretations may arise from terms that are not very clear. Terms that originated in the evaluation of MT hypotheses lose their original meaning when used in HER, so we suggest that they should be replaced by terms that describe the before and after of the editing process, such as "unedited" (to replace "reference") and "edited" (to replace "hypothesis"). Another suggestion to improve these edit distances is to check the information provided by the reports each tool creates, clarifying the direction in which the estimates are performed, so that a deletion refers to the process of deleting a word, and the movement is reported in the correct direction.

The purpose of our research (identifying correctly which words are edited and how) is not yet fulfilled at this stage. More tests are needed to know the precision of all edits estimated by a process like the one performed by TERcom.

Simulating more complex editing – After having concluded that HER is a more accurate estimate of editing, we wanted to know how precise could this be when editing became more complex. So, we set up a sequence of 50 edit patterns, applied to the same 15-word sentence, starting by applying one edit to a word and finishing by applying random patterns of the four edits.[13]

In these increasingly complex scenarios, we also want to test whether HER treats sequences of edits as one edit: if two contiguous words are deleted, is that one edit, or two edits? This is relevant because this action may identify a cohesive unit. The following tables present a visual comparison of the number of edit actions actually performed, together with the words effectively edited, followed by the edit scores

[13]We edited a new sentence, this time in Portuguese, which reads "Você pode acessar as configurações de tela selecionando Configurações a partir do menu principal."

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
1	Delete 1 word	1	1		1				1	6.67%
2	Insert 1 word	1	1	1					1	6.67%
3	Replace 1 word	1	1			1			1	6.67%
4	Move 1 word 1 position forward	1	1				1	1	1	6.67%
5	Move 1 word 1 position back	1	1				1	1	1	6.67%
6	Move 1 word 2 positions forward	1	1				1	1	1	6.67%
7	Move 1 word 2 positions back	1	1				1	1	1	6.67%

Fig. 7 Full HER(pe,mt) analysis of sentences with one edit

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
8	Delete 1 phrase (2 words)	1	2		*2				2	13.33%
9	Delete 1 phrase (3 words)	1	3		*3				3	20.00%
10	Insert 1 phrase (2 words)	1	2	*2					2	13.33%
11	Insert 1 phrase (3 words)	1	3	*3					3	20.00%
12	Replace 1 phrase (2 words)	1	2			*2			2	13.33%
13	Replace 1 phrase (3 words)	1	3			*3			3	20.00%
14	Move 1 phrase (2 words) 1 position forward	1	2				1	*1	1	6.67%
15	Move 1 phrase (2 words) 1 position back	1	2				1	2	1	6.67%
16	Move 1 phrase (3 words) 1 position forward	1	3				1	*1	1	6.67%
17	Move 1 phrase (3 words) 1 position back	1	3				1	*1	1	6.67%
18	Move 1 phrase (2 words) 2 positions forward	1	2				1	2	1	6.67%
19	Move 1 phrase (2 words) 2 positions back	1	2				1	2	1	6.67%
20	Move 1 phrase (3 words) 2 positions forward	1	3				1	*2	1	6.67%
21	Move 1 phrase (3 words) 2 positions back	1	3				1	*2	1	6.67%

Fig. 8 Full HER(pe,mt) analysis of sentences with one edit to phrases

presented by TERcom, set up as HER(pe,mt).[14] When the number of edit actions estimated did not correspond to the edit actions performed, we highlighted the cell and inserted a *.

Figure 7 presents a sequence of seven sentences each with one single editing action. Movement is tested in shifts (movements to contiguous positions) and in two positions and two directions, back and forward. In such simple situations, TERcom can correctly identify all the edits that were produced.

In the next stage of our tests, we experimented with editing phrases, i.e. sets of two or three contiguous words (see Fig. 8). It is interesting to see a very different behavior for the first three actions (deletion, insertion, and replacement) and for movement. We can see that TERcom counts deletions, insertions, and replacements according to the number of words that are edited, whereas movements, like a

[14] The titles of the columns in these tables are as follows. Under "EDITED": Actions: number of editing actions effectively performed; and Words: number of words effectively edited. Under TERCOM SCORES: I, D, S, and Sh: number of Insertions, Deletions, Substitutions, and Shifts estimated by TERcom; WdSh: number of words estimated as being shifted, NumEr: number of total edits or errors in segment; and HER: the HER score that results from dividing NumEr by the 15 words in the reference.

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
22	Delete 1 + 1 word (dift positions)	2	2		2				2	13.33%
23	Insert 1 + 1 word (dift positions)	2	2	2					2	13.33%
24	Replace 1 + 1 word (dift positions)	2	2			2			2	13.33%
25	Move 1 + 1 word one pos. fwd (dift positions)	2	2				2	2	2	13.33%
26	Delete 1 word + Insert 1 word (dift positions)	2	2	1	1				2	13.33%
27	Delete 1 word + Replace 1 word (dift positions)	2	2		1	1			2	13.33%
28	Delete 1 word + Move 1 word one pos. fwd (dift positions)	2	2		1		1	1	2	13.33%
29	Insert 1 word + Delete 1 phrase (2wd) (dift positions)	2	3	1	*2				3	20.00%
30	Insert 1 word + Replace 1 phrase (2wd) (dift positions)	2	3	1		*2			3	20.00%
31	Insert 1 word + Move 1 phrase (2wd) one pos. fwd (dift positions)	2	3	1			1	*1	2	13.33%
32	Replace 1 word + Delete 1 phrase (3wd) (dift positions)	2	4		*3	1			4	26.67%
33	Replace 1 word + Insert 1 phrase (3wd) (dift positions)	2	4	*3		1			4	26.67%
34	Replace 1 word + Move 1 phrase (3wd) one pos. (dift positions)	2	4			1	1	*1	2	13.33%
35	Delete 1 phrase (2wd) + Insert 1 phrase (2wd)	2	4	*2	*2				4	26.67%
36	Delete 1 phrase (2wd) + Replace 1 phrase (2wd)	2	4		*2	*2			4	26.67%
37	Delete 1 phrase (2wd) + Move 1 phrase (2wd) 2 positions	2	4		*2		1	2	*3	20.00%

Fig. 9 Full HER(pe,mt) analysis of sentences with a combination of two edits

shift operation in TER, count as one, even if we move two or three words. We can, furthermore, observe that the number of words moved (column WdSh) often describes problems in correctly identifying the words that were moved.

However, the most important observation in this table is the number of estimated edits (NumEr): this is a combination of number of edit actions (for movement) and number of edited words (for the other actions). This explains the lower HER scores for sentences with movement.[15] We only highlighted this column when this number was not the number of actions nor the number of edited words.

Next, we combined more than one edit in the same sentence. For example, in sentence 26 in Fig. 9, we deleted one word and then inserted a different word in another position. Note that a method that considers movement as a deletion followed by an insertion does not allow us to distinguish between what we are simulating in this sentence and a movement of a word.

Let us look at sentence 29, in Fig. 9. In this sentence, we inserted one word and then deleted a two-word phrase. According to the number of edits, this should count as two, but according to the number of edited words, this should be three. TERcom counts it according to the number of edited words. However, in sentence 31, TERcom's count is only correct if we look at the number of edits (in this case, two edits: an insertion and a movement, which affected three words). As we move down the table, we see that wrong information increases, especially in terms of the number of edits: where we count one edit, TERcom counts more, usually when an edit implies more than one word. The worst case is the example at the end of the table (sentence 37); here, two edits that affected four words are counted as three. This is explained by the combination of an edit that counts as two (the deletion of

[15] Note that we are always using the same 15-word sentence as reference, which is the denominator for all these estimates. Besides, the number of words shifted does not affect the calculation of the edit score.

		EDITED		TERCOM SCORES						
	ACTIONS	Actions	Words	I	D	S	Sh	WdSh	NumEr	HER
38	Delete 1 word + Insert 1 word + Replace 1 word	3	3	1	1	*2			*4	26.67%
39	Delete 1 word + Insert 1 word + Move 1 word 5 positions	3	3	*	*	*1	*2	*2	3	20.00%
40	Delete 1 word + Replace 1 word + Move 1 word 7 positions	3	3		1	1	1	1	3	20.00%
41	Replace 1 word + Insert 1 phrase (2wd)	2	3	*2		1			3	20.00%
42	Replace 1 phrase (2wd) + Insert 1 word	2	3	1		*2			3	20.00%
43	Replace 1 phrase (2wd) + Insert 1 phrase (2wd)	2	4	*2		*2			4	26.67%
44	Insert 1 word + Delete 1 phrase (2wd) + Replace 1 word	3	4	*	*1	*2	*1	*1	4	26.67%
45	Insert 1 phrase (2wd) + Delete 1 word + Replace 1 word	3	4	*2	1	1			4	26.67%
46	Insert 1 word + Delete 1 word + Replace 1 phrase (2wd)	3	4	1	1	*2			4	26.67%
47	Insert 1 ph (2wd) + Delete 1 ph (2wd) + Replace 1 ph (2wd)	3	6	*2	*2	*2			6	40.00%
48	Insert 1 word + Delete 1 ph (2wd) + Replace 1 word + Move 1 word	4	5	*	1	*2	*2	*2	5	33.33%
49	Insert 1 ph (2wd) + Delete 1 word + Replace 1 word + Move 1ph (2wd)	4	6	*2	1	1	1	2	*5	33.33%
50	Insert 1 word + Delete 1 word + Replace 1 ph (2wd) + Move 1 word	4	5	*	*	*3	*2	*2	5	33.33%

Fig. 10 Full HER(pe,mt) analysis of sentences with a combination of three or more edits

TOTALS	I	D	S	Sh	Edits	Words	NumEr	NumWd	HER
Performed	23	23	23	23	92	142		750	
HER scores	30	31	38	27			126	750	16.80%

Fig. 11 Total HER scores for the set of 50 simulated edited sentences

a two-word phrase) and an edit that counts as one (the movement of a two-word phrase).

In this last set of sentences, there are sentences with random combinations of two, three, or four editing actions. Figure 10 shows a much higher number of problems in identifying the number of edits performed. This is also where the first "ghost" edits appear, i.e. the number of estimated edits is higher than that actually performed. In sentence 38, an extra fourth action is estimated: a replacement was added to the correct insertion, deletion, and replacement. In sentence 49, five actions are estimated, one more than what was done, but the number of edited words is higher.

This demonstration included many examples of relatively simple issues, which accumulated to give us a good grasp of the limits to how we can interpret edit distances. Still, we could see how the problems in identifying the edits grow as the edit scores increase, especially above 20%.

Let us now check the global scores (see Fig. 11). We simulated the same number of edits for each type, because we wanted to check whether TERcom reflected this distribution. In the real world, this distribution is naturally rare. Several studies have identified a preponderance of replacements in user data (deletions are usually the second most frequent edit, then insertions and finally movement, which normally lags behind).[16] We can see in our data that HER overestimated all edits, but in an unbalanced way, replacement being the edit that sees the highest increase (65% more edits) and movement the one with the lowest increase (17%), compared to the

[16] Snover et al. (2006) reported fewer insertions than shifts, when they used TER for the first time, but well-balanced APE systems consistently report this proportion (Carmo et al. 2020).

actual edits performed. So, it seems that there is a bias in the edit distance estimation method toward replacement and against movement.

The total edit score for this set of sentences presented by TERcom was 16.80%. As we commented before, this score is calculated by dividing the total number of edits (NumEr: 126) by the total number of words (NumWd: 750). But the number of edits estimated by HER (NumEr) is neither the sum of editing actions performed nor the sum of edited words. Indeed, this score combines some editing actions, some edited words, and a certain number of incorrectly estimated edits. If we consider the number of actual edits performed (EdActions: 92), the estimated HER number of edits is overestimated by 37%. If we consider the number of edited words (EdWords: 142), it is underestimated by 11%.

Conclusions from the experiments – We can conclude from this set of experiments that edit scores, even when measured as HER, do not have the explanatory capacity we usually attribute to them. The simulations in our experiment naturally did not cover the full extent of editing that occurs in real-life projects. This being a very systematic and constrained simulation, our editing scores only reach 40% (and in a single edited sentence). In reality, editing work, which supposedly is reported by edit distances, does not follow such regular patterns as the ones we simulated, and it often results in a number of words which surpasses the 100% of words in the hypothesis.

4 Discussion

In the previous sections, we have discussed data collected by keyloggers and edit distances as subjects of academic research. Before we discuss the usefulness and reach of this research, it is important to broaden the discussion so that it encompasses current uses of these concepts and instruments outside research.

4.1 Current Uses of Process Data and of Error/Edit Rates

Process data collected by keyloggers is rarely used outside of research projects. To our knowledge, there are no companies collecting process data containing editing work patterns from translators and using these to develop their systems. The situation is very different for edit distances, since error and edit rates are used in many applications in research and in the industry.

In this volume, there are examples of research using edit distances as a tool that may help identify lexical similarities between languages, such as cognates (Heilmann and Llorca-Bofí, Chap. 8), or to identify shifts in syntactic structures (Vanroy et al., Chap. 10). The volume also includes research that highlights the limitations of edit distances, showing, for instance, how no metric can be used alone to measure effort (Cumbreño and Aranberri, Chap. 3).

It is important to note that the main issue of edit distances is not their shortcomings to describe the editing process. These shortcomings are reasonable and expected, since these are estimates, approximations, a heuristic of a process for which there is not enough data. The use of edit distances without awareness of their limitations is our main concern. This section debates a few examples of such uses.

Evaluation of quality – Edit distances, and TER in particular, are often used as proxies for human effort, in different evaluation initiatives, mainly in academic contexts, as in the APE shared tasks we alluded to before (Sect. 2.4). But in professional scenarios, edit distances are also generally considered, in conjunction with records of total time spent, as "meaningful data points to estimate translation effort" (Meer 2019).

The book "Translation Quality Assessment" (Moorkens et al. 2018) is a comprehensive and fundamental compilation of different perspectives on translation quality management. Error metrics and error typologies play an important role in this theme, both in MT evaluation and in human translation workflows. Popović (2018) presents an extensive list of error typologies used in research and by the industry and mentions that some unification and generalization are necessary. One of the suggestions of the paper is to think of PE as a form of error annotation, as if each edit identified an error. If there was such a transparency between edit and error, and if there was a one-to-one match between edit types and error typologies, it would be easy to collect error annotations from the PE activity alone. The fact that there are so many error typologies shows that PE and error annotation are different processes with different objectives, so adding error annotation to PE is in fact adding a new process to an already complex one.

The assumption that there is a one-to-one alignment between edit and error is intrinsic to automated evaluation, as this relies on another assumption, one that sustains that high-quality translation data has a very high equivalence between source and target units.[17]This reductionist view, that all linguistic units in a translation are bound by paradigms, being replaceable by equivalent units, leaves out complex relations sustaining the editing decision process. (Besides the paradigmatic level, there are, for example, syntagmatic relations, those that are determined by the relation between the words in the target sequence.)

Since editing processes are more complex than error rates are capable of describing, we advocate that evaluation of human translation should not rely so much on these rates and typologies. Instead, it should aim at being a step toward understanding and improving translation production processes. And while we do not claim that a study of editing as a technical operation allows us to study the whole complexity of the process, a consistent model of the surface level will provide a good foundation to other layers of analysis, a fundamental requirement to approach such a complex dimension as that of quality.

[17]This may be called an "artificial form of equivalence" (Pym 2010, 135); Carl (2009) refers to this as an isomorphism, which only exists in MT, between alignment units and production units.

Interactivity – After collecting learnt models of processes, these can be used for the development of systems in which users interact with MT (Carl and Planas 2020). Most of these interactive environments, like TransType (Langlais et al. 2000) or Lilt (Wuebker et al. 2016), apply predictive writing aids that conform to left-to-right translating. This form of prefix-based interactive translation has also been recently tested with success as an application of NMT (Knowles et al. 2019; Domingo et al. 2019). Some of the experiments in this field, however, break away from predictive writing, and they model editing according to the four edits. Some examples are CATalog (Pal et al. 2016)—which logs the four editing actions for use in project management and APE, Kanjingo (Moorkens et al. 2016)—a mobile platform for editing, and LFPE (Simianer et al. 2016)—which has only been used for research purposes. Domingo et al. (2017) developed a prototype that allowed for the application of the four actions. However, all work in this system were simulated by an automated process that followed a programed left-to-right word editing sequence. It is still not clear to what extent knowledge from the editing process can be used to develop these systems.

Prediction – Interactive translation systems aim at supporting the user by predicting the next step, so as to provide a useful suggestion and thus reduce the effort of the translator (Macklovitch et al. 2005). In 2009, Carl commented on the purpose to make TPR data productive enough to enable prediction of editing behaviors (Carl 2009, 245). Ten years later, the "predictive turn" seemed to have been officially reached (Schaeffer et al. 2019). However, TPR data has a descriptive nature, and it needs to be carefully processed before it can be used to build effective predictive models. Schaeffer et al. (2019, 19) claim that "the main obstacle for implementing a model which is capable of predicting both the product, that is, the target text, and the human cognitive processes which led to the latter, is that those researchers who model the product chase the human gold standard in terms of its quality." In this chapter, our focus is not on the cognitive process or on a notion of quality, but still it is not easy to see how process data can be engineered to predict even the next step in an editing process.

Prediction is also the purpose of applied tasks of MT methods such as APE and Quality Estimation (QE) of the output of MT systems (Specia et al. 2018). Both of these tasks rely on metrics like TER to measure and improve the accuracy of their systems and models.

Post-editor modeling – MT researchers also want to know how translators and post-editors work. One of the applications of the knowledge acquired is to build post-editor profiles, also called models or continuous representations of post-editors. The paper that presented "Translator2Vec" (Góis and Martins 2019) deserves a careful analysis, as it constitutes a good example of sharing of knowledge about the application of editing process data in automated applications.[18]

[18]The authors have released their dataset, which may be used for further research: https://github.com/Unbabel/translator2vec/.

The formalization of the process that was used to learn editor models in this project, by using modern representation techniques, makes it possible to embed the knowledge of the specific representations of each post-editor in a vector space, which can later be integrated to design better interfaces, to improve QE and APE, and to better allocate human resources in a production workflow. This project started with keystroke data, because, as the authors say, one cannot learn how translators work by looking at the initial and the final text. This data was then converted into what the authors call "action sequences" to describe that process: "Overall, five text-editing actions are considered: inserting (I), deleting (D), and replacing (R) a single word, and inserting (BI) and deleting (BD) a block of words" (Góis and Martins 2019, 45). "BI" and "BD," assuming that they refer to the same edited unit, naturally correspond to movement. The data also includes mouse actions, relevant to detect selected words and repeated editing.

The paper concludes that the representations of action sequences are effective predictors of editing time, that they can be used to identify and cluster human post-editors by behavior, and also that this data can be used to improve other tasks. We believe that the paper also shows that there is still a lot to learn about such a simple process as editing, from which it follows that it is premature to say that computer-aided translation is shifting toward "human-aided translation" by which the translation process is performed by the machine and the "human in the loop only intervenes when needed" (Góis and Martins 2019, 43).

4.2 Best Methods to Study Editing

We believe that there is consensus in that four editing actions are an adequate form of describing editing work. Why four? Because these fulfill all possibilities according to the discrete properties of unit and position, as explained at the end of Sect. 2.2. Two-edit models, which only consider deletion and insertion, do not allow for a distinction between a movement and a deletion of a word that coexists with an insertion. As for edit metrics with more than four operations, like TERp, these usually add sub-types of substitution, by resorting to manipulations of unit length or to levels of linguistic analysis, going below the surface level in which edits occur. Proposals of character-based edit distances, namely CharacTER (Wang et al. 2016) and chrF (Popović 2015), have shown their relevance for evaluation, especially for morphology-rich languages, but for the description of editing they can only help by adding stem matching, which is a sub-type of substitution.

We analyzed TPR logs, which convey the fundamental information about what goes on in editing. These logs also have the added value of including chronological information. We may use that information to solve questions in our research: if we are measuring effort or difficulty, more than looking at the products of a translation process, we should study the editing actions that led to that final translation product. To do that, we need models that incorporate micro- and macro-units (Sect. 2.3).

Another criterion to consider in a discussion about the best methods to describe editing is the perspective on errors or edits. These terms are two perspectives over the same reality, so it is easy to fall into equivocal statements that may jeopardize the descriptive capacity of our instruments of analysis. An illustration of this is the decision to cap TERp at 100% (see Sect. 2.4). This decision is explained in these terms: "[A score above 1] is unfair, since the hypothesis cannot be more than 100% wrong" (Snover et al. 2006, 6). We might agree with this, from a strictly quantifiable perspective. However, editing may increase the length of sentences, not only because one word may be replaced by two or more words, but also because one word may contain more than one error or edit (which is visible when we consider several, equally good, references). Therefore, a distance that aims at describing editing effort cannot be constrained by the 100% limit. A capped metric like TERp may be more useful as an error rate, but it has a diminished value as an edit rate, since it explicitly disregards any editing effort that creates more words than the ones the hypothesis contains.

Error rates are naturally useful. If we are only interested in obtaining edit scores, we can decide that it does not make a big difference if we use TER or HER. For comparison of static data, like in genome decoding, this is indeed irrelevant. Furthermore, the use of TER previously mentioned, as the main metric for system comparison in many evaluation tasks, shows that a metric does not need to be rigorous in the details of what it aims to describe; as long as it is consistent and considers all systems equally, it can be used to compare systems. But there are other situations in which this use of metrics as evaluation tools goes beyond system comparison.

If we are studying a dynamic process, it is relevant to get as close as possible to what happens in that process. Using the metric that best shows the number of edits, which types of edits and which words were edited, is an important step toward that goal. The main purpose of suggesting an inverted metric like HER is to move the focus from errors to edits, which is a relevant move when that is our object of study. Besides, an edit rate presents an improved perspective on actual editing, and when researchers aim at improving process estimation methods, any process that brings them closer to that process should be preferred.

4.3 Open Questions for Research

Analyzing complex editing – The fact that editing in real scenarios is non-linear (Sect. 2.2) complicates any method of collection and interpretation of process data.

An editing model based on discrete properties can allow for a progressive analysis of different dimensions of the complexity of editing. Information collected on the editing process can be used separately to learn units of editing and patterns of manipulation of positions. Length and time can then be used to study other levels of complexity in editing. Time may be used to work with repetitive edits to the same units, according to the micro-/macro-unit model; length may serve to guide studies

on manipulation of segmentation, when words are transformed into phrases, clauses, or even sentences.

There are other approaches to complex editing. When there are multiple references, we can approach those as different paradigms, options to fill in equivalent positions, but other edits are propagated throughout a sentence. A syntagmatic approach may help us understand the effects of movement: does the movement of a phrase imply more changes to other units in the same sentence?

One of the features that edit distances provide is manipulation of costs. We may use costs to study differences between the four editing actions. What if we make secondary actions costlier than primary ones (namely, because they may force other edits, by syntagmatic contamination)? Or if we make replacement less expensive, so that we reduce its weight in the models? And what effects will there be over our metrics if we standardize costs for phrase-level edits for all actions, not just movement? Another possibility is to assign weights to edits by adding a factor related to the time each one takes to perform or to the pause before it is applied. We can also study models of costs that vary according to the number of paradigmatic alternatives available for each unit. These alternatives may be revealed by the use of several references or by macro-units that report a translator replacing the same word several times.[19]

In our TER/HER model of editing, we did not consider the semantic and linguistic content affected by these actions, because we see edits as units of information in an engineering problem, like Shannon in his Information Theory (Shannon 1948). Linguistic information may be integrated in further analyses, but, before we do that, we should make sure that our models of actions happening on surface units are as accurate and useful as possible. With a well-explored and detailed model, it is easier to see that the addition of linguistic information to a surface analysis would ideally be done consistently across the whole model: adding lexical, semantic, or other linguistic information to word replacement is just as valuable to the study of movement, deletion, or insertion.

This strategy might seem a choice for the easiest path, but even the definition of a surface unit is not exempt of difficulties. The transformation created by editing may highlight units that were not considered otherwise – that is what Toury refers to when he includes manipulation of segmentation in his list of matricial norms (see Sect. 2.1). In fact, when confronted with a word that is transformed into a phrase or a clause, or when the concept such a word conveyed is spread into different words, occupying scattered positions in the same sentence, we are faced with difficulties in describing not just the unit but also the process. We can therefore claim that even at the surface level, in which our study of editing occurs, there is a complexity that is still not adequately explored by research.

[19] Schaeffer et al. (2016) call this gamut of choices "word translation entropy," which is an approach that is explored in detail in other chapters of this volume, namely by Wei (Chap. 7) and Carl (Chap. 5).

Editing may be complex enough when we work with only one language, replacing and acting on words and sentences to improve and correct them, without concerns of how these relate to other languages. But when we consider the word as the unit of editing in the context of translation, we must recall that word is not synonym with unit of alignment, i.e. there is no one-to-one relation between words in two languages. Studying editing in translation raises several complex questions. Lacruz et al. (this volume, Chap. 11), for example, identify a specific unit in Japanese that cannot be aligned with one word in English, and Carl (this volume, Chap. 9) discusses segmentation of units as a cognitive process. Furthermore, any research that deals with processes of alignment or with the concepts of literality and equivalence between source and target words and sentences, bring to the fore the complexity of the actions we perform when we work with words that establish relations, not just paradigmatic and syntagmatic within one language, but crossing levels between languages. Huang and Carl (this volume, Chap. 2), Vanroy et al. (this volume, Chap. 10), and Carl (this volume, Chaps. 5 and 13) evidence the complexity of tapping into the different dimensions of these processes.

How informative can these models be? – One of the difficulties in learning action models is that actions alone may not be informative enough, and when we add content to those actions (lists of the words that were deleted, inserted, etc.), they become too sparse. Previous studies have highlighted how edits tend to be unique and inconsistent (Wisniewski et al. 2015).

When we have informative models from our editing process data that include word-based editing actions, maybe we will be able to build predictive models, as Góis and Martins (2019) propose. If these models are adapted to human activities, maybe we can even use them to inform our translator training programs, helping future translators become more efficient (post-)editors. Still, this is an open question, and more research is needed to evaluate the usefulness of these models.

4.4 Why Should TPR and MT Communicate?

If process data answered all the questions raised by TPR, and if MT could tap into the process from bilingual data, each discipline could work with its own models and methods, and it would easily fulfill its objectives. In the current situation, what is there to gain for each approach from communicating with the other, through the lens of edit rates?

The best method to study editing is unlikely to rise from developing either process logs or edit distances in isolation. We included in our analysis several details that developers may use to improve the use of edit distances in translation. Examples of suggestions concern the bias toward replacement, and the way movement is processed and reported differently from the other edits and the combined final score, which is based on neither number of edits nor number of edited words. Most of these

details come from the analysis of the actual editing process, instead of relying on the capacity to learn from product data.

Gains for TPR from edit distances – Actual process data, as the one collected by CRITT TPR-DB, is more reliable and accurate than any estimates of editing work. However, this process data is more easily interpreted and applicable after it is converted into units of action that are associated with units of form, as the example of edit distances shows. An investment in research into the discrete properties of units, positions, lengths, and time may help TPR gain a higher capacity to explain the complexity of the translation process. This may be improved by increasing the availability of process data interpreted by these methods. Another positive effect may be the increased reuse of the data produced by TPR, in MT research and outside of academia, eventually even in the development of translation support systems.

Besides, this data may be applied to the exploration of how translators work between source texts, target texts, MT output, and other support resources. It is very different to work on tasks performed over target text (like editing) and to work on a process that goes from source to target text (like translation). Equivalence in the first case is established between units in the target language; in the second case, it must be established between units in two different languages. For translation, equivalence cannot be established at the surface level, so process models must be more complex.

An approach that looks for aligned units by collecting editing actions may help us understand how these units are selected: do translators choose words that share the same stem to over-type, or do they select units of meaning, or cohesive syntactic groups? With a model of editing actions inspired by MT methods, TPR may try to answer questions like these.

Gains for MT for incorporating process data – APE and interactive translation are two of the areas of MT in which we may find citations from Translation Studies literature, most of it involving TPR (see, for example, Góis and Martins (2019)). Still, the models and methods that are employed most frequently are not determined by the knowledge gained from these studies. The discourses about edit processes are usually only employed to interpret results, while the methodologies are oriented by the evaluation of MT errors. As we have seen, there would be great gains if MT researchers could use more efficiently the knowledge about the processes of translation that TPR contains.

Furthermore, there are several examples of studies that, useful as they may be, are still supported by simulations and estimations, when it is recognized that these are approximations to real process data. The assumed cost of tapping into this data may be reduced if more contact with process studies is maintained.

Popović et al. (2016) have demonstrated that to tune a system, the best data comes from the original PE, with the caveat that other data must be added to avoid overfitting. Forcada (2017a) has also called for more tuning and optimization of systems, to avoid some of the haphazard circles that MT research seems to fall in. A good deal of pressure is created on method and model development in the pursuit of correlations with human judgment. The use of reliable data from actual process could alleviate some of these issues, replacing imperfect proxies, helping

tune systems, and establishing the correlation with human processes, instead of aiming at it during system evaluation.

5 Closing Remarks

As so many other disciplines that deal with the same subjects but which do not interact much, maybe all it takes for more communication to happen between TPR and MT is a common language. On one of the sides, we have accumulated knowledge on translation processes and, on the other, efficient and intelligent methods to learn from data. The units of action in edit distances may provide the common vocabulary and the focus on themes that benefit both TPR and MT.

This chapter discussed some of the reasons for the limited development of common projects about the translation process, and it presents suggestions and questions that justify more investment in such projects. The suggestion to transform error rates into edit rates (TER becoming HER) is a contribution to improve the interpretation capacity of our methodologies. Research on the translation process in Translation Studies may gain from adopting models that have proven its effectiveness in automated tasks in MT. Furthermore, TPR may help research and application of MT models achieve a better grounding on knowledge of human processes. In our view, the communication between these two approaches, more than contributing to automating and speeding processes, will give us better insight into how complex these processes are. Ultimately, it will contribute to our understanding of translation as a process at the scale of human complexity.

Acknowledgments The author thanks the collaboration of Maja Popović, Leonardo Zílio, Sabrina Girletti, Michael Carl, and the anonymous reviewers of the chapter, for the help with technical issues and the very useful comments.

References

Alves F, Hurtado Albir A (2010) Cognitive approaches. In: Gambier Y, van Doorslaer L (eds) Handbook of translation studies – volume I. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp 28–35

Alves F, Vale D (2009) Probing the unit of translation in time: aspects of the design and development of a web application for storing, annotating, and querying translation process data. Across Lang Cult 10(2):251–273

Aziz W, Castilho S, Specia L (2012) PET: a tool for post-editing and assessing machine translation. In: Proceedings of the eighth international conference on language resources and evaluation. European Language Resources Association (ELRA), Istanbul, pp 3983–3987

Bojar O, Chatterjee R, Federmann C, Haddow B, Huck M, Hokamp C, Koehn P, Logacheva V, Monz C, Negri M, Post M, Scarton C, Specia L, Turchi M (2015) Findings of the 2015 workshop on statistical machine translation. In: Proceedings of the tenth workshop on statistical machine translation (WMT 2015), Lisbon, pp 1–46

Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Yepes AJ, Koehn P, Logacheva V, Monz C, Negri M, Neveol A, Neves M, Popel M, Post M, Rubino R, Scarton C, Specia L, Turchi M, Verspoor K, Zampieri M (2016) Findings of the 2016 conference on machine translation (WMT16) 2:131–198

Carl M (2009) Triangulating product and process data: quantifying alignment units with keystroke data. In: Mees I, Göpferich S, Alves F (eds) Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen, Samfundslitteratur, Copenhagen, pp 225–247

Carl M (2012) Translog-II: a program for recording user activity data for empirical reading and writing research. In: Proceedings of the eighth international conference on language resources and evaluation. European Language Resources Association (ELRA), Istanbul, pp 4108–4112

Carl M, Jakobsen AL (2009) Towards statistical modelling of translators' activity data. Int J Speech Technol 12(4):125–138

Carl M, Planas E (2020) Advances in interactive translation technology. In: Ehrensberger-Dow M, Massey G, (eds) The bloomsbury companion to language industry studies. Bloomsbury, London, pp 361–386

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer Science+Business Media, Cham, pp 13–54

do Carmo F (2017) Post-editing: a theoretical and practical challenge for translation studies and machine learning. Ph.D. thesis, Universidade do Porto. https://repositorio-aberto.up.pt/handle/10216/107518

do Carmo F, Shterionov D, Wagner J, Hossari M, Paquin E, Moorkens J, Way A (2020) A review of the state-of-the-art in automatic post-editing. Machine Translation. https://doi.org/10.1007/s10590-020-09252-y

Dam-Jensen H, Heine C, Schrijver I (2019) The nature of text production – similarities and differences between writing and translation. Across Lang Cult 20(2):155–172

Damerau FJ (1964) A technique for computer detection and correction of spelling errors. Commun ACM 7(3):171–176

Domingo M, Peris A, Casacuberta F (2017) Segment-based interactive-predictive machine translation. Mach Transl 31(4):163–185

Domingo M, García-Martínez M, Alvaro Peris, Helle A, Estela A, Bié L, Casacuberta F, Herranz M (2019) Incremental adaptation of NMT for professional post-editors: a user study. In: Proceedings of machine translation summit XVII volume 2: translator, project and user tracks. European Association for Machine Translation, Dublin, pp 219–227

Dorr B (2011) Part 5: machine translation evaluation. In: Olive J, Christianson C, McCary J (eds) Handbook of natural language processing and machine translation. Springer, New York, pp 801–894

Forcada M (2017a) Is machine translation research running around in circles? In: CCL25 – 25 years of the Centre for Computational Linguistics at KU Leuven, Leuven. http://www.ccl.kuleuven.be/CCL25/slmikel.pdf

Forcada ML (2017b) Making sense of neural machine translation. Transl Spaces 6(2):291–309

Góis A, Martins AFT (2019) Translator2Vec: understanding and representing human post-editors. In: Proceedings of machine translation summit XVII volume 1: research track. European Association for Machine Translation, Dublin, pp 43–54

Hokamp C, Liu Q (2015) Handycat: the flexible CAT tool for translation research. In: Proceedings of the 18th annual conference of the European association for machine translation. European Association for Machine Translation, Antalya, p 216

Knowles R, Sanchez-Torron M, Koehn P (2019) A user study of neural interactive translation prediction. Mach Transl 33(1–2):135–154

Kollberg P, Eklundh KS (2002) Studying writers' revising patterns with S-notation analysis. In: Olive T, Levy CM (eds) Studies in writing: volume 10. Contemporary tools and techniques for studying writing. Kluwer Academic Publishers, Dordrecht, pp 89–104

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. The Kent State University Press, Kent, London
Lacruz I (2017) Cognitive effort in translation, editing, and post-editing. In: Schwieter JW, Ferreira A (eds) Handbook of translation and cognition. Wiley, Malden, pp 386–401
Langlais P, Foster G, Lapalme G (2000) Transtype: a computer-aided translation typing system. In: Embedded machine translation systems – workshop II. Seattle, Washington, pp 46–51
Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the second workshop on statistical machine translation, Prague, pp 228–231
Leijten M, Waes LV (2013) Keystroke logging in writing research: using inputlog to analyze and visualize writing processes. Writ Commun 30(3):358–392
Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 10(8):707–710
Lopresti D, Tomkins A (1997) Block edit models for approximate string matching. Theor Comput Sci 181(1):159–179
Macklovitch E, Nguyen NT, Lapalme G (2005) Tracing translations in the making. In: Proceeding of the MT Summit X, Phucket, pp 323–330
van der Meer J (2019) Eight years of DQF bring us closer to fixing the operational gap. https://blog.taus.net/eight-years-of-dqf-bring-us-closer-to-fixing-the-operational-gap
Moorkens J, Way A (2016) Comparing translator acceptability of TM and SMT outputs. Balt J Mod Comput 4(2):141–151
Moorkens J, O'Brien S, Vreeke J (2016) Developing and testing Kanjingo: a mobile app for post-editing. Tradumàtica: Tecnologies de la traducció (14):58–66
Moorkens J, Castilho S, Gaspari F, Doherty S (2018) Translation quality assessment: from principles to practice. Springer International Publishing AG, Cham
Moran J, Saam C, Lewis D (2014) Towards desktop-based CAT tool instrumentation. In: Third workshop on post-editing technology and practice. In: The 11th conference of the association for machine translation in the Americas, pp 99–112
Niessen S, Och FJ, Leusch G, Ney H, Niessen S, Och FJ, Leusch G, Ney H (2000) An evaluation tool for machine translation. In: Proceedings of the second international conference on language resources and evaluation. European Language Resources Association (ELRA), Athens, pp 39–45
Pal S, Zampieri M, Naskar SK, Nayak T, Vela M, Van Genabith J (2016) CATaLog online: porting a post-editing tool to the web. In: Proceedings of the international conference on language resources and evaluation. European Language Resources Association (ELRA), Portoroz, pp 599–604
Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics (ACL 2002), 6–12 July 2002, Philadelphia, pp 311–318
Popović M (2011) Hjerson: an open source tool for automatic error classification of machine translation output. Prague Bull Math Linguist 96(1):59–68
Popović M (2015) chrF: character n-gram f-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation (WMT@EMNLP 2015), Lisbon, pp 392–395
Popović M (2018) Error classification and analysis for machine translation quality assessment. In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds), Translation quality assessment: from principles to practice. Springer International Publishing AG, Cham, pp 129–158
Popović M, Arčan M, Lommel AR (2016) Potential and limits of using postedits as reference translations for MT evaluation. Balt J Mod Comput 4(2):218–229
Pym A (2010) Exploring translation theories. Routledge, Abingdon/New York
Schaeffer M, Dragsted B, Winther Balling L, Carl M (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer Science+Business Media, Cham, pp 183–210

Schaeffer M, Nitzke J, Hansen-Schirra S (2019) Predictive turn in translation studies: review and prospects. Springer International Publishing, Cham, pp 1–23

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Shapira D, Storer JA (2007) Edit distance with move operations. J Discrete Algoritm 5(2):380–392

Simianer P, Karimova S, Riezler S (2016) A post-editing interface for immediate adaptation in statistical machine translation. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations, Osaka, pp 16–20

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation of the Americas (AMTA 2006). Visions for the Future of Machine Translation, Cambridge, pp 223–231

Snover M, Madnani N, Dorr B, Schwartz R (2009a) TER-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Mach Transl 23(2–3):117–127

Snover M, Madnani N, Dorr BJ, Schwartz R (2009b) Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In: Proceedings of the fourth workshop on statistical machine translation. Association for Computational Linguistics, Athens, pp 259–268

Specia L, Scarton C, Paetzold GH (2018) Quality estimation for machine translation, vol 11. Morgan and Claypool, San Rafael

Tillmann C, Vogel S, Ney H, Zubiaga A, Sawaf H (1997) Accelerated DP based search for statistical translation. In: European conference on speech communication and technology, Rhodes, pp 2667–2670

Toury G (1995) Descriptive translation studies – and beyond. John Benjamins Publishing Company, Amsterdam/Philadelphia

Wang W, Peter JT, Rosendahl H, Ney H (2016) CharacTer: translation edit rate on character level. In: Proceedings of the first conference on machine translation: volume 2, Shared Task Papers. Association for Computational Linguistics, Berlin, pp 505–510

Wisniewski G, Pécheux N, Yvon F (2015) Why predicting post-edition is so hard? Failure analysis of LIMSI submission to the APE shared task. In: Proceedings of the tenth workshop on statistical machine translation, Lisbon, pp 222–227

Wuebker J, Green S, DeNero J, Hasan S, Luong MT (2016) Models and inference for prefix-constrained machine translation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, pp 66–75

Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort

Jie Huang and Michael Carl

Abstract Estimating post-editing effort is essential to identify translation difficulties and decide the payment for post-editors. Keystrokes, fixations, and production duration, as well as lexical and syntactic variations of the translation product, are frequently used as indicators of post-editing effort. This chapter introduces Word-based Human Edit Rate (WHER), a measure derived from HTER, as a new predictor to measure post-editors' effort on the word-level. The original HTER metric calculates an edit distance between MT output and its post-edited version from the minimum number of assumed edit operations. The WHER metric matches these edit operations to the corresponding words in the TT segment and maps them via alignment links to ST words. WHER thus provides the minimum number of expected edit operations for each ST word given the MT output. The chapter describes an experiment in which 21 student translators were invited to post-edit audiovisual texts and their translation processes were recorded with eye-tracking and keystroke-logging devices. After correlating WHER operations with the other common effort indicators derived from the process and product, we find that WHER is a reliable predictor for word-level post-editing effort.

Keywords Word-based Human Edit Rate (WHER) · Cognitive effort · Product analysis · Process analysis · Post-editing effort

1 Introduction

Several studies using MT for audiovisual texts have shown that the direct MT output cannot meet high-quality standards in the domain of audiovisual products (Armstrong et al. 2006, Melero et al. 2006,\vadjust{\pagebreak} Volk 2008, Bywood

J. Huang (✉)
Renmin University of China, Beijing, China

M. Carl
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_2

et al. 2013, Burchardt et al. 2016). However, MT post-editing can reduce translators' effort and increase productivity (de Sousa et al. 2011, Ortiz-boix and Matamala 2016, DePalma et al. 2019). Estimating post-editing effort relates to the identification of translation difficulty (e.g., Dragsted 2012; see also Vanroy et al. this volume, Chap. 10) and also impacts the pay rates of post-editors (Vieira 2014).

According to Krings (2001), the post-editing effort consists of three aspects: technical, temporal, and cognitive effort. Technical and temporal effort can be measured by the number of keystrokes typed and the time spent (Dragsted 2012, Jia et al. 2019). However, cognitive effort cannot be directly observed but only estimated through the process of reading and writing and also post-edited texts (Campbell 2017). For example, eye-tracking techniques are used to collect reading activities (Koglin 2015, Vieira 2016a), and the lexical and syntactic variations of post-edited texts are measured to reflect the cognitive effort in writing (Nitzke 2019, Vanroy et al. 2019).

This study defines a new product-based measure to assess translation difficulty, the Word-based Human Edit Rate (WHER). WHER is a derivation of HTER which measures the minimum edit distance between MT output and its post-edited version on the sentence level (Snover et al. 2006). Do Carmo (this volume, Chap. 1) inverts the reference and the hypothesis in the computation of HTER which makes the result better interpretable to assess post-editing effort and calls his new measure HER. In this chapter, we learn from do Carmo's practice and extend HER into WHER by mapping the edit operations in the TT words via word-alignment links to the equivalent ST positions. We assess to what extent WHER can be used to indicate post-editing effort as exerted during the process of post-editing. While keystroke data is recorded during the post-editing process to indicate the "real" amount of technical effort, WHER measures the number of minimum TT edit operations per ST word. We, thus, expect to find a correspondence between the minimum possible and the really performed edit operations.

The objective of this study is to assess the extent to which WHER correlates with post-editing behavior and, thus, may be suited to estimate post-editing effort. In addition, the WHER score might point to positions that are difficult to post-edit, which can be helpful to evaluate post-editors' effort without observations from the post-editing process. Another potential of WHER is to predict whether a word is correctly translated by the MT system and thus helps estimate the MT quality. Translation quality estimation (QE) is increasingly important in developing Natural Language Processing (NLP) and MT engines; various models have been created to fulfill the task without human ratings (Martins et al. 2017, Basu et al. 2018, Xenouleas et al. 2019). In the following section, we give an overview of frequently used effort indicators in the translation process and product. In Sect. 3, we present our experimental setting, data collection method, and WHER computing method. The correlation between WHER and several effort indicators of post-editing effort will be discussed in Sect. 4, and the key findings and future scenarios for WHER usage will be summarized in Sect. 5.

2 Related Research

Three approaches are commonly used to estimate effort in translation which are sometimes combined to triangulate the data. The first approach is to observe reading and writing activities from keystroke-logging and eye-tracking data (Dragsted 2012, Koglin 2015, Vieira 2016a). The second approach is to estimate the effort from the post-edited results. Lexical and syntactic variations of the TT indicate cognitive effort in translation production, which has been addressed with multiple measures including word-level entropy scores such as HTra and Word Distortion Entropy (HCross) (Carl et al. 2016). The last approach is subjective ratings from human post-editors (de Sousa et al. 2011, Vieira 2016b). As a traditional method, subjective reflection is used to elicit the perceived effort of post-editors during the translation or post-editing task (Moorkens et al. 2015).

However, previous studies found that a single measure from any of the above approaches is not robust enough to explain the cognitive effort (Koponen 2012, Guerberof 2014). Some measures are more sensitive to individual differences than others (Vieira 2016a). Also, there may not be strong correlations between the different approaches to measuring cognitive effort. For example, average fixation duration per sentence and subjective ratings are not strongly associated with each other in correlation tests and principal component analyses (Schaeffer and Carl 2014, Vieira 2016a). That is why multiple approaches are normally used together to measure the post-editing effort. In this section, we present commonly used effort indicators and related studies from the approaches of process and product analysis. As human subjective ratings concern larger segments and are usually not applicable to word-level analysis, the last approach is less relevant to this study and not further discussed.

2.1 Process Indicators

The first approach is to estimate post-editors' effort by their eye movements and keystroke behavior during reading and writing activities. Detailed information including keystrokes (e.g., number of insertions, number of deletions), fixations (e.g., number of fixations on ST/TT, duration of fixations on ST/TT). and duration (e.g., total production duration) are commonly used metrics in translation and writing studies (Mossop 2007, Dragsted 2012, Leijten and Van Waes 2013, Koglin 2015).

Keystroke information, such as the number of insertions and deletions and the total number of all keystrokes, are related to the technical effort of post-editors as they reflect "the actual linguistic changes to correct the machine translation errors" (Koglin 2015: 129). Naturally, the more MT output is modified, the more effort is required during post-editing. Eye movements reflect reading behavior on the ST and TT. According to the eye-mind hypothesis, fixations are usually linked to attention

(Just and Carpenter 1980) such that the attention follows eye movements. Empirical translation studies have shown that the average fixation duration and the number of fixations per word correlate with other effort indicators such as the pause-to-word ratio and production duration per word (Vieira 2016a). Also, the first fixation duration and the total fixation duration are used as indicators of effort (Schaeffer et al. 2019). Research show that differences are observed in post-editors' fixation behavior, with TT usually attracting more attention than the ST (Sanchis-Trilles et al. 2014, Vieira 2014).

2.2 Product Indicators

Lexical variations and syntactic distortions (reordering) in the TT are indicators of post-editing effort, as they reflect activities of text modification or structure adjustment during post-editing (see, e.g., Vanroy et al. (this volume, Chap. 1); Lacruz et al. (this volume, Chap. 11); Ogawa et al. (this volume, Chap. 6)). The larger the number of alternative translations, the more effort is expected for post-editors to make the modification. Similar logic applies to the syntactic variations, which are measured by the vectorized word sequence distortion from the ST to the TT sentence. The concept of entropy (Carl this volume, Chap. 5), borrowed from information theory, is used to show the degree of translation variations of each ST word (Schaeffer and Carl 2014, Schaeffer et al. 2016). The features are generated by the analysis toolkit integrated into the CRITT TPR-DB, a large repository of translation process data (Carl et al. 2016). Overall, the indicators of lexical and syntactic modifications of the MT output give us a glimpse of post-editors' effort for each ST word, which is facilitated by the word-level analysis.

AltT, ITra, and HTra are measures of lexical translation choices. AltT is the number of alternative translations for an ST word in a given context across different participants and sessions. Higher AltT values indicate a wider range of TT words corresponding to the ST word. ITra is the self-information of a translation, computed as ITra $= \log_2(1/\text{ProbT})$, where ProbT is the probability of the translation, as provided in the CRITT TPR-DB table. Higher ITra values correspond to higher information content indicating that the current TT word is less frequently used. These two indicators can be calculated after the alignment of ST and TT words (or phrases) to display the translation choices of each participant for each ST word. HTra is the word translation entropy that multiplies the sum of ITra with their expectation (Schaeffer et al. 2016, Carl this volume, Chap. 5). The three measures differ in the sense that AltT and HTra values are identical among participants on the same ST token, while ITra is a participant-specific metric that relates to a particular translation. Previous studies have found that the HTra value correlates with process measures such as the duration of production, the number of insertions, and the number of fixations on ST (Bangalore et al. 2016, Vanroy et al. 2019, Wei this volume, Chap. 7; Lacruz et al. this volume, Chap. 11). It shows that the measures of lexical variations are robust and reliable to estimate post-editing effort.

Cross and HCross are measures for syntactic distortions. Cross is a vector of relative cross-linguistic distortion of word position between ST and TT. Its absolute value thus indicates the degree of adjustment made on the sentence syntactic structure. HCross is based on the Cross values of all participants on the same ST word and is calculated as the word order entropy (Carl and Schaeffer 2017). If the HCross value of a given ST word is 0, it means that the same relative translation re-ordering is chosen by all participants in the TT (Nitzke 2019, Carl this volume, Chap. 5). On the contrary, a high HCross value implies a larger variance among participants in choosing different relative TT word position. Therefore, Cross is a participant-unique and session-unique metric, and HCross is unique among ST words only. The two metrics are both used to measure the effort of adjusting sentence syntactic structures during post-editing.

3 Method

3.1 Material

Twenty-one participants post-edited extracts of two audiovisual texts on the topic of law. One text was a documentary film, and the other was an episode of TV drama. In each text, two extracts of comparable length and durations were selected and allocated to participants in a random sequence. Although legal texts may be more specialized than general texts, the pieces did not contain difficult terminologies. For the material to be as authentic as possible for the sake of ecological validity (Orero et al. 2018), all selected scenes were self-contained clips with subtitles referring to a complete scenario. The documentary extracts have a mean duration of 25 s (sd = 3) and a mean length of 53 words (sd = 1). The drama extracts have a mean duration of 36.5 s (sd = 2.5) and a mean length of 96.5 words (sd = 2.5). As both subtitles and videos of source material are required for the professional working process of AVT (Díaz Cintas and Remael 2014), the corresponding video clip with subtitles in the SL was played twice before participants started the task. The video-watching process was not eye-tracked or key-logged. No time constraint was set on participants' post-editing process.

3.2 Participants

All 21 participants were Masters students of translation. All of them were Chinese native speakers with English as the second language, between 22 and 32 years old. They had a similar level of English proficiency, with 81% having passed the China Accreditation Test for Translators and Interpreters (CATTI) between the English and Chinese language pair and all the rest scoring over 7 in International English

Language Testing System (IELTS).[1] None of them had professional experience or training on AVT or post-editing. Therefore, they were targeted as novice translators without expertise in AVT or post-editing. Our sampling method considers that post-editing audiovisual texts is currently not a common practice for professionals, so we only selected novice translators in this post-editing experiment. To imitate the actual working scenario, we provided participants with a brief guideline of quality and technical requirements in subtitle translation, i.e., the Code of Good Subtitling Practice (Ivarsson and Carroll 1998). They were asked to read and understand the guideline carefully before they start.

3.3 Apparatus

This experiment used a portable Tobii X2–60 eye-tracker (60 Hz) and Translog-II software for the recording of both eye-tracking and keystroke-logging data. The eye-tracker was placed on the bottom of a 24″ monitor as shown in Fig. 1.

After watching the video and calibrating the eye-tracker, participants post-edited a transcript of the video text in Translog-II with the task setup as follows: as shown in Fig. 2, the English source texts (i.e., the subtitles) were displayed on the left

Fig. 1 Eye-tracker setup of the experiment

[1]For more information, see http://www.catticenter.com/wzjs/452 (CATTI) and https://www.ielts.org/ (IELTS).

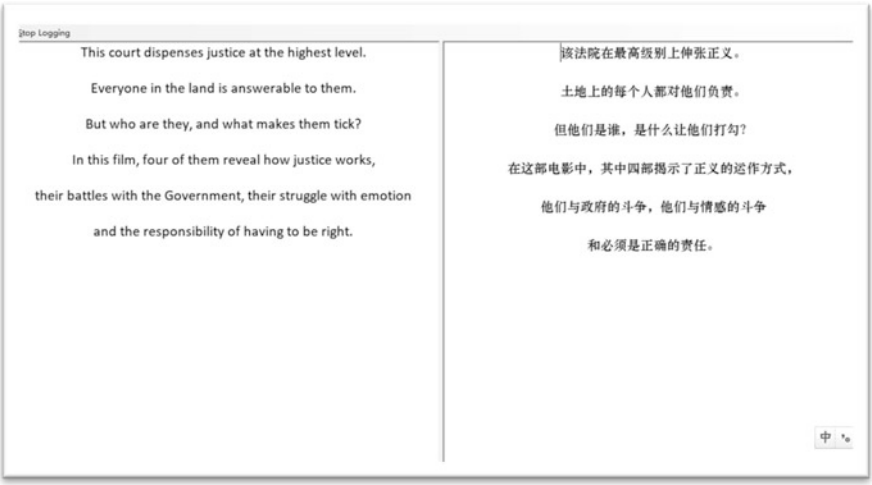


Fig. 2 Translog-II task setup for participants

window of Translog-II in Calibri, 20-point size, and 1.5 line spacing, while the Chinese target texts were positioned on the right side with similar settings except that the font was changed to STZhongsong. The line breaks of both English and Chinese texts were displayed in the same way as in the original videos, in which a short sentence occupies a line and a long sentence is broken into two lines. After collecting the XML files produced by Translog-II, we added the data to the CRITT TPR-DB for further alignment and analysis.

3.4 Data Alignment

The ST and TT were first sentence-segmented and tokenized[2] and aligned using the YAWAT tool (Germann 2008). To ensure consistency, a single researcher aligned each meaning unit in ST (a word or phrase) to its corresponding TT unit in all sessions by participants. Figure 3 shows a tokenized segment from the interface of the manual alignment tool with the ST on the left and the TT on the right. Successively, we aligned the ST and TT tokens on a level of minimal translation equivalence. In Fig. 3, the ST-TT alignment groups are as follows: *"Everyone"*—*"每个/人"*, *"in the land"*—*"国 土/上"*, *"is answerable to"*—*"回/应/. . . 的/诉求"*. The ST tokens *them* and punctuation "." are not translated and have no TT equivalents. As shown in this example, all of the ST and TT units represent the minimal meaning group in each language, including basic collocations. Phrase-level

[2]For Chinese, we used the Stanford Tokenizer (https://nlp.stanford.edu/software/tokenizer.shtml).

Fig. 3 Alignment of ST and TT units in Yawat

units avoid the arbitrary separation of meanings and allow us to have consistent alignment in both ST and TT.

3.5 Computation of WHER

The TER script computes an edit distance at the segment-level between a translation hypothesis (usually the MT output) and a translation reference (usually a human translation). It produces an *adjusted hypothesis*, which is tagged with edit operations on specific word positions. The adjusted hypothesis encodes edits and shifts, which represent the minimum edit distance between hypothesis and reference. The TER script also computes a cumulative TER score where each of the edits has a cost of 1, thus representing the distance between the hypothesis and the reference. Snover et al. also introduce the HTER score in which the reference is the post-edited MT hypothesis.[3] Given that TER is designed to be a measure of MT quality and there are many ways to translate a sentence, HTER (i.e., edit distance between MT output and its post-edited version) is believed to give more prominence to the "real" quality of the MT output, as compared to a perhaps very different human translation. However, in order to assess the amount of human editing (as compared to MT errors), do Carmo suggests reversing the reference and the hypothesis, so that the post-edited version becomes the hypothesis and the MT output the reference. do Carmo calls his reversed edit distance metric the HER which is presumably better suited to assess the post-editing effort. We extend the HTER score into a word-level HER (WHER), which (1) reverses the reference and the hypothesis in the HTER calculation following do Carmo's suggestion, (2) breaks down the segment-level HER operations to the word level, and (3) maps the operations on TT words to the aligned ST equivalents.

The edit distance between the hypothesis and the reference text consists of shifts of word positions (H), insertion (I), deletion (D), and substitutions (S). We take the hypothesis text to be the MT output and the reference text to be the final post-edited result. The minimum edit operations between the reference text and the adjusted MT are calculated to represent the minimum effort of post-editing from the product-wise perspective. A sample sentence containing all of the four types of WHER operations is presented in Fig. 4 illustrating the mappings between WHER operations and the exact TT tokens.

[3]But see do Carmo (this volume, Chap. 1) for a discussion on the terminological confusion of different definitions of HTER.

ST word.	Everybody		in the land				is answerable to				
WHER			SH				ISDDDSS				
Keystrokes	都对他们负责。		国家的的家土上				回应的诉求				
TT (Hyp.)	回	应		国土	上	每个	人				的	诉求
Adjusted Hyp.	I	S		S	H			D	D	D	S	S
MT (Ref.)		土地	上	的		每个	人	都	对	他们	负责	。

Fig. 4 Example for calculating the minimum edit operations in WHER. Identical background colors indicate translation equivalence. Font colors green (I), insertion; blue (D), deletion; red (S), substitution; yellow (H), shift

Figure 4 reproduces the alignment from Fig. 3. It shows the ST segment with the WHER operations, the Chinese post-edited translation, an adjusted hypothesis which consists of the string of edit operations that are projected to the ST words, and the raw MT output. In addition, we add a row of the real keystroke activities below WHER to make comparisons. The adjusted MT is an intermediate construct that illustrates the assumed operations which have presumably occurred in the mapping of the post-edited hypothesis on the MT reference. Figure 4 also shows a sequence of deletions, substitutions, and insertion operations and a shift between the hypothesis and the MT reference. In addition, it shows the three aligned chunks between the ST segment and the TT hypothesis (i.e., the post-edited version). Notice that there is a syntactic inversion between "*Everybody*"—"每个/人" and *"in the land"*—"国土/上" and a discontinuous translation, *"is answerable to"*—"回/应/. 的/诉求".

Let's illustrate the calculation of WHER by taking an example of the ST phrase "is answerable to" which is aligned with two discontinuous Chinese segments, "回/应/" and "的/诉求" consisting of a total of four TT tokens.

Step 1: The edit operations are mapped to Chinese TT words. The shift operation (H) only happens in the position change of the hypothesis, so it is found here below "上" of the post-edited TT. Substitutions are identified as tokens occupying the same positions but with different contents between TT and MT. In this case, four substitutions are detected for "土地", "的", "负责", and "。". Deletions and insertions happen in places where tokens are missing or added from the MT to the TT. The vertical display of token positions in Fig. 4 shows that "回" is an insertion and "都", "对", and "他们" are deletions.

Step 2: The edit operations associated with the TT tokens are then mapped onto the ST words through their alignment links. The WHER operations for "is answerable to" is, consequently, "ISDDDSS."

Step 3: The WHER score for each ST token in an alignment unit is calculated as the sum of WHER operations because each operation type costs 1 in the same way. Then, the WHER score for "is answerable to" is 7, as a sum of all edit operations.

Next, we can compare WHER operations to the actual modifications[4] that were produced during post-editing session. The row of actual modifications as produced in the post-editing process is shown below the three ST chunks. Overall, there are 19 modifications of Chinese characters in the segment (which required 38 keystrokes with the Chinese input tool). Seven modifications (deletions), which make up five Chinese words, are assigned to "Everybody." These operations represent the deletion of the erroneous MT output "都/对/他们/负责/。/", at the end of the MT output. WHER assigns the modifications of these five words as three deletions and two substitutions to *is answerable to* (DDDSS). There is thus a discrepancy in the allocation of edit operations between WHER and the TPR-DB analysis. There are also seven modifications assigned to *in the land*. The translation of *in the land* actually consists of two tokens and three Chinese characters "国土/上/". However, the post-editor seems to have revised his/her own editing activities and inserted and deleted the two characters "家的". Finally, there are five insertions "回应的诉求" for *is answerable to* which corresponds to the discontinuous chunk "回应" and "的诉求". Notice that the first part (回应) consists of one insertion and one substitution in terms of WHER operations, but were produced as insertions in the post-editing process.

Comparing WHER to HER, do Carmo says "HER is to move the focus from errors to edits" and "an edit rate presents an improved perspective on actual editing" (see do Carmo this volume, Chap. 1). We pursue the same aim with the WHER score but on a word level. With WHER, we are looking at the human editing effort instead of MT errors on the word level. However, the only thing that changes from errors to edits is an inversion of labels for insertions and deletions: what is a deletion error in the MT output in TER becomes an insertion operation in the editing pattern for HER, and what is an insertion error in the MT output for TER becomes a deletion operation in HER. Shifts (H) and substitutions (S) are independent of the error/edit view as they are symmetrical. Figure 4 shows this reversed relation in the errors vs. edits view: the "回" in the post-edited translation of *is answerable to* was an omission error in the MT output and appears as insertion (I) in the string of edit operations. Similarly, "都对他们" was an erroneous MT insertion but appears as deletion (D) in the edit string. Shifts (H) and substitutions (S) also change directionality depending on whether we look from MT hypothesis to TT reference or the other way around.

[4]Note that there is a difference in produced keystrokes and modifications for Chinese: with an input method editor (IME), there are usually more (i.e., 2–3) keystrokes required to produce one Chinese character. In this analysis, we count the number of character modifications in the text, as opposed to keystrokes.

3.6 Features

We correlate effort indicators from both the process and product data with WHER score. As mentioned above, keystroke, fixation, and duration information are indicators of post-editing effort during the process. The details of the process and product features are displayed in the following two tables.

As shown in Table 1, ten features from the three groups of process data are used in our experiment. For keystroke, Key_ins and Key_del measure the insertion and deletion activities recorded on the TT window by Translog-II, and Key_all is the total number of keystrokes that reflect the overall typing effort. Fixation features including Fix_S, Fix_T, Trt_S, and Trt_T cover both temporal and count data with a separation on ST and TT windows. The total values Fix_all and Trt_all are also included to provide a holistic view of the reading effort. Measuring the time spent from the first keystroke to the last keystroke, Dur is a temporal record of the technical effort.

Table 2 shows five features used for measuring lexical and syntactic variations of post-edited texts. AltT counts the number of alternatives for each ST token; therefore, it is not a participant-specific value. ProbT is the probability of the current translation choice relating to the ST token, which is sensitive to participants' individual differences. Based on the two values, HTra calculates the information entropy of each ST word across the whole data set and indicates the variations of translations for each ST token regardless of the participants' differences (Carl and

Table 1 Process features

Category	Feature name	Description
Keystroke	Key_ins	Number of keystroke insertions
	Key_del	Number of keystroke deletions
	Key_all	Total number of keystroke insertions and deletions
Fixation	Fix_S	Number of fixations on ST
	Fix_T	Number of fixations on TT
	Fix_all	Total number of fixations on both ST and TT
	Trt_S	Fixation duration on ST
	Trt_T	Fixation duration on TT
	Trt_all	Total fixation duration on both ST and TT
Duration	Dur	Production duration from the first keystroke to the last keystroke

Table 2 Product features

Category	Feature name	Description
Lexical variation	AltT	Number of alternatives for ST tokens
	ITra	Self-information of current translation choice
	HTra	Word translation entropy
Syntactic variation	Cross	Cross value for ST tokens
	HCross	Word distortion entropy

Schaeffer 2014, Schaeffer et al. 2016). Similarly, Cross and HCross are included to indicate the variation related to syntactic adjustments.

4 Results

This section discusses the results of the correlation tests. The collected data of the above features are not normally distributed except HTra, ITra, and HCross. The distributions of WHER, all of the process features, and the remaining product features are skewed to the right. As most of these data have 0 s, we decide to transform the above right-skewed data by adding a constant 1 to all values and taking their log transformation (Hancock et al. 2018). In this way, the features are more comparable, and we can use Pearson's r as the correlation metric in all tests. Among all of the correlation results below, the asterisks indicate the significance levels of the correlations, where one asterisk (*) refers to a significant effect (p-value <0.05), two asterisks (**) for a highly significant effect (p-value <0.01), and three asterisks (***) for a very highly significant effect (p-value <0.001).

4.1 Process Features

We collect process data from keystroke, duration, and fixation as indicators of technical, temporal, and cognitive effort during post-editing. With log transformations on both WHER and the other features, Table 3 shows the correlation results between them.

Table 3 shows that LogWHER strongly correlates with the insertion activities (LogKey_ins) and the number of overall keystrokes (LogKey_all). The correlation of LogKey_ins ($r = 0.76$) is higher than that of LogKey_all ($r = 0.68$), indicating a major contribution from the insertion activities. The number of deletions (LogKey_del) is only weakly but significantly correlated with LogWHER

Table 3 Correlations between WHER score and the process features (log-transformed)

Category	Feature name	Pearson's r with LogWHER
Keystroke	LogKey_ins	0.76***
	LogKey_del	0.13***
	LogKey_all	0.68***
Fixation	LogFix_S	0.01
	LogFix_T	0.15***
	LogFix_all	0.13***
	LogTrt_S	0.01
	LogTrt_T	0.12***
	LogTrt_all	0.11***
Duration	LogDur	0.35***

($r = 0.13$). The strong correlation suggests that WHER score is a good indicator of performed keystroke operations.

However, the correlation with fixation data is relatively weak. Fixations are counted separately for the ST and TT texts; we find higher scores in TT than ST where LogFix_T ($r = 0.15$) and LogTrt_T ($r = 0.12$) have significant correlations with LogWHER. The results corroborate the findings of previous studies that TT is likely to attract more attention than ST (Daems et al. 2017, Schaeffer et al. 2019). Overall, the correlations between fixation data and WHER are not as strong as those in keystroke data. One possible explanation might be the noise of fixation data and the gaze-to-word mapping errors. WHER is thus not predictive of the reading effort as indicated by the number and duration of fixations in this study.

We used the production duration of an ST token (Dur) as an indicator of temporal effort. The correlation result between LogDur and LogWHER is moderate ($r = 0.33$), which is higher than the scores for fixations and lower than those for keystrokes. It is also indicated that the WHER feature reflects, to some extent, the temporal effort during post-editing.

Overall, the keystroke and duration features are found to be at least moderately correlated with the WHER score, but the fixation data are only weakly correlated. This suggests that WHER is more indicative of the technical and temporal effort than for cognitive effort, as gathered from reading activities.

4.2 Product Features

As mentioned above, AltT and Cross are distributed with skews to the right, so we added a constant 1 to all values and used their log-transformed data to correlate with LogWHER. In particular, Cross has both positive and negative numbers because it is a vector between the relative word positions in ST and TT. Its absolute value is thus indicative of the number of word distortions between ST and TT. Therefore, we take the absolute value of Cross before the log transformation to enable the correlation tests.

Table 4 shows relatively strong correlations between WHER and the three features of lexical variation. Although with slight differences, HTra, LogAltT, and ITra show a consistent moderate correlation with LogWHER, with the highest

Table 4 Correlations between WHER and the indicators of product features (log-transformed)

Category	Feature name	Pearson's *r* with LogWHER
Lexical variation	LogAltT	0.51***
	ITra	0.70***
	HTra	0.54***
Syntactic variation	LogCross	0.30***
	HCross	0.36***

absolute correlation score being 0.61 for ITra and the lowest value being 0.54 for HTra. This means that less common translations require more operations than more frequent ones. For post-editing, it implies that rare translation alternatives are likely to relate with more edit operations. In general, the correlation results of the three indicators suggest that the effort of editing MT output can be well reflected by WHER.

For the features of syntactic variation, LogWHER correlates moderately but significantly with both HCross (0.36) and LogCross (0.30). The two correlation values are both lower compared with the above indicators of lexical variation. As the results of lexical and syntactic variation reflect the editing effort, the above results imply that lexical modifications on the MT output are more prominently reflected in the WHER score than syntactic modification. In other words, the overall editing effort indicated by WHER operations only overlaps partly with the effort of making syntactic choices.

5 Discussion and Conclusion

From the results presented above, we can see that the WHER correlates with both process and product measures of post-editing effort. Moderate to strong correlations are found between WHER and several features in keystroke, duration, and lexical variation. The correlation results support our assumption that the number of WHER operations reflects the post-editing effort.

Keystroke activities have stronger correlations with WHER than gaze data. The number of insertions has the strongest correlation with WHER, while the number of deletions is not correlated at all. The production duration indicating typing and pause effort has a moderate correlation with WHER. However, the weak correlation of the number of fixations and total fixation durations shows that the reading effort might not be well represented by the WHER operations. Effort indicators of making lexical modifications in the MT output have stronger correlations with WHER than the indicators of adjusting sentence structures. While all three lexical indicators are moderately correlated with WHER, ITra has the highest correlation, while HCross and Cross only have weak correlations with WHER. The correlation results are in line with the previous findings that a single measure is not robust enough to estimate the post-editing effort (Koponen 2012, Guerberof 2014). Our new WHER metric correlates better with keystroke activities (Key_ins) and the self-information of translation (ITra).

To sum up, this chapter introduces WHER as a measure to quantify the minimum per word edit operations. The study aims at finding out whether WHER is a reliable indicator of post-editing effort. By experimenting with audiovisual texts, we recorded participants' post-editing activities with keystroke-logging and eye-tracking devices. We collected the MT output and the final post-edited results and aligned the ST and TT. We computed the WHER score and correlated it with the process and product data on the word level. Our main contribution is the

development of the WHER score which extends the HTER score by Snover et al. (2006) and the HER score, as suggested by do Carmo (this volume, Chap. 1).

We have shown that WHER correlates with both process and product measures which indicate multiple aspects of post-editing effort. Measuring the edit operations mapped to each ST word, WHER provides a new perspective of looking at post-editors' effort. We find that it is possible to use WHER to estimate post-editors' typing activities, lexical modifications, and, to a lesser extent, word order changes in the target texts. However, as our data show, reading activities indicated by fixations are not associated with WHER. For future research, we would like to include more translation process recordings, language pairs, and text types to corroborate the usability of WHER. If WHER is proven to be stable in correlating with multiple process and product indicators of post-editing effort, it might be used as an indicator to estimate word post-editing difficulty and/or generate indications which ST fragments might be troublesome in MT and post-editing. Besides, as the sentence-based HTER score has been used in the translation industry (see Cumbreño and Aranberri this volume, Chap. 3, for a sentence-level HTER assessment) to predict post-editing difficulty and decide human pay rates, the WHER metric can help specify words that are harder to translate and thus encourage language service providers to make flexible strategies on translation and pricing.

References

Armstrong S, Way A, Caffrey C et al (2006) Improving the quality of automated DVD subtitles via example-based machine translation. In: Proceedings of translating and the computer. Aslib, London. http://www.mt-archive.info/Aslib-2006-Armstrong.ppdf

Bangalore S, Behrens B, Carl M et al (2016) Syntactic variance and priming effects in translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research exploring the CRITT TPR-DB. Springer, Cham, pp 211–238

Basu P, Pal S, Naskar SK (2018) Keep it or not: word level quality estimation for post-editing. In: Proceedings of the third conference on machine translation (WMT). Association for Computational Linguistics, Brussels, pp 772–777

Burchardt A, Lommel A, Bywood L et al (2016) Machine translation quality in an audiovisual context. Target 28(2):206–221

Bywood L, Volk M, Fishel M, Georgakopoulou P (2013) Parallel subtitle corpora and their applications in machine translation and translatology. Perspectives 21(4):595–610

Campbell S (2017) Choice network analysis in translation research. In: Olohan M (ed) Intercultural faultlines: research models in translation studies I. Textual and cognitive aspects. Routledge, London, pp 29–42

Carl M (this volume) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Schaeffer M (2014) Word transition entropy as an indicator for expected machine translation quality. In: Miller KJ, Specia L, Harris K, Bailey S (eds) Proceedings of the workshop on automatic and manual metrics for operational translation evaluation. European Language Resources Association, Paris, pp 45–50

Carl M, Schaeffer MJ (2017) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. J Lang Commun Bus 56:43–57

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research exploring the CRITT TPR-DB. Springer, Cham, pp 13–54

Cumbreño C, Aranberri N (this volume) What do you say? Comparison of metrics for post-editing effort. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Daems J, Vandepitte S, Hartsuiker R, Macken L (2017) Translation methods and experience: a comparative analysis of human translation and post-editing with students and professional translators. Meta 2:245–270

de Sousa SCM, Aziz W, Specia L (2011) Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In: Proceedings of recent advances in natural language processing. Association for Computational Linguistics, Hissar, pp 97–103

DePalma DA, Pielmeier H, O'Mara P (2019) Who's who in language services and technology: 2019 rankings. Common Sense Advisory, Boston, MA. https://insights.csa-research.com/reportaction/305013039/Marketing

Díaz-Cintas J, Remael A (2014) Audiovisual translation subtitling. Routledge, New York, NY

do Carmo F (this volume) Editing actions: a missing link between translation process research and machine translation research. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Dragsted B (2012) Indicators of difficulty in translation—correlating product and process data. Across Lang Cult 13(1):81–98

Germann U (2008) Yawat: yet another word alignment tool. In: Proceedings of the 46th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Columbus, pp 20–23

Guerberof A (2014) The role of professional experience in post-editing from a quality and productivity perspective. In: O'Brien S, Winther Balling L, Carl M et al (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars, Newcastle Upon Tyne, pp 51–76

Hancock GR, Stapleton LM, Mueller RO (2018) The Reviewer's guide to quantitative methods in the social sciences, 2nd edn. Routledge, London

Ivarsson J, Carroll M (1998) Code of good subtitling practice. European Association for Studies in screen translation, Berlin. https://www.esist.org/wp-content/uploads/2016/06/code-of-good-subtitling-practice.pdf.pdf

Jia Y, Carl M, Wang X (2019) How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. J Spec Transl 31:60–86

Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87(4):329–354

Koglin A (2015) An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. Transl Interpret 7(1):126–141

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the seventh workshop on statistical machine translation. Association for Computational Linguistics, Montreal, QC, pp 181–190

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes, Kent State University Press, Kent

Lacruz I, Ogawa H, Yoshida R, Yamada M, Martinez DR (this volume) Using a product metric to identify differential cognitive effort in translation from Japanese to English and Spanish. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Leijten M, Van Waes L (2013) Keystroke logging in writing research: using inputlog to analyze and visualize writing processes. Writ Commun 30(3):358–392

Martins AFT, Junczys-Dowmunt M, Kepler FN et al (2017) Pushing the limits of translation quality estimation. Trans Assoc Comput Ling 5:205–218

Melero M, Oliver A, Badia T (2006) Automatic multilingual subtitling in the eTITLE project. In: Proceedings of translating and the computer. Aslib, London. http://www.mt-archive.info/Aslib-2006-Melero.pdf

Moorkens J, O'Brien S, da Silva IAL et al (2015) Correlations of perceived post-editing effort with measurements of actual effort. Mach Transl 29:267–284

Mossop B (2007) Empirical studies of revision: what we know and need to know. J Spec Transl 8:5–20

Nitzke J (2019) Problem solving activities in post-editing and translation from scratch: a multi-method study. Language Science Press, Berlin

Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and ST-based information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Orero P, Doherty S, Kruger JL et al (2018) Conducting experimental research in audiovisual translation (AVT): a position paper. J Spec Transl 30:105–126

Ortiz-boix C, Matamala A (2016) Post-editing wildlife documentary films: a new possible scenario? J Spec Transl 26:187–210

Sanchis-Trilles G, Alabau V, Buck C et al (2014) Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. Mach Transl 28:217–235

Schaeffer M, Carl M (2014) Measuring the cognitive effort of literal translation processes. In: Proceedings of the workshop on humans and computer-assisted translation (HaCaT). Association for Computational Linguistics, Gothenburg, pp 29–37

Schaeffer M, Dragsted B, Hvelplund KT et al (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 183–210

Schaeffer M, Nitzke J, Tardel A et al (2019) Eye-tracking revision processes of translation students and professional translators. Perspectives 27(4):589–603

Snover M, Dorr B, Schwartz R et al (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the Association for Machine Translation in the Americas. Association for Machine Translation in the Americas, Cambridge, pp 223–231

Vanroy B, De Clercq O, Macken L (2019) Correlating process and product data to get an insight into translation difficulty. Perspectives 27(6):924–941

Vanroy B, De Clercq O, Tezcan A, Daems J, Macken L (this volume) Metrics of syntactic equivalence to assess translation difficulty. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Mach Transl 28:187–216

Vieira LN (2016a) How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. Mach Transl 30:41–62

Vieira LN (2016b) Cognitive effort in post-editing of machine translation: evidence from eye movements, subjective ratings, and think-aloud protocols. Dissertation. Newcastle University, Newcastle Upon Tyne

Volk M (2008) The automatic translation of film subtitles: a machine translation success story? In: Nivre J, Dahllöf M, Megyesi B (eds) Resourceful language technology: Festschrift in honor of Anna Sågvall Hein. Uppsala University, Uppsala

Wei Y (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Xenouleas S, Malakasiotis P, Apidianaki M, Androutsopoulos I (2019) SUM-QE: a BERT-based summary quality estimation model. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 6005–6011

What Do You Say? Comparison of Metrics for Post-editing Effort

Cristina Cumbreño and Nora Aranberri

Abstract The improvement in machine translation quality is creating a constantly increasing number of post-editing jobs. As a result, research geared toward ensuring an efficient translation process for post-editors has become more important than ever. To this end, being able to measure and predict the effort involved during the post-editing activity is essential. This work aims to assess whether simple post-editing effort metrics associated with the three effort dimensions (temporal, cognitive, and technical) correlate among themselves. Also, it seeks to examine whether these simple metrics are able to capture the variation in effort involved in addressing different error types. To address these objectives, we asked professional translators to post-edit a test suite of sentences that include one pre-selected error each and used a set of simple metrics to measure the post-editing effort. Results seem to indicate that the correlation between the metrics is rather low, which suggests that the use of a single metric to measure the effort might produce biased measurements. We also observe that, overall, metrics report very similar effort values for the different error types but some distinctions are noticeable, which allow us to rank error types per difficulty.

Keywords Post-editing effort · Metrics · Correlations · Error types

C. Cumbreño
University of the Basque Country UPV/EHU, Leioa, Spain
e-mail: ccumbreno001@ikasle.ehu.eus

N. Aranberri (✉)
HiTZ Basque Center for Language Technologies – Ixa NLP Group, University of the Basque Country UPV/EHU, Leioa, Spain
e-mail: nora.aranberri@ehu.eus

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_3

1 Introduction

With the increase in machine translation (MT) quality, post-editing is becoming commonplace within the translation industry. The wide array of studies showing that MT systems increase translators' productivity (Guerberof 2009; Plitt and Masselot 2010; Parra Escartín and Arcedillo 2015), among others, has encouraged many translation companies to integrate them into their workflow. While the situation varies greatly in different countries with respect to language pairs, content types, etc., as a general trend, professional translators are progressively getting fewer translation jobs and more post-editing offers (Gaspari et al. 2015).

(Machine) translation research is following suit to try to understand what this post-editing activity involves. Studies in this area cover many aspects. Some of them are concerned with editing analyses that look for insights into the actual changes performed by post-editors (De Almeida 2013; Aranberri 2017; Koponen et al. 2019). Based on this experience, other efforts concentrate on establishing guidelines and on observing to what extent these are complied with (Flanagan and Christensen 2014; Hu and Cadwell 2016; Massardo et al. 2016). Finally, other works seek to ensure an efficient workflow for post-editors, measuring post-editing effort (O'Brien et al. 2014; Daems et al. 2017; Moorkens 2018) and developing QE tools (Specia and Farzindar 2010; Aranberri and Pascual 2018).

As mentioned above, one of the main requirements to guarantee an efficient workflow for post-editors is the accurate measurement of the post-editing effort. This is commonly measured according to the proposal of Krings (2001), who claimed that post-editing involved three types of effort, namely, temporal, cognitive, and technical. To that end, researchers (and industry players) have come up with diverse metrics to measure post-editing effort that tends to be directed at one of the dimensions. For example, working times are associated with the temporal dimension, perceptions of effort are related to the cognitive dimension, and keystrokes are connected to the technical dimension (see also Vanroy et al., this volume, Chapter, on metrics to assess translation difficulty).

From that context, given that each metric mainly focuses on one dimension and that each dimension addresses a different aspect of the effort, it is unclear that single metrics report the same post-editing effort (de Gibert and Aranberri 2019). For that reason, we could expect that it will only be possible to know the full effort by considering metrics from all three dimensions. To put it differently, we believe that only considering a single metric will not provide us with enough information about the full effort involved in post-editing, at least with the widespread metrics used nowadays (see Related Work). As a result, the decisions we make based on this information might be biased. That is precisely the objective of this work, the evaluation of how different metrics for the three dimensions compare, and whether the measurements capture variations in post-editing effort for different error types. This small study has been carried out using post-edits provided by professional translators and 11 metrics that measure the post-editing effort for different dimensions. Results display different effort outcomes, suggesting that

using a single metric or focusing on a single dimension might not capture the real post-editing effort.

2 Related Work

Let us consider the different metrics used by academic researchers and industry players to measure post-editing (PE) effort. It can be argued that the temporal aspect is the most straightforward to compute. Total post-editing times, which may be computed at text or sentence level, are common in academic work (Tatsumi and Roturier 2010; Specia 2011). Research has also explored new ways of using time measurements to predict cognitive effort by capturing peaks in difficulty at different moments of the post-editing process. For example, while some have focused on pause times (O'Brien 2006b), others have used gaze duration (Dragsted and Hansen 2009; Carl et al. 2011) as metrics to capture the effort. In contrast, the experiments performed in industrial settings concentrate almost exclusively on capturing total post-editing time with the goal of measuring the improvement gain between translation and post-editing (Plitt and Masselot 2010; Parra Escartín and Arcedillo 2015).

Regarding the technical dimension, the post-editing effort is captured using different metrics. One of the most widespread is the count of keystrokes, where a stroke corresponds to one physical action taken by the editor (De Almeida and O'Brien 2010; Koponen 2016; Nitzke and Oster 2016). In order to easily capture this information several tools have been developed, for example, PET (Aziz et al. 2012), which resembles the working environment of a translation memory (TM) tool or Translog (Carl 2012), which records the reading and writing activity of an editor on a text file.

Another popular metric for this dimension is the edit distance between the MT output and the post-edited version (Temnikova 2010; Koponen 2012; Wisniewski et al. 2013). This is mainly calculated using the automatic metric Translation Edit Rate (TER) (Snover et al. 2006). Called HTER when post-edited versions are used as reference instead of from-scratch translations, this metric computes the minimum number of editing operations (i.e., insertions, deletions, word substitutions, and phrase shifts) to be performed on a given MT output so that it becomes an exact match of its reference, normalised by the number of words in it.[1] The advantage of this metric is that it is non-invasive for the post-editing work and very easy and fast to implement. In practice, it is enough to run a simple language-independent program to compare the MT output and the post-editing version, after the translator has completed the task, to obtain a result.

[1] See Do Carmo (this volume, Chap. 1) for a discussion on the terminological confusion of different definitions of HTER.

Nevertheless, HTER has various shortcomings. For example, it does not take into account the total edits performed during the post-editing process, but rather restricts its count to those changes that are visible at the end of the process. It can be argued that the effort captured by this metric is limited, as it only pays attention to the product, but not to the technical effort carried out during the process. Also, it gives every edit the same score, even if some edits may be more challenging than others, as research seems to indicate (Temnikova 2010).

To mention another weakness, we should turn to the way in which the edits are computed. The calculation of HTER is language-agnostic, that is, it calculates the shortest sequence of edits without considering the linguistically motivated changes a human translator would perform. do Carmo (this volume, Chap. 1) discusses editing actions and their interpretation. In an attempt to address this, in Blain et al. (2011), the authors proposed a protocol to automatically extract minimal and logical edits, called "Post-Editing Actions," which are based on linguistically oriented error classifications. However, this approach requires linguistic annotation of the texts to be compared, which makes it less straightforward to use. Despite evidence against the appropriateness of the metric, interestingly, Huang and Carl (this volume, Chap. 2) report a moderate but significant correlation between HTER operations with production duration when studying the operations assigned to the words in the translated text mapped via alignment links to the source text words.

When considering research that emerges from industrial settings, we observe that HTER is the metric that is more widely used to account for the technical effort, or even post-editing effort in general (Tatsumi 2009; Specia 2011; Parra Escartín and Arcedillo 2015). This is not surprising, given the technical difficulty of using keystroke loggers in real settings. This type of feature is usually not integrated in major TM tools, and external software must be used alongside them. HTER is also being widely used for QE, which aims at automatically predicting the quality of MT output (Specia and Farzindar 2010).

Cognitive effort is the most difficult aspect to quantify, as it delves into subconscious processes and mental strain (see Wei, this volume, Chap. 7, for a discussion of an example in detail). In academic research, attempts have been made to measure it through a number of complex techniques. Among others, in Krings (2001), the author used think-aloud protocols, which involve having translators to explain their edits as they perform them. While very informative, this strategy affects the natural flow of the translation process, fails to tap into the subconscious processes, and does not offer comparable results. A completely different approach was followed by O'Brien (2006b), where the choice network analysis technique was used to explore the different ways a segment can be edited, with the assumption that the more options there are, the more effort it takes to choose among them. This is also the underlying assumption in work such as Carl and Schaeffer (2017) and in various chapters in this volume (e.g., Lacruz et al., Chap. 11; Ogawa et al, Chap. 6; Wei, Chap. 7). What must be carefully considered when using this technique is that the options available to each post-editor might differ.

Another approach to measuring cognitive effort involves analyzing the presence of pauses during post-editing based on the assumption that the more a translator

pauses before an edit, the more cognitively challenging the edit is, or, indeed, that the translator is assessing the previous (or more distant) edit, among other possibilities. As a result, researchers have studied pause-typing ratios, as well as the duration, frequency, and distribution of pauses in the sentence. For example, this has allowed researchers to link the presence of clusters of short pauses with cognitively challenging edits (Lacruz et al. 2012; Lacruz and Shreve 2014). Similarly, Probst (2017) found differences in the pause length prior to post-editing certain error types, and works such as O'Brien (2005, 2006a) examined pauses in segments containing specific source text features believed to increase cognitive effort and segments without them but found no significant differences.

Finally, we should mention eye-trackers, which follow the editor's gaze assuming that the segments where the gaze stays longer are more cognitively demanding. Eye-trackers have gained momentum in recent years, gathering a growing source of reliable cognitive effort measurements. The average fixation time and counts have been used to explore various aspects of post-editing. For example, O'Brien (2011) and Moorkens (2018) used this metric to determine the quality of MT output, while others have assessed translators' reactions to new TM tools (Mesa-Lao 2013) and translation expertise (Martínez-Gómez et al. 2018). Eye-trackers have also been applied to measure productivity; other works, such as Carl et al. (2011) and da Silva et al. (2017), have reported a significant increase in cognitive effort in translation from scratch as opposed to post-editing. In turn, in Alves et al. (2016), eye-trackers are used to compare interactive MT (i.e., where the tool displays suggestions as the translator writes) with non-interactive MT and find that the former requires a lower cognitive effort. Finally, eye-trackers have been employed to determine when and how different types of errors are recognized by post-editors (Schaeffer et al. 2019) and their impact on cognitive effort (Daems et al. 2017).

Another way of investigating cognitive effort in research is to simply ask the participant to assess the difficulty of the task, either before or after performing it (Koponen 2012; Lacruz et al. 2012); this is often referred to as manual evaluation or perceived effort. The advantage of this method is a lower cost compared to the other metrics. However, inter-annotator agreement tends to be very low and as is the correlation with objective measures.

In spite of their reliability, the metrics addressing cognitive effort remain largely confined to academic research, and they tend to be overlooked in industrial settings, with the exception of perceived effort. This may be partly due to the difficulty and expertise required to apply them, and even to the cost of acquiring and using specialized software on a large scale.

3 Experimental Setup

This work aims to explore two questions, namely, how metrics belonging to each post-editing effort dimension compare, and whether such metrics capture differences in effort for different error types. In this section, we present the

experimental setup, which focuses on the English–Spanish language pair. Firstly, the characteristics of the data set and the error classification used are presented; secondly, the metrics used for the effort measurement are outlined; and finally, the profile of the participants is described.

3.1 Data Set and Error Categories

To address the two objectives of this work, it was necessary to collect English source sentences and their Spanish machine translations containing specific errors. A quick review of commonly used data sets revealed that they include source sentences that vary greatly in length and that their MT output displays a wide range of error types and frequencies. Consequently, we opted for compiling a test suite that included sentences with the specific characteristics the experiment required. The use of test suites has been proposed in past studies, such as Arnold et al. (1993), Burchardt et al. (2016) or Guillou and Hardmeier (2016), as the best way to analyze specific linguistic features. For example, recently, Schaeffer et al. (2019) used a test suite to analyze errors in human translation proofreading, which allowed them to limit the total and local error frequencies.

As shown in Moorkens et al. (2015) and Probst (2017), among others, controlling the number of errors to be studied is key to draw solid conclusions. Therefore, we took steps to ensure that all the error types appeared enough times to make the results comparable. We also decided to restrict the number of errors present in each segment to one so that we could analyze the effort required by each of them without the interference of other errors in the near context.

It was our aim to gather existing sentences, rather than creating them artificially, and consequently, to focus on naturally occurring MT errors. Within this setup, we set additional constraints that we expected would homogenize the sentences, reducing at least to a certain extent, the difference in effort necessary to address the error-free words in the sentences.

To this end, we introduced restrictions on topic, formality, and style by extracting all the sentences from the same source, time period, and topic. Specifically, a corpus was created with the news articles about the Venezuelan crisis spanning from January 23rd to February 15th 2019 from the online international edition of the newspaper *The Guardian*.

Another feature we considered was the sentence length, as it can negatively affect the human perception of the MT output and the post-editing activity (Tatsumi 2009; Koponen 2012; Popovic et al. 2014). We set the length of the sentences to be included in our test suite from 20 to 25 words. Establishing maximum and minimum sentence length limits should help to reduce the impact of this feature.

Once the sentences that did not meet the above-mentioned criteria were discarded from the original corpus, we machine translated it using Google Translate[2] and carried out an error analysis. For this purpose, we took the classification proposed by Temnikova (2010) as a starting point, which includes ten classes ranked by cognitive effort. Given the good quality of the MT system and the low focus on stylistic issues of the error classification, the use of a single annotator was deemed sufficient for this work. From this analysis, 600 errors were identified. Because we wanted to focus on sentences that included a single error, those with multiple errors or error-free sentences were discarded from the corpus. Based on the frequency and distribution of errors encountered in the remaining sentences, we decided to restrict and modify the error classes of the original classification.

Out of the ten original classes, we discarded *wrong punctuation* and *missing punctuation*, as well as *incorrect style synonym* due to their low occurrence rates. We also discarded *word order at word level* and *word order at phrase level* because this type of error generally co-occurred with others. The classes *incorrect word* (mistranslation 1), *extra word* (extra word), and *missing word* (missing word) were used as originally defined. The class *correct word, incorrect form* was divided into two classes, namely *agreement of number/gender* (N/G agreement) and *agreement of time/aspect* (T/A agreement). Finally, the definition of the *idiomatic expression* class was restricted to refer to *mistranslations of 2 or more words* (mistranslation 2+). Table 1 displays the final six error classes we used, together with their frequency of occurrence in the analyzed MT output.

Finally, we randomly selected 10 sentences for each error type, amounting to 60 in total, which were to be included in the test suite. The sentences were arranged so that the sequence would have some degree of cohesion, and so that it would always be clear what person or situation they were referring to.

Table 1 Description of the final error categories used together with occurrence proportions

Error type	Description	Errors	Prop.
N/G agreement	Wrong number or gender of one or more words	20	0.03
T/A agreement	Wrong tense or mode (aspect) of one or more verbs	53	0.08
Mistranslation 1	Mistranslation of one word	89	0.14
Extra word	Extra word (not present in source sentence)	51	0.08
Missing word	Word present in source sentence but missing in machine-translated output	32	0.05
Mistranslation 2+	Mistranslation of two or more words (multi-word expressions)	67	0.11
Others	No errors / other errors / more than one error	288	0.48
Total		600	1

[2]https://translate.google.com/.

3.2 Metrics

Post-editing effort, as previously stated, is claimed to have three main dimensions, specifically, temporal, technical, and cognitive. In order to measure it, several metrics of varying complexity and cost have been proposed and used over the years for each of the dimensions. This experiment focuses on metrics commonly used in industry because they are relatively easy to implement during a post-editing job. Next, we describe the specific metrics used, classified by dimension.

Temporal effort

- **Total time.** It is the time spent working on a sentence, computed as the time elapsed since translators start working on a segment, until they finish.
- **Editing time.** It is the total time minus the pause time (see definition below). It is considered as the time spent typing and editing.

Cognitive effort

- **Pause time.** it is any lapse of time between keystrokes that surpasses a certain threshold. That threshold, which is included in the count, was established at 0.3 seconds, following Lacruz et al. (2012), who determined it as the shortest time elapsed for a pause to be considered as such.
- **Editing pause time.** It is the length of the pauses that take place between the first and last edits.
- **Initial pause.** It is the length of the pause before the first edit. This is assumed to be time spent reading and finding the error.
- **Final pause.** It is the length of the pause after the final edit. This is assumed to be revision time. If no editing has been carried out during an annotation, the total time is considered as revision time.
- **Pause count.** It is the number of pauses per segment.
- **Editing pause count.** It is the number of pauses that take place between the first and last edits.
- **Perceived effort.** It is a rating provided by participants about the difficulty of the segment on a 1 to 3 scale, with 1 being easy and 3 difficult, after finishing it.

Technical effort

- **Keystrokes.** It is the number of keyboard keys pressed. These include digit, symbol, and letter keys; copy, cut, and paste keys; navigation keys; any action keys (enter, delete, shift, etc.); and the space bar.
- **HTER.** It is computed as the edit distance between the machine-translated output and the final human post-edited version (Snover et al. 2006).

While some TM tools have integrated plug-ins that allow measuring post-editing effort, we decided to look for an open-source tool that would include several metrics. We finally opted for PET (Aziz et al. 2012), a graphical user interface for translation and post-editing, which allows gathering effort indicators and is highly customizable to the researchers' needs. The front end of PET resembles the working environment

Fig. 1 Screenshot from the test PET task showing an open segment

of a TM tool, displaying the source segments on one column and the segments to post-edit on the other. This is an advantage because participants are presented with a familiar setup for the task.

We availed of the customization possibilities of PET in several ways. Firstly, in order to ensure that the post-editing effort was allocated to the relevant sentence, we configured the task so that only one sentence could be active and visible at a time (see Fig. 1). Needless to say that the participants could return to previous sentences if needed (note that we count the number of revisions and add the extra time to the corresponding segment). Secondly, the task was customized so that, after working on each segment, participants were asked about the perceived difficulty. This question was presented in a pop-up window, halting the recording of all the other metrics. This enabled us to clearly distinguish between post-editing activity and assessment time.

PET generates an XML file where the performance of the participants is recorded. From that file, it is possible to extract post-editing time, keystrokes, perceived effort, and HTER. From that information, it is also possible to compute the remaining metrics we set to study.

3.3 Participants

The participants were recruited through the professional translator job posting website *ProZ*.[3] Out of all the applications, 7 participants were selected whose main

[3] https://www.proz.com/.

Table 2 Survey statements together with average score and standard deviation, where 1 equals strongly disagree and 5 strongly agree

Statement	Response score
It takes me less time to post-edit a text than to translate it from scratch	3.36 (0.95)
I enjoy post-editing	3.48 (1.10)
I like translating more than post-editing	4.10 (0.68)
I accept all post-editing jobs proposed to me	3.36 (0.81)
Post-editing jobs tend to be frustrating	3.48 (1.06)
I cannot assess the difficulty of a post-editing job before accepting it	4.10 (1.24)
The quality of machine-translated text tends not to be good enough so that the job is profitable for me	3.91 (1.01)
The remuneration for post-editing jobs is..	2.56 (0.48)

working languages were English–Spanish and had at least one year of experience in translation and at least 3 months' experience in post-editing.

Prior to being selected for the experiment, participants were asked to fill in a short survey to ensure that they did not have extreme opinions about post-editing. This way we expected to avoid participants unintentionally introducing bias into the experiment, and to provide us with insights into their general attitude toward MT and post-editing. The survey consisted of a series of statements that they rated from fully disagreeing (1) to fully agreeing (5), as well as a question concerning the fairness of post-editing remuneration, which they could assess from very unfair (1) to very lucrative (5).

Table 2 shows the specific statements and the average score assigned to each of them. In general, translators reported enjoying translating more than post-editing, even when they admitted that post-editing is less time-consuming. Regarding post-editing remuneration, opinions were divided between 3 (fair) and 2 (unfair); none of the translators regarded post-editing as either very lucrative or underpaid. It is also worth noting that translators considered the quality of MT output not to be good enough for post-editing, yet they reported not always being able to check it before accepting a job offer. All this information seems consistent with previous research into translators' opinions (Guerberof 2013).

In the additional comments section in the survey, several translators pointed out that post-editing jobs were diverse: where some could be enjoyable and profitable, others would be very frustrating, depending on the quality of the MT output. They agreed that MT could help but it is not useful in every situation. Often, they commented that when accepting post-editing jobs they would end up translating segments from scratch, but for a reduced fee. Another participant added that this situation was dangerous because some translators would try to skim through the text as fast as possible and, as a result, let mistakes and false friends slide. The same translator concluded that a representative sample of the text should always be

provided for post-editing jobs in order to properly assess the quality of the machine-translated output, but that this is not yet common practice in the industry.

The task presented to participants was the post-editing of our set of 60 segments, each containing a pre-established error type. Participants were paid for performing the task with a view to increasing reliability. They completed the job off-site, and to ensure a smooth experience during the task, they received a number of directions. First, they were provided with guidelines that described, among others, the nature of the task, the number of segments to be addressed, that there was only one error per segment, and the type of editing to be performed together with examples. Even if we are aware that this approach distances from the conditions of a real setup, given that the objective of this work was to measure the post-editing effort of different types of error in isolation, it was deemed adequate to warn translators that their task involved identifying such error and editing it. Moreover, a separate document included a step-by-step guide to installing and using PET, and a test task to familiarize themselves with the process. Finally, a document with a timeline of the Venezuelan crisis, the topic of the test suite, was also included so that the participants could focus on the task without disrupting it to check information about the content. The participants reported a good experience and confirmed that no issues were encountered.

4 Results

This section presents the results of the task performed by the professional translators. We first report the inconsistencies found in the work of participants. Next, we compare the results obtained using the different metrics. Finally, we analyze the results of the metrics per error type to observe whether the metrics capture differences in post-editing effort.

4.1 Inconsistencies in Editing Work

When analyzing the work performed by each participant, we noticed excessive or insufficient editing in a high number of segments. As it is summarized in Table 3, a closer inspection revealed that in 61 cases, the participants had edited not only the one error but also more elements (*additional elements*), introducing unnecessary stylistic changes, mainly word order or punctuation changes. For example, one participant had changed all the formatting of quotation marks from *“”* to *«»*. Additionally, in 34 segments, we detected that no editing was performed (*no elements*), and in another 62 segments, participants had corrected something different from the intended error (*wrong element*).

A degree of subjectivity in the identification of errors cannot be totally ruled out from both the researchers’ and the post-editors’ perspectives even if a well-defined classification was used during error selection and concise guidelines provided for

Table 3 Distribution of the sentences according to the elements addressed in each by each participant

Participants	Additional elements	No elements	Wrong element	Correct element
1	2	1	12	45
2	2	3	10	45
3	29	4	3	24
4	4	5	11	40
5	16	6	10	28
6	8	4	6	42
7	0	11	8	41
Total	61	34	62	265

post-editing. Given those inconsistencies, and in order to avoid introducing noise to our data, we decided to discard the segments where post-editors performed unexpected changes. In total, 157 segments had to be discarded, leaving 265 sentences for the analysis: 44 *N/G agreement* errors, 43 *T/A agreement* errors, 34 *extra word* errors, 40 *missing word* errors, 52 *mistranslation 1* errors, and 52 *mistranslation 2+* errors. We collected the values of the different metrics for these sentences.

4.2 Comparison of Metric Results

The first objective of this work is to compare the results provided by each metric with a view to examining if they all reported a similar post-editing effort for the evaluated sentences. Since different metrics are aimed at measuring different aspects of effort, we expect to see metrics that correlate well within their effort dimension and have worse correlations with metrics associated with other dimensions. Indeed, Huang and Carl (this volume, Chap. 2) draw this conclusion based on their investigation of word-based HTER (WHER). To that end, we calculated Pearson correlations between all the different metrics (see Fig. 2). Out of the 55 metric pairs, the differences between them are statistically significant at 95% confidence interval for 45 pairs.[4]

Looking at the results it is clear that the correlation between the metrics is overall quite low, with only a few exceptions. For example, *keystrokes* correlate moderately to well with *editing time*, *pause count*, and *editing pause count*. Whereas it is interesting to see these metrics (associated with all the dimensions) correlate,

[4]The differences are not statistically significant for the following pairs: total time and pause time; total time and editing pause time; total time and first pause; total time and last pause; editing time and pause count; editing time and editing pause count; editing time and keystrokes; pause time and editing pause time; pause time and first pause; pause time and last pause.

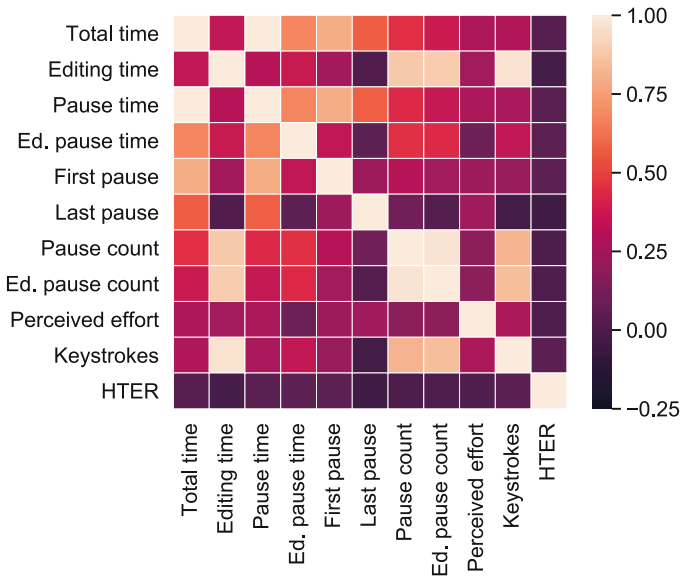


Fig. 2 Pearson correlation heat map between all the metrics where darker blocks represent lower correlations and lighter blocks represent higher correlations

this outcome is not unexpected. While it is true that correlations emerge, it could be argued that these metrics only address the effort of a limited part of the post-editing activity and do not consider the complete process. To be precise, the activity monitored by these specific metrics is restricted to the editing period of the post-editing process, which is when keystrokes are used, and when pauses are computed.

We can also observe that *total time* correlates from moderately to well with *pause time* and some other pause-related metrics such as *editing pause time* and *first pause*. However, these high correlations might be artifacts of the methodology used, as the presence of a single error to address is expected to result in similar measurements for the time-related metrics addressing non-editing periods.

It is interesting to note that *HTER* has particularly low correlations with all the metrics. It could be argued that, given the poor correlation between *HTER* and all the other metrics, including *keystrokes* and *total time*, this could indicate that this metric is not generally representative of post-editing effort and that it should be used with caution. A similar case could be made of the *perceived effort*, whose results do not correlate with any other metric, including those also associated with the cognitive dimension.

If we consider the correlations between the metrics for each of the error types separately (see Fig. 3), we can observe that the patterns remain the same. However, in this case, we can see that the level of correlation is slightly higher for some categories except for *HTER*, which remains at very poor levels. For example,

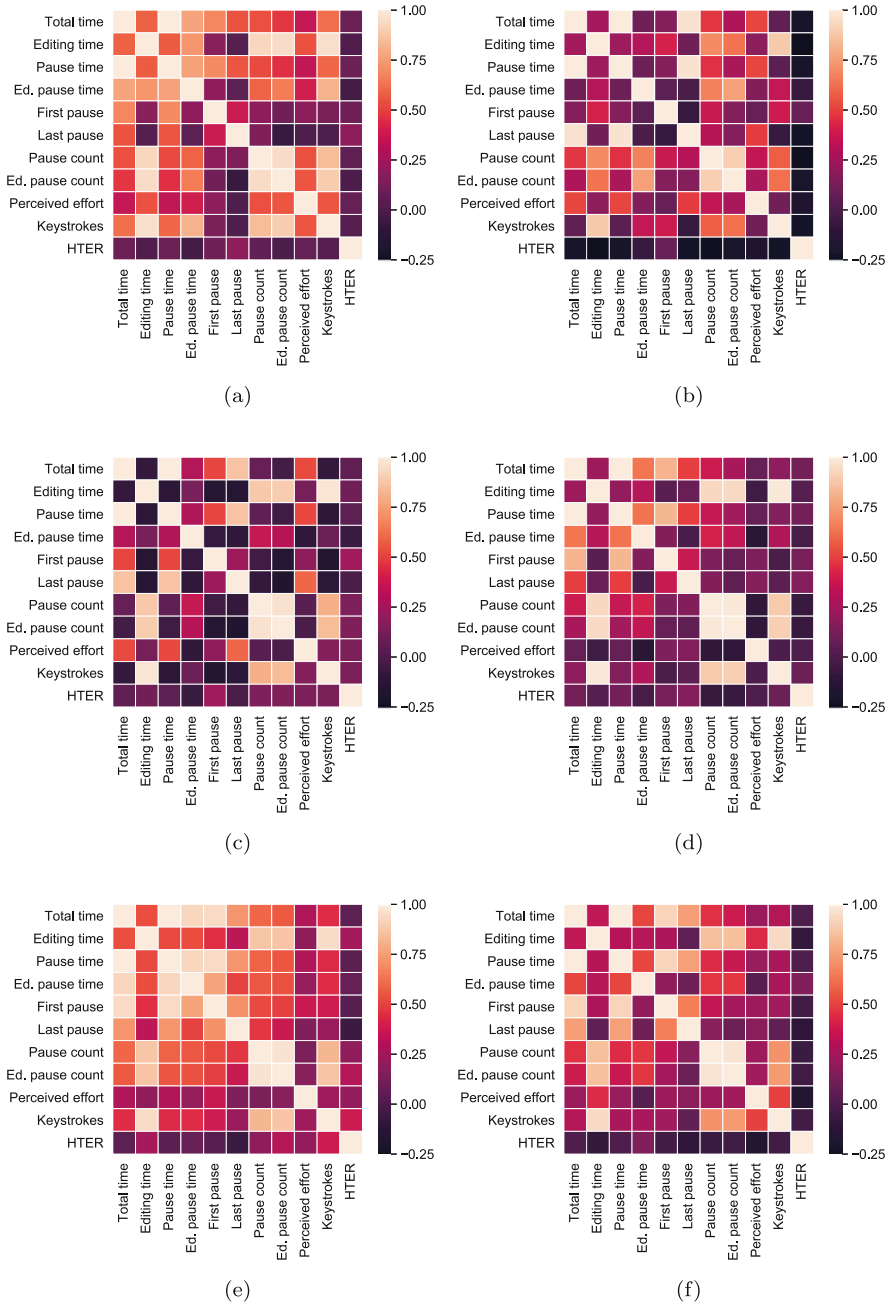


Fig. 3 Pearson correlation heat map between all the metrics per error category where darker blocks represent lower correlations and lighter blocks represent higher correlations. (**a**) Agreement N/G (**b**) Agreement T/A (**c**) Extra word (**d**) Missing word (**e**) Mistranslation 1 (**f**) Mistranslation 2+

T/A agreement and *extra word* have overall low correlations, whereas correlations between metrics for *N/G agreement* and *mistranslation 1* improve.

4.3 Distributions of Errors Per Metric

The second objective of this work is to examine whether metrics can detect differences in post-editing effort for the different error types. Since errors are assumed to have an impact on the difficulty of the sentence, we assume the results will show noticeable differences between them. We also assume that Temnikova's (2010) error ranking will be confirmed by the results. This section presents a series of box plots that compare the results for the error types per metric, which have been grouped by effort dimension. Finally, we discuss the general patterns that were found. The divisions by effort dimensions will allow us to compare the metrics that should show the most similar behavior and check whether there are patterns in the ways errors affect them.

4.3.1 Temporal Effort

The two box plots in this section represent the *total time* and *editing time* (see Fig. 4). The units in the y axes of these plots are milliseconds, and they have been normalized by the number of words in the post-edited versions of the sentences. We decided to normalize the time metrics because not all sentences have the same number of words, so this transformation was necessary if we were to compare them. The x axes show the different errors, ordered following Temnikova's ranking (Temnikova 2010), from *N/G agreement*, the easiest to perform, to *mistranslation 2+*, the hardest.

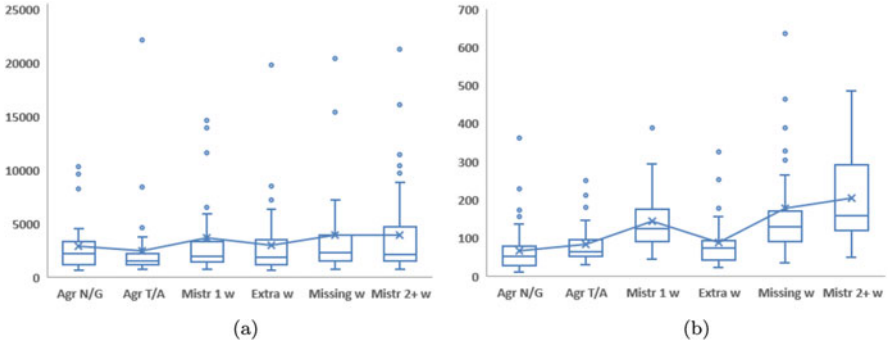


Fig. 4 Distribution graph for total time and editing time metrics where the y axis indicates milliseconds, for each of the error categories. (**a**) Total time (**b**) Editing time

If we look at the *total time* (see Fig. 4a), we observe that the results are very similar across error types. However, differences are more noticeable for *editing time* (see Fig. 4b). *mistranslation 1*, *mistranslation 2+*, and *missing word* require higher times to address than the rest. If we consider the order of difficulty suggested by the *editing time* metric, which repeats over several metrics, we observe that it seems to challenge Temnikova's ranking, proposing an alternative order of difficulty.

4.3.2 Cognitive Effort

We now turn to the metrics used to measure cognitive effort (see Figs. 5 and 6). These were divided into several pause-related metrics and perceived effort. Let us first focus on the pause-related metrics. We can see that the results for *pause time* and *last pause* are very similar for all the error types but the rest show noticeable differences. It is interesting to note that where differences are clear, the results for *extra word* are lower than those for the preceding category, *mistranslation 1*, as happened with the metrics for the temporal dimension. Also, we can observe a tendency for *T/A agreement* to display lower scores than *N/G agreement*. Finally, it is worth mentioning the lower results for *mistranslations 2+* for the *editing pause time*.

If we turn to the perceived effort metric, where 1 is the best possible score and 3 is the worst, we see that the pattern we obtain with regard to the effort ranking for the different error types is similar to those obtained for the other metrics (see Fig. 6). On average, the perceived effort ranged between 1.2 and 1.4, meaning that the difficulty in addressing most sentences was considered between easy to medium. It is also interesting to note that in all cases except *T/A agreement* there were instances of translators choosing all three different possible scores.

4.3.3 Technical Effort

Technical effort was measured using two different metrics: keystroke logging and HTER. Figure 7 shows the results for both metrics, which are rather different. Keystrokes display a similar pattern to the other metrics, considering *mistranslation 1* and *mistranslation 2+* and *missing word* more difficult to post-edit than the rest. These error types imply, on average, 10 more keystrokes than *extra word*, *T/A agreement*, and *N/G agreement*. Considering that addressing mistranslations implies writing entire words, while addressing agreement issues involves either correcting, adding, or deleting only word endings, these differences are to be expected. In contrast, *HTER* results are the ones that differ from the pattern followed by the other metrics to a greater or lesser extent. This does not come as a surprise, as this metric measures the product and so does not account for words that have been changed more than once, while the rest focus on the process. In this case, *N/G agreement*

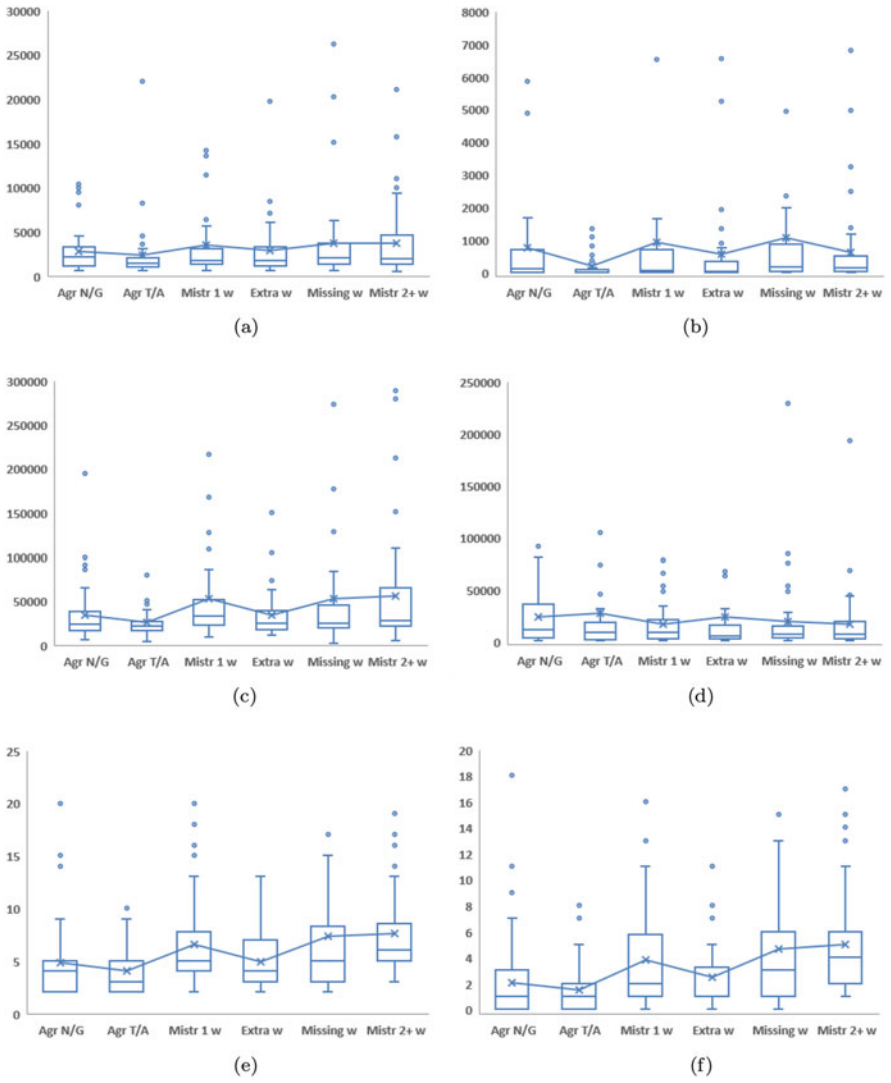


Fig. 5 Distribution graphs for pause-related metrics, where the y axis indicates milliseconds for *pause time*, *editing pause time*, *first pause*, and *last pause*, and counts for *pause counts* and *editing pause count*, for each of the error categories. (**a**) Pause time (**b**) Editing pause time (**c**) First pause (**d**) Last pause (**e**) Pause count (**f**) Editing pause count

gets the worst scores (highest HTER scores), while *mistranslation 2+* obtains the best. It must be noted, however, that in general, the results are quite similar, with averages ranging between 25 and 28. These results seem to show, again, that *HTER* should be used with caution and full awareness of its shortcomings.

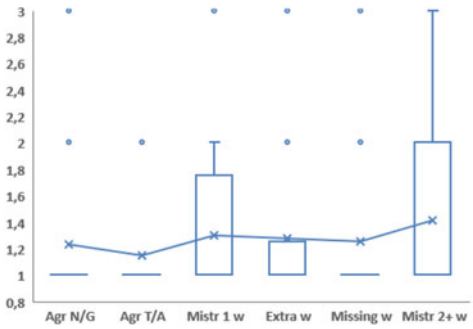


Fig. 6 Distribution graph for the perceived effort metric for each of the error categories where a higher number in the y axis indicates a higher perceived difficulty

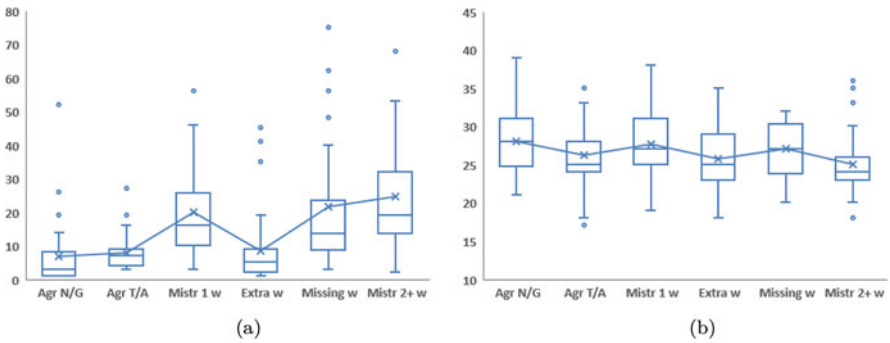


Fig. 7 Distribution graphs for the keystrokes and HTER metrics, where the y axis refers to the number of keystrokes for *keystrokes* and the HTER score for *HTER*. (**a**) Keystrokes (**b**) HTER

4.3.4 Discussion of Results

The comparison of the results obtained by the different metrics showed that, with few exceptions, the correlations between the metrics were rather low. This indicates that any decision about post-editing (schedules, post-editor allocation, difficulty ranking, remuneration, etc.) made based on effort might differ depending on the metric used to measure it. The analysis of correlations done at error type level confirms the previous results, but the results for the metrics related to pauses seem to improve for the *N/G agreement* and *mistranslation 1* error types in particular.

Whereas metrics display very similar results regardless of the error type, some indication that the *mistranslation 1* errors are more difficult to post-edit than *extra word* or *missing word* error types seems to emerge, which partly challenges Temnikova's ranking. Moreover, *T/A agreement* is reported to be easier to address than *N/G agreement*.

Although a pattern seems to emerge in the results, repeating this experiment with more participants, a larger test suite, and more error types would be necessary to confirm these trends and establish a definite effort ranking for error types. It is nevertheless important to acknowledge that the differences between errors seem very limited. While research has focused on linking the presence of errors to

increased effort, it might be the case that when the error is isolated (i.e., no additional errors are present in its near context) its effects on effort are not as distinct as previously assumed. This might be because detecting the error and/or implementing its correction might be more straightforward. It must also be remembered that this research has not considered the presence of specific words or structures in the sentences, different sentence lengths, or the combined effect of different errors within a sentence. It would be interesting to compare the results of this experiment with others that control those aspects.

5 Conclusions

This work aimed to examine whether commonly used metrics to measure post-editing effort correlate with each other, in particular, those associated with different effort dimensions. To that end, we created a purposely built test suite consisting of sentences with one specific type of error each and asked a group of professional translators to edit it. Even when some translators exceeded or downplayed the required editing, we gathered a rather homogeneous set of results for the different error types. The activity of the translators was recorded using PET, which allowed us to collect diverse metrics associated with the different dimensions. In particular, we obtained total time and editing time connected to the temporal effort; pause time, editing pause time, initial and final pauses, pause and editing pause counts, and perceived effort as metrics for the cognitive effort; and keystroke count and HTER for the technical dimension.

Results showed that metrics associated with different dimensions do not correlate well with each other with the exception of total time and pause count. Interestingly, they also showed that metrics addressing the same effort dimension do not correlate either. For cognitive effort, and in line with (Vieira 2016), perceived effort did not correlate well with the different pause metrics. This probably corroborates the fact that asking people to rate the post-editing difficulty of sentences is not a good strategy to obtain reliable results. For the technical dimension, keystrokes and HTER returned very different results and did not correlate. This is probably because that the focus of both metrics is different, as previously pinpointed by authors such as Daems et al. (2017); the keystrokes pay attention to the process, while HTER limits its scope to the final product. It is interesting to see that Huang and Carl (this volume, Chap. 2) report different results when working with multiple errors per segment. Although the adequacy of HTER to measure post-editing effort is not yet fully understood, it is a very common tool in industry, and it is used as the basis for quality estimation models. The fact that HTER correlates so poorly with all the other metrics regardless of the dimension should cause those using this metric as the sole source of post-editing effort measurements to pause and reflect on their decision. Regarding keystrokes, as expected, they did correlate quite well with editing time, pause count, and editing pause count, probably because all of them restrict their focus to key-pressing time, and pauses happen between edits.

Results also showed that the differences reported by the metrics studied are similar regardless of the error type. This might indicate that the capacity of the above-mentioned metrics to detect effort differences by error is rather limited. Interestingly, previous research has concluded that the presence of certain errors greatly affects the post-editing difficulty (Koponen 2012; Lacruz et al. 2014). Considering that the main difference between those studies and the present experiment is that we restricted to one the number of errors in each sentence while the others considered sentences with multiple errors, this may indicate that the difficulty contributed by each error is accentuated when several are present. Concerning the error ranking proposed in Temnikova (2010), even if differences are minimal, some patterns emerge in the results that repeat themselves over most metrics, suggesting a more refined rank. *N/G agreement* seems to require more effort to fix than *T/A agreement*, and *mistranslation 1* appears to be more difficult to deal with than *extra word* or *missing word*.

Within its limited scope, this work provides further evidence to suggest that commonly used metrics for measuring post-editing effort might not correlate well with each other nor are they able to consistently differentiate the effort required by different error types. Moving forward, research that focuses on capturing post-editing effort creating metrics that consider the different dimensions is still necessary, specially with regards to obtaining a metric that is easily scalable and affordable. Additionally, a way of linking post-editing effort with source text and MT output characteristics in the line of O'Brien (2005), Temnikova (2010), and Daems et al. (2015), for example, would facilitate the identification of MT output that is adequate for post-editing.

Acknowledgments The research leading to this work was partially funded by the Spanish MEIC and MCIU (*UnsupNMT* TIN2017-91692-EXP and *DOMINO* PGC2018-102041-B-I00, co-funded by EU FEDER), and the *BigKnowledge* project (BBVA foundation grant 2018).

References

Alves F, Szpak K, Luiz Gonçalve J, Sekino K, Aquino M, Castro R, Koglin A, Fonseca N, Mesa-Lao B (2016) Investigating cognitive effort in post-editing: a relevance-theoretical approach. In: Hansen-Schirra S, Grucza S (eds) Eye-tracking and applied linguistics. Language Science Press, pp 109–142

Aranberri N (2017) What do professional translators do when post-editing for the first time? First insight into the Spanish-Basque language pair. HERMES J Lang Commun Bus 56:89–110

Aranberri N, Pascual JA (2018) Towards a post-editing recommendation system for Spanish–Basque machine translation. In: Proceedings of the 21st annual conference of the European association for machine translation. European Association for Machine Translation, pp 21–30

Arnold D, Moffat D, Sadler L, Way A (1993) Automatic test suite generation. Mach Transl 8(1–2):29–38

Aziz W, Castilho S, Specia L (2012) PET: a tool for post-editing and assessing machine translation. In: Eighth international conference on language resources and evaluation. European Language Resources Association (ELRA), pp 3982–3987

Blain F, Senellart J, Schwenk H, Plitt M, Roturier J (2011) Qualitative analysis of post-editing for high quality machine translation. In: Proceedings of the XIII machine translation summit. Asia-Pacific Association for Machine Translation, pp 164–171

Burchardt A, Harris K, Rehm G, Uszkoreit H (2016) Towards a systematic and human-informed paradigm for high-quality machine translation. In: Proceedings of the LREC 2016 workshop "translation evaluation – from fragmented tools and data sets to an integrated ecosystem". European Language Resources Association (ELRA), pp 35–42

Carl M (2012) Translog-II: a program for recording user activity data for empirical reading and writing research. In: Eighth international conference on language resources and evaluation. European Language Resources Association (ELRA), pp 4108–4112

Carl M, Schaeffer MJ (2017) Models of the translation process. The handbook of translation and cognition, pp 50–70

Carl M, Dragsted B, Elming J, Hardt D, Jakobsen AL (2011) The process of post-editing: a pilot study. Copenhagen Stud Lang 41:131–142

Daems J, Vandepitte S, Hartsuiker R, Macken L (2015) The impact of machine translation error types on post-editing effort indicators. In: 4th workshop on post-editing technology and practice (WPTP4). Association for Machine Translation in the Americas, pp 31–45

Daems J, Vandepitte S, Hartsuiker R, Macken L (2017) Identifying the machine translation error types with the greatest impact on post-editing effort. Front Psychol 8:1282

De Almeida G (2013) Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two romance languages. Ph.D. thesis, Dublin City University

De Almeida G, O'Brien S (2010) Analysing post-editing performance: correlations with years of translation experience. In: Proceedings of the 14th annual conference of the European association for machine translation, pp 26–28

Dragsted B, Hansen I (2009) Exploring translation and interpreting hybrids. the case of sight translation. Meta: journal des traducteurs/Meta: Translators' Journal 54(3):588–604

Flanagan M, Christensen TP (2014) Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. Interpreter Transl Trainer 8(2):257–275

Gaspari F, Almaghout H, Doherty S (2015) A survey of machine translation competences: insights for translation technology educators and practitioners. Perspect Stud Translatology 23:1–26

de Gibert O, Aranberri N (2019) Estrategia multidimensional para la selección de candidatos de traducción automática para posedición. Linguamática 11(2):3–16

Guerberof A (2009) Productivity and quality in mt post-editing. In: Proceedings of the MT summit XII-Workshop: beyond translation memories: new tools for translators MT. Association for Machine Translation in the Americas, pp 1–9

Guerberof A (2013) What do professional translators think about post-editing? J Specialised Transl 19:75–95

Guillou L, Hardmeier C (2016) Protest: a test suite for evaluating pronouns in machine translation. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16). European Language Resources Association (ELRA), pp 636–643

Hu K, Cadwell P (2016) A comparative study of post-editing guidelines. Baltic J Modern Comput 2:346–353

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the seventh workshop on statistical machine translation. Association for Computational Linguistics, pp 181–190

Koponen M (2016) Machine translation post-editing and effort: Empirical studies on the post-editing process. University of Helsinki, Helsinki

Koponen M, Salmi L, Nikulin M (2019) A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. Mach Transl 33:61–90

Krings H (2001) Repairing texts: empirical investigations of machine translation post-editing processes. The Kent State University Press

Lacruz I, Shreve G (2914) Pauses and cognitive effort in post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (Eds.), Post-editing of Machine Translation: Processes and Applications. Cambridge Scholars Publishing, pp 246–272

Lacruz I, Shreve G, Angelone E (2012) Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In: Proceedings of the workshop on post-editing technology and practice (WPTP). Association for Machine Translation in the Americas, pp 1–10

Lacruz I, Denkowski M, Lavie A (2014) Cognitive demand and cognitive effort in post-editing. In: Proceedings of the third workshop on post-editing technology and practice (WPTP-3). Association for Machine Translation in the Americas, pp 73–84

Martínez-Gómez P, Han D, Carl M, Aizawa A (2018) Recognition and characterization of translator attributes using sequences of fixations and keystrokes. Eye tracking and multidisciplinary studies on translation, pp 97–120

Massardo I, van der Meer J, O'Brien S, Hollowood F, Aranberri N, Drescher K (2016) MT post-editing guidelines. TAUS Signature Editions

Mesa-Lao B (2013) Eye-tracking post-editing behaviour in an interactive translation prediction environment. J Eye Mov Res 6(3):541

Moorkens J (2018) Eye tracking as a measure of cognitive effort for post-editing of machine translation. In: Walker C, Federici F (eds) Eye tracking and multidisciplinary studies on translation. John Benjamins, pp 55–70

Moorkens J, O'Brien S, da Silva I, de Lima Fonseca N, Alves F (2015) Correlation of perceived post-editing effort with measurements of actual effort. Mach Transl 29:267–284

Nitzke J, Oster K (2016) Comparing translation and post-editing: an annotation schema for activity units. In: Carl M, Srinivas B, Moritz S (eds) New directions in empirical translation process research. Springer, pp 293–308

O'Brien S (2005) Methodologies for measuring the correlations between post-editing effort and machine translatability. Mach Transl 19:37–58

O'Brien S (2006a) Eye-tracking and translation memory matches. Perspect Stud Translatol 14:185–205

O'Brien S (2006b) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Lang Cult 7(1):1–21

O'Brien S (2011) Towards predicting post-editing productivity. Mach Transl 25:197–215

O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) (2014) Post-editing of machine translation: processes and applications. Cambridge Schoolars Publishing

Parra Escartín C, Arcedillo M (2015) Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In: Proceedings of 4th workshop on post-editing technology and practice (WPTP4). Association for Machine Translation in the Americas, pp 46–56

Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull Math Clin Linguist 93:7–16

Popovic M, Lommel A, Burchardt A, Avramidis E, Uszkoreit H (2014) Relations between different types of post-editing operations, cognitive effort and temporal effort. In: Proceedings of the 17th annual conference of the european association for machine translation, pp 191–198

Probst A (2017) The effect of error type on pause length in post-editing machine translation output. Master's thesis, Tilburg University

Schaeffer M, Nitzke J, Tardel A, oster K, Gutermuth S, Hansen-Schirra S (2019) Eye-tracking revision processes of translation students and professional translators. Perspectives Studies in Translation Theory and Practice 27(4): 589–603

da Silva IAL, Alves F, Schmaltz M, Pagano A, Wong D, Chao L, Leal ALV, Quaresma P, Garcia C, da Silva GE (2017) Translation, post-editing and directionality. Transl Transit Between Cogn Comput Technol 133:107–134

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation in the Americas. Association for Machine Translation in the Americas, pp 223–231

Specia L (2011) Exploiting objective annotations for measuring translation post-editing effort. In: Proceedings of the 15th conference of the European association for machine translation. European Association for Machine Translation, pp 73–80

Specia L, Farzindar A (2010) Estimating machine translation post-editing effort with hter. In: Proceedings of the second joint EM+/CNGL workshop: bringing MT to the user: research on integrating MT in the translation industry (JEG 10), pp 33–41

Tatsumi M (2009) Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In: Proceedings of the XII machine translation summit, pp 332–339

Tatsumi M, Roturier J (2010) Source text characteristics and technical and temporal post-editing effort: what is their relationship. In: Proceedings of the second joint EM+/CNGL workshop bringing MT to the user: research on integrating MT in the translation industry (JEC 10), pp 43–51

Temnikova I (2010) Cognitive evaluation approach for a controlled language post–editing experiment. In: Proceedings of the seventh international conference on language resources and evaluation. European Language Resources Association, pp 3485–3490

Vieira LN (2016) How do measures of cognitive effort relate to each other? a multivariate analysis of post-editing process data. Mach Transl 30:41–62

Wisniewski G, Singh AK, Segal N, Yvon F (2013) Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In: Proceedings of the XIV machine translation summit. European Association for Machine Translation, pp 117–124

Measuring Effort in Subprocesses of Subtitling

The Case of Post-editing via Pivot Language

Anke Tardel

Abstract There has been noticeable growth in the use and production of intralingual and interlingual subtitles due to technological advances and accessibility legislation. While the reception of subtitles has been increasingly studied over the years, there are only a few empirical studies that investigate the process of subtitling. This contribution gives initial results from a study that investigates the impact of reference material during post-editing of NMT of audiovisual content via language. The focus is on transcription and translation processes, the two main subprocesses of the complex task of interlingual subtitling. Applying well-established methods from TPR, key-logging and eye tracking, this study takes a first look at how the integration of language technology, specifically how NMT impacts these subprocesses when used in an indirect translation or pivot setup. In the study, 25 professional subtitlers and translation students were recorded working in three different conditions when post-editing the German NMT output of Swedish movie excerpts. Results evaluate the impact of post-editing, with and without English reference script and/or original Swedish video, on temporal, technical, and cognitive effort which is estimated with established measures based on gaze, typing, and session duration data. All sessions were recorded with Translog-II and eye tracking which to the author's knowledge has not been applied to an audiovisual context yet apart from Huang and Carl (this volume, chapter "Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort").

Keywords Audiovisual translation · Post-editing · Effort · Eye tracking · Key-logging · Process research

A. Tardel (✉)
Tra & Co Center, Johannes Gutenberg-Universität Mainz, Germersheim, Germany
e-mail: antardel@uni-mainz.de

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_4

1 Introduction

The continuous growth of media production, technological advances, and accessibility legislation have had a huge impact on the AVT industry, particularly on intralingual and interlingual subtitling. While the reception of subtitles has been increasingly studied over the years with eye tracking studies ranging from processing native and foreign subtitles to comparing different style guides (see Doherty and Kruger 2018), there are only a few empirical studies that investigate the process of subtitle production. The focus has been mainly on subtitle processing rather than on production processes of subtitles.

A closer analysis of the processes involved in subtitling becomes more imperative with subtitling being under increasing pressure regarding higher demand and throughput expectations which result in an increased use of language technology such as *Automatic Speech Recognition* (ASR) and NMT in AVT, specifically subtitling. Innovations in ASR, and substantial quality gains in NMT for creative texts (e.g., Moorkens et al. 2018; Toral et al. 2018) suggest that PE could boost productivity not only in written translation of literary texts but in AVT contexts as well (Huang and Carl: this volume, chapter "Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort"). While the industry continuously introduces new tools and platforms that include the mentioned language technology, empirical research is only slowly catching up to these developments.

So far if subtitle production was studied, it was in small-scale case studies and surveys such as the ones carried out by Beuchert (2017) and Künzli (2017) with a focus on traditional subtitling without language technology. Only two process studies were carried out using eye tracking and key-logging: Hvelplund (2017) investigated attention distribution and cognitive effort during the translation of audiovisual (AV) content for dubbing, and Orrego-Carmona et al. (2018) compared subtitling students and professional subtitlers regarding their temporal, cognitive, and technical effort in interlingual subtitling. In contrast to these two studies, the investigation presented in this chapter focuses only on subprocesses involved in subtitling, i.e. transcription and translation of movie dialogue in short video sequences by PE of ASR and NMT. The aim is to gain insights that can assist in interpreting integrated subtitling processes which apply innovative language technology such as ASR and NMT but also practices such as indirect translation, i.e. subtitling via a pivot language (Díaz-Cintas and Remael 2014; Vermeulen 2011, 32).

Indirect translation or pivot translation is the practice of translating not directly from the source language but via a more popular intermediate language (e.g. English) to counter lower resource language combinations (Assis Rosa et al. 2017). Thus, instead of translating from language A (Swedish) to language C (German), the text is first translated from Swedish to language B (English) and then from English to German. In pivot translation this indirect translation is performed in both directions (A-B-C and C-B-A). Indirect translation has also been studied in the MT context by e.g. Liu et al. (2018) who found that pivot MT translation

compared to direct MT did not necessary yield better results. They therefore suggest to incorporate quality estimation and/or a PE step, which is what was done in the present study (Liu et al. 2018, 10).

There have been a number of research projects that attempted integrating ASR and NMT in captioning processes as presented in Sect. 2 but none of them studied the process with behavioral data. This development is also central to the EU-funded COMPASS[1] project. The study was carried out within this project, which is described in Sect. 3 together with the associated conditions and research questions.

To the best of the author's knowledge, there are no empirical studies yet that specifically look at the impact of PE of NMT in indirect transcription processes for subtitling nor at the impact that experience in subtitling might have on it. It should be noted that this chapter mainly focuses on the presentation of the methodology as well as initial results on effort during PE via a pivot language. More detailed results than presented within the scope of this chapter, i.e. analyses on the segment level as well as a look on TT quality, and on efficiency in the conditions will be presented and thoroughly discussed in the author's dissertation (Tardel forthcoming).

The outline of this chapter includes, first, a review of relevant empirical research on PE in AVT (Sect. 2), followed by a description of the project COMPASS and a presentation of the research questions regarding the part of the main study that is relevant to this chapter (Sect. 3). The second half of the chapter is concerned with the methodology (Sect. 4) and study design (Sect. 5) as well as the presentation of initial results (Sect. 6). This is then followed by a brief discussion (Sect. 7), as well as a conclusion and outlook to further research (Sect. 8).

2 Research on Computer-Assisted Subtitling

Since the advent of NMT, the productivity gain found in PE has increased, while still heavily depending on the sentence length, domain, and language pair. Toral and Sánchez-Cartagena (2017), for example, found using "the best PBMT and NMT constrained systems submitted to the news translation task of WMT16" that the NMT systems performed better than the PBMT systems. They tested with various language combinations and especially on shorter segments and their results are in line with findings from Bentivogli et al. (2016) focusing on the English–German language pair. In their small-scale study with technical texts, they found a decrease in PE effort of 26% compared to the best phrase-based system.

In AV content, verbal utterances can be expected to be shorter than in written communication which can be confirmed by looking at the transcripts in this study which were made up of 22–34 segments with an average segment length between 6 and 9 words per segment. Similar to the findings of Bentivogli et al. (2016), when

[1]COMPASS – Computer-Assisted Subtitling, https://www.compass-subtitling.com/project-1 (last accessed: November 2020).

comparing NMT and phrase-based MT, Klubička et al. (2017) found a decrease in fluency errors but there is also evidence that errors of NMT are more difficult to spot.

Toral et al. (2018) investigated NMT with creative texts and compared effort during PE of different MT approaches (NMT and phrase-based MT) to effort during translation from scratch and found that NMT outperformed phrase-based MT. In a key-logging study, Jia et al. (2019) compared effort during PE of NMT to translation from scratch with domain-specific and general texts. They found that PE of NMT significantly reduces temporal effort only for domain-specific texts. Cognitive effort was reduced during PE of NMT compared to translation from scratch for both text types while maintaining equivalent fluency and accuracy. For an extensive overview of studies that measured cognitive effort during PE with eye tracking in written translation of technical and general domains, see Moorkens (2018).

After the successful application of PE in written translation over the past two decades, researchers have also looked into the use of MT in automatic subtitling and captioning. While PE is standard in written translation, it is typically not used for subtitles, although the industry is adapting and Bywood et al. (2017) even propose the new role of subtitling post-editor.

With a growing research interest and advances in both ASR and MT, these technologies find increasing application in the field of captioning massive open online courses and motivated research projects such as TransLectures[2] with phrase-based MT (Silvestre Cerdà et al. 2012; Valor Miró et al. 2014), and more recently TraMOOC[3] (Kordoni et al. 2016; Castilho et al. 2017, 2018), where statistical SMT and NMT were compared in educational contexts. In Castilho et al. (2018), the authors found overall fewer errors as well as reduced technical effort and slightly reduced temporal effort for PE of NMT across four language pairs: English to German, Greek, Portuguese, and Russian.

Similar attempts have been made in the domain of TV subtitling by training phrase-based MT systems on subtitle data (cf. Armstrong et al. 2006; Flanagan 2009; Volk et al. 2010). The concept of combining ASR and MT with translation memories (TM) is not a new idea in AVT and was tested in projects such as MUSA IST[4] (Multilingual Subtitling of Multimedia Content) (Piperidis et al. 2004) and eTITLE[5] (Melero et al. 2006). While, at the time, the ASR did not perform well enough, results showed that systems performed better when combining MT with a TM.

[2] TransLectures – Transcription and Translation of Video Lectures, https://www.mllp.upv.es/projects/translectures/ (last accessed: November 2020).

[3] TraMOOC – Translation for Massive Open Online Courses, https://www.k4all.org/project/tramooc/ (last accessed: November 2020, original project website no longer available).

[4] MUSA IST Project, http://sifnos.ilsp.gr/musa/ (last accessed: November 2020).

[5] eTITLE – European multilingual transcription and subtitling services for digital media content, https://cordis.europa.eu/project/id/22160 (last accessed: November 2020, original project website no longer available).

De Sousa et al. (2011) carried out a study comparing PE effort in the translation of subtitles assisted by a black-box rule-based MT system, a TM, and an in-domain phrase-based MT engine as well as Google Translate SMT finding similar efficiency gains as in written translation. Even in the smaller corpus, the results showed high correlations between temporal effort and participants' perceived effort, better scoring for shorter segments, and a clear productivity gain for pre-translated segments either by TM or MT.

In the more recent SUMAT[6]project, a large-scale evaluation of phrase-based MT trained on professional and crowd-sourced open domain subtitles for both scripted and non-scripted material was carried out for eleven language pairs (Etchegoyhen et al. 2014). They found varying results for quality and productivity gain/loss depending on the language pair and direction but overall a higher percentage of positive evaluations, and a 40% productivity gain based on data of two subtitlers per language pair.

These studies mostly relied on SMT and original pre-segmented subtitles in template files with fixed spotting to be uploaded for the PE of MT and TM segments. PE effort was measured with translation time and perceived effort in subjective ratings, but also with automatic measures comparing the post-edited subtitles to the raw MT output. A recent study by Matusov et al. (2019) looked into the customization of NMT to subtitling with a particular focus on automatic segmentation. In this small-scale study with two translators, they found a productivity gain of 37% compared to subtitling from scratch. Also, they recorded lower HTER scores compared to the non-adapted NMT with edit rates based on a comparison of MT output and post-edited text, which according to Cumbreño and Aranberri (this volume, chapter "What Do You Say? Comparison of Metrics for Post-editing Effort") might not be the best indicator for PE effort. Unfortunately, studies of NMT in AVT are still scarce and typically no process data such as gaze and typing behavior is recorded which can provide more evidence on the actual technical or cognitive effort in these processes (cf. Huang and Carl: this volume, chapter "Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort").

All the abovementioned studies have in common that they worked with pre-segmented and spotted subtitles to be translated or post-edited. The use of template files, most commonly based on English, has been practiced for quite a while in subtitling. In AVT research, however, it has been picked up only recently (Georgakopoulou 2019). Instead of translating subtitles from scratch, subtitlers work with a pre-spotted empty or often English template file in which they type the target subtitles without adjusting the timing. The use of English as pivot language in template files is not widely accepted but still practiced especially in large multilingual projects to cover more language combinations. In this form of indirect translation, the English subtitles become the new ST and during translation subtitlers sometimes do not have access to the video as they work with fixed subtitles. These

[6]SUMAT: An Online Service for Subtitling by Machine Translation, http://www.fp7-sumat-project.eu/ (last accessed: November 2020).

practices have shown to increase productivity in the process but have always given rise to debates particularly regarding TT quality and information loss (Artegiani and Kapsaskis 2014; Nikolić 2015).

Subtitles are condensed versions of the spoken content leaving out, among others, easily omittable items such as repetitions, exclamations, or emphatic phrases (Georgakopoulou 2019), or they contain shorter equivalents that might not be the first translation choice in a written translation. This, in combination with practices where access to the original video is limited due to copyright regulations, can have a big impact on effort during indirect translation but also on quality regarding content errors. The study, which is described in more detail in the next section, enters the conversation on pivot subtitling as it investigates the indirect translation with PE of transcripts instead of subtitles with and without access to the video.

3 COMPASS Project

The study presented in this chapter was carried out within the scope of the project COMPASS (Computer-Assisted Subtitling) which was managed by ZDF Digital and Johannes Gutenberg University of Mainz and funded by the European Commission between January 2018 and June 2019. The project's aim was to optimize the overall multilingual subtitling processes for offline public TV programs and video-on-demand platforms. For this, conventional workflows for the creation of intralingual and interlingual subtitles were reviewed and transferred to a uniform process model within a platform leveraging state-of-the-art ASR and NMT.

The focus of the COMPASS project was that involved processes should be assisted by automization wherever possible, while a combination of technology, human–machine interaction, and machine learning approaches optimizes and continuously helps improve processes at crucial points such as import, transcription, translation, segmentation, and quality assurance. COMPASS attempted to combine human and machine input to make the process of creating subtitles as efficient and fit for purpose as possible, from uploading the original video until burning in the final subtitles.

3.1 A Proposed Workflow for Subtitling

There is not much empirical process data available on subtitling and even less on the integration of language technology in the subtitling process. Therefore, within the framework of the COMPASS project, two eye tracking studies were carried out.

The first small-scale analysis of the intralingual subtitling process with eight subtitlers helped defining bottlenecks by triangulating data from eye tracking and questionnaires with the produced subtitles (Tardel et al. 2020, forthcoming). The results were used to propose the COMPASS pipeline foreseeing the use of ASR in

order to extract a transcript of the films as a first step, followed by human PE of the ASR texts before subtitlers convert them into intralingual subtitles if requested.

As the workflow model should also be applied in larger multilingual projects, it foresees that the transcripts (or subtitles) are then translated via NMT into English as pivot language or target languages. After human PE of the English NMT transcript (or subtitles), more NMT and PE steps into other target languages may follow. This is in contrast to pivot MT for lower resource languages, where no intermediate PE step is included (Liu et al. 2018). By post-editing subtitles, or transcripts as support for subtitling, the hope is to make significant gains in productivity while maintaining acceptable quality standards even in indirect translation via a language different from the AV content's original language.

At this point, the question arises whether NMT of subtitles, especially with translation via pivot language, is the right workflow or whether transcripts should be translated and post-edited before being converted into subtitles instead. A comparison of the two workflows is not tested in this study, but the assumption is that PE of transcripts instead of subtitles could be performed by translators instead of highly specialized subtitlers. This could free up resources for the actual subtitling and is the reason why a between-groups comparison was included in the main study described below.

3.2 Study on Subprocesses in Subtitling

The second study fits into the young field of *subtitling process research* (SPR), a sub-discipline of TPR and AVT (Beuchert 2017; Orrego-Carmona et al. 2018). In SPR, empirical methods such as eye tracking and key-logging are applied to describe and predict behavior in audiovisual translation and subtitling processes. The focus of this main study is not on subtitling but on transcription and translation as subprocesses involved in interlingual subtitling, and the impact of ASR and NMT on different levels of effort within these subprocesses.

In the study, professional subtitlers and student translators were recorded during computer-assisted interlingual transcript generation for AV content. Given that subtitling is a complex task, interpreting key-logging and eye tracking data in the overall subtitling process can be complicated because timing, space constraints, and segmentation influence the behavior in a way that is difficult to control and to compare.

Furthermore, because of its polysemiotic nature, AV content is not as easy to manipulate as written text, because more aspects come into play (e.g. videos) that make it difficult to select suitable STs. Besides word count and frequency, videos can have differing video duration, words per minute, number of speakers, background music, and other non-verbal elements that impact the comprehension of the AV ST and how it is rendered in the written representations. Subtitling as well as transcription, as a form of polysemiotic translation (Gottlieb 1994), involves the processing of information from different semiotic channels.

In the present study, the following abbreviations are used to differentiate between the different resources involved in the tasks which are motivated by the proposed COMPASS pipeline described above:

- *video*: The *audiovisual ST*, i.e. the Swedish video in the PE task. In the conditions, video is abbreviated with V.
- *ST*: The *written representation of the video*, i.e. a transcript in the same language as the video. The ST served as input to the NMT for translation into the pivot language English (see PTT below).
- *PTT*: The *written representation of the video in the pivot language*. The PTT is a post-edited transcript of the Swedish to English NMT output which served as input for the NMT into German and as English reference script which is indicated by S in the conditions.
- *TT*: The *written representation of the video in the target language*, in this study the German transcript. In the PE task, TT is the NMT output of the PTT and is the text that is being post-edited.

Similar to translation for dubbing, subtitlers need to "coordinate and organize information from acoustic as well as visual channels [of the video] during the translation process" in addition to dealing with the ST or PTT (if available), and the evolving TT, (Hvelplund 2017, 110). Access to the video is therefore critical even if the audio might be in a language the subtitler does not understand as is the case in pivot subtitling. Even in interlingual transcription processes, where the segmentation, condensing, and synchronization do not play a role, the video provides relevant context to the subtitler. In transcription for subtitling as proposed in COMPASS, the ST, PTT, or TT should serve as a useful point of departure for subtitlers to efficiently turn the written and translated dialogue in the target language into subtitles which convey the original message to the target audience. This can be compared to translation for dubbing which as investigated with student translators by Hvelplund (2017).

Subtitling and also transcription can be quite time-consuming even for short video sequences, which poses a challenge for the application of eye tracking to the task. Therefore, among other reasons, this study focuses on transcription, which is less time-consuming than interlingual subtitling because it does not involve spotting, condensing, and segmentation of the text. This can be seen as a limitation of this exploratory study because subprocesses in subtitling cannot be analyzed in isolation. Nevertheless, this study contributes to fundamental research in this field with results guiding the conception and interpretation of more applied subtitling studies.

Transcription and translation are two subprocesses involved in interlingual subtitling and in this study, these tasks were recorded with eye tracking and key-logging to investigate the impact of ASR and NMT on different levels of effort. For the verbatim transcription, the definition "written representation of audible speech" by Matamala et al. (2017, 2) was taken and extended with translation into the target language. In the transcripts, speaker changes are indicated, but not labeled with names and sound is not described. Overall in the main study, three tasks were recorded: intralingual transcription, translation, and PE. First results from the main

study comparing all three tasks were presented in Hansen-Schirra et al. (2020). The results of the first two tasks (transcription and translation) with a focus on the impact of ASR and experience are discussed in Tardel (2020, forthcoming). This chapter complements these initial findings with a focus on the PE task described below. In the PE task, participants post-edited three German NMT transcripts (TT) of Swedish videos under three conditions:

1. *PE+V+S*: assisted by the Swedish video and a reference script in the pivot language English (PTT), i.e. maximum support
2. *PE+V*: assisted only by the Swedish video, thus performing monolingual PE with video
3. *PE+S*: assisted only by a reference script in the pivot language English (PTT), thus performing traditional PE without access to the video

The approach is to apply PE to transcripts that can be used in a next step as support in pivot-based interlingual subtitling. Translating already segmented and condensed subtitles from a pivot language could introduce more errors, as the translator cannot confirm the original verbal content in the foreign language. At the same time, if translators only work with the written representations of the video but do not have access to the video itself, the missing context information might lead to mistakes in the translation. Temporal, technical, and cognitive effort in subtitle production (transcript creation) has not been empirically tested with recent NMT. With the above listed conditions, this study addresses the following research questions:

1. *RQ1: Temporal Effort*
   Is PE with more support (English PTT and the foreign video) significantly slower than PE with only one reference?
2. *RQ2: Technical Effort*
   Does PE with more support require significantly more technical effort than PE with only one reference?
3. *RQ3: Cognitive Effort and Visual Attention*
   Is PE with more support significantly more demanding and does it require switching attention more often than PE with only one reference?
4. *RQ4: Experience in AVT*
   Is effort on these three levels lower for professional subtitlers than for translation students not used to AVT?

The present study investigates how the amount of reference material (video and English PTT) provided during PE of the NMT video transcripts impacts effort on three levels and regarding to participants' experience. The three conditions are inspired by practices and study setups that include monolingual PE (Nitzke 2019) and subtitle template translation without access to the video (De Sousa et al. 2011). The aim is to gain a better understanding of the role of the video in the polysemiotic PE process and the role of the English PTT which can be seen as a partial ST in the polysemiotic pivot translation.

For the interlingual transcript creation of AV content, the assumption is that the first condition with the maximum support (both video and English PTT) is similar to translation (or PE) for dubbing (Hvelplund 2017). Participants need to coordinate multiple resources. The second condition with only the English PTT as support, but no video, is similar to written translation or PE, and the third condition with only the foreign Swedish video can be considered to be monolingual PE.

Professional subtitlers are used to working with AV content and written texts while translation students mainly learn to work with written text, unless following specialized training for AVT. In contrast to subtitling, during the translation of transcripts, no additional condensing and segmentation skills or technical handling of subtitling software regarding spotting are necessary. This suggests that this task could also be performed by translators or even translation students. Given the polysemiotic task, however, contrary to written translation, additional information from different semiotic channels comes into play, especially when the video is available, which professional subtitlers are more used to and expected to use more efficient than translation students. Therefore, differences in both groups' behavior and effort are to be expected.

4 Methodology

The presented study applies a combination of well-established methods from TPR to the verbatim transcription and translation of film dialogue. The triangulation of gaze data from eye tracking, typing activities from key-logging, and product data in linear mixed models allows a detailed analysis of the processes on different levels and under varying conditions while taking into account the participant- and text-inherent variances.

In this analysis, the gaze and typing data is used to measure effort on the three levels suggested by Krings (2001) who applied them to compare PE and translation processes. Krings distinguishes between temporal, technical, and cognitive effort which are compared to each other in Cumbreño and Aranberri (this volume, chapter "What Do You Say? Comparison of Metrics for Post-editing Effort"). In this study effort on the three levels is described with the following measures; for a detailed overview of the variables see Table 2 in Sect. 5.3:

- Temporal effort: session completion time
- Technical effort: insertions and deletions (general number of keystrokes)
- Cognitive effort: total fixation count and average visit duration, total reading time, relative attention to video and video replay time

With eye tracking two primary measures are recorded: fixations, where the eye is nearly still and, according to the Eye Mind Assumption (Just and Carpenter 1976), information is assumed to be processed, and saccades indicating rapid eye movements between fixations. Fixation measures include fixation count and duration as well as visit count and duration which are analyzed according to a

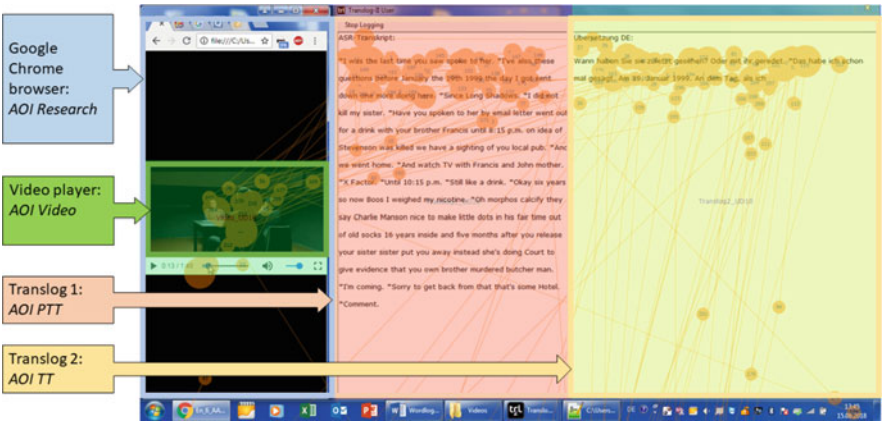


Fig. 1 Screenshot of the setup with gaze plots and respective AOIs set in Tobii Studio

reference area on screen, also known as area of interest (AOI). The AOIs defined in this study are indicated in Fig. 1.

For the analysis, the screen was divided into the four AOIs *Research*, *Video*, *PTT*, and *TT*. Not all AOIs were visible at all times and not in all conditions:

- AOI Research was only visible and active, i.e. collecting gaze data, when participants performed online research.
- AOI Video was visible in condition P+V+S and P+V and activated whenever the participant replayed the Swedish video. The rest of the time no gaze data on the video was collected.
- AOI PTT was visible only in conditions P+V+S and P+S and activated from the first click into Translog-II until the participant clicked "stop logging." This AOI is on Translog window 1, which contained the English transcript and was not editable by participants.
- AOI TT was always visible in all three conditions and also activated from the first click into Translog-II until the participant clicked "stop logging." This AOI is on Translog window 2, which contained the NMT which participants post-edited.

A Tobii TX300 remote eye tracker was used with Tobii Studio (version 3.3) and the eye tracking plug-in for Translog-II (Carl 2012, version 2.0). The default Tobii Fixation Filter was applied recording the strict average of both eyes and with a velocity threshold of 35 px/window and a distance threshold of 35 px. Tobii Studio was used for screen recording, eye tracking, definition of AOIs, and descriptive AOI-based analyses covering the overall setup on screen.

The AOIs on the left side of the screen (Fig. 1) include the Google Chrome video player (excluding navigation bar) and the browser window (online research). These AOIs were activated when the participant replayed the video or performed online research. The rest of the time these AOIs were deactivated, i.e. not collecting gaze data.

Table 1 Overview of tasks and conditions with metadata on language (SV-Swedish, EN-English, and DE-German) and number of recordings (N) included in the analysis in total and per video respectively

Nr.	SL>TL	Condition	Description	N (V1/V2/V3)
6	SV/EN>DE	PE+V+S	PE with video and EN script	25 (9/9/7)
7	SV>DE	PE+V	PE with video only	25 (7/8/9)
8	EN>DE	PE+S	PE with EN script only	24 (8/5/9)

The other AOIs on the right side of the screen (Fig. 1) were drawn on the Translog-II window(s) in which participants were presented with the transcripts and MT output and post-edited the TTs. Depending on the condition, PTT displayed an assisting English transcript and TT contained the German MT output where the participants typed and post-edited the translated transcript. The Translog AOIs were activated from the first click into the editor until participants clicked "stop logging."

Translog-II is a screen-based key-logging tool that logs insertions and deletions in combination with eye tracking data on both windows within the Translog-II editor while the text evolves. Data is recorded in a format that allows post-processing such as the alignment of PTT and TT segments and tokens as well as its integration in the CRITT TPR-DB (Carl et al. 2016). Translog-II is widely used in TPR and makes it possible to record and analyze process data of the evolving TT in relation to the fixated PTT. A drawback is that it cannot collect data outside of Translog-II, such as online research or the replaying of the video, which in this study is the true ST.

Therefore and also because not all conditions contained a ST (or in this case PTT), the analysis of the ET data in this study was performed with the AOIs in Tobii studio. Calibration was performed both in Tobii Studio and Translog-II with 5 points on screen with participants seated at a distance of 60 cm to the eye tracker. The eye tracking data in this analysis is only from Tobii Studio, leaving the more thorough, segment and word-based analysis with the ET mapping in Translog-II for the author's dissertation (Tardel forthcoming).

5 Study Design and Procedure

The overall study design consists of three tasks containing a total of eight conditions. Participants performed all three tasks and conditions always in the same order with videos alternating in a balanced pseudo-randomized manner from participant to participant. The tasks were subprocesses in intra- and interlingual subtitling processes: intralingual verbatim transcription, interlingual verbatim transcription, and PE of NMT via English as pivot language.

In the different conditions, the tasks were modified by introducing transcripts of the videos produced by ASR, human, or pre-translated with NMT. In this chapter only the three PE conditions described in Sect. 3 are relevant. An overview of the

three PE conditions with the languages and number of recordings per video (1, 2, and 3) is given in Table 1.

The PE conditions were the last three of the overall eight sessions recorded per participant. In the first PE condition PE+V+S, the German TT was post-edited with the assistance of an English PTT and the original Swedish video. In the second PE condition PE+V, participants performed monolingual PE of the German TT assisted only by the Swedish video. In the final PE condition PE+S, participants post-edited the German TT without the Swedish video and only with the English PTT, basically performing regular written PE.

Per condition, a total of 25 participants were recorded except for one translation student who did not have a valid recording for condition PE+S (P29). The target language was always participants' native language German. For the PE task, the source language was Swedish and the pivot language English. Because participants did not understand Swedish, English-from-Swedish which was used for the pivot translation was considered the new source language. It should be noted that the pivot translation is subject to interference between the actual source language and the target language (or pivot language) which impacts translation choices in the target language as suggested by Toury (1995/2012, 310) in the *Law of Interference*. For a recent work on indirect translation in literary translation (cf. Ivaska 2020).

Prior to the recording, participants were informed about the methodology and filled out an informed consent form before completing a questionnaire for metadata on language and training background regarding subtitling and translation experience. Every participant started with a copying task in Translog-II to become familiar with the setup in Translog-II and to record typing speeds in the tool. Participants could adjust their headphone volume and get used to the video navigation before the first task. Calibrations both in Tobii Studio and Translog-II were performed prior to every new session to ensure comparable data quality.

The total duration of the experiment, including the two transcription, the three translation, and the three PE conditions, was approximately three hours for all sessions including short breaks between sessions. Fatigue was prevented by brief breaks after each session and by keeping video sequences short with only two minutes per video. Since subtitlers usually work on much longer subtitling tasks without breaks, this is also a limitation to the study, but with longer videos, the eye tracking recordings would require repeated calibration which interrupts the workflow. Furthermore, in Translog-II, text lengths are limited to one page as scrolling negatively impacts eye tracking data collection (Jakobsen et al. 2009).

In the brief for the experiment, participants were provided with the titles of the TV series, and received general instructions on how to post-edit the transcripts which included not to describe sounds or background noises, not to use paragraphs, and to indicate speaker changes with an asterisk, and incomplete utterances or long pauses with ellipses. In the end, the final TT should not contain any placeholders. Specifically regarding the PE conditions, participants were asked to retain as much of the MT output as possible, to refrain from rewriting the entire transcript, and to produce correct transcripts with a focus on content, register, and language, so that it would be useful to a subtitler later in the overall subtitling process.

Furthermore, online research was allowed, as long as they did not search for scripts. In addition to the demographics questionnaire, participants were asked to answer post-task questions relating to the subjective perception of the role and quality of the scripts after each of the three PE conditions.

5.1 Sampling

Participants were sampled by convenience sampling from two groups with differing experience which, in the analysis, is referred to as *status*: translation student (S) or professional subtitler (P).

Translation students ($N = 13$) were all from the translation studies program at University of Mainz and the group of professional subtitlers ($N = 12$) were all freelance subtitlers working in the Berlin area. For the study, a total of 13 professional participants were recorded, but one participant had to be excluded due to poor eye tracking quality. Therefore, only 12 of them were included in the analysis which brings the total number of participants to 25.

All participants were German L1 speakers with English as their L2 and active working language. None of the participants had any knowledge of Swedish, which was important for the PE task. For the participation in the study, all participants were remunerated for the time they invested in the study similar to taking a translation or subtitling job.

There was a bias for female participants in both data sets which is not expected to impact the results and represents the current trends in the industry. The 12 female translation students were between the 3rd and 6th semester ($SD = 4.4$) and the one male translation student was in his 5th semester. Three of the students were in the MA translation and the rest in the middle of their BA studies. The three MA students had only little subtitling and PE experience ($N = 4$). The rest had no subtitling or PE experience at all.

The professional subtitlers (nine female, three male) had at least two years of professional experience in interlingual subtitling with an average of 6.7 years of experience ($SD = 3.5$). All professional subtitlers had either formal training in translation or in AVT, and currently work as freelance subtitlers. Only four of the subtitlers had little PE experience.

5.2 Material

The videos used in this study were all short scenes from three different crime series available on online video-on-demand streaming platforms. For the PE tasks via pivot language, the scenes were taken from three episodes of the Swedish series Before We Die (SVT/ZDF 2017). Participants were asked prior to the study whether they had heard or even watched the series, and only those who answered with no were

allowed to participate. The videos were chosen to be comparable in audio quality, length, text content, and number of speakers. The three Swedish videos were 1:02 min with 196 words (V1), 1:26 with 201 words (V2), and 1:35 with 185 words (V3). Each video made up an internally coherent scene and was taken from different episodes.

The Swedish transcripts were human generated, then machine translated by Google Translate (June 2018) into English and post-edited before further translated with Google Translate into German (TT). The English PTTs were post-edited by a Swedish–English translator and underwent a quality check to provide a correct English reference script (PTT) in the PE tasks of the study.

5.3 Data Analysis

The statistical analysis of the recorded user activity data, i.e., typing and gaze behavior, was carried out in *R* (V. 3.6) with descriptive statistics and LMMs (Baayen 2008). These mixed-effects regression models make it possible to generalize samples (participants and translated videos) to populations (any translator belonging to that group post-editing any video) by distinguishing between fixed effects (effects of interest) and random effects (factors to generalize over).

The models were created using the packages *languageR* and *lme4* (Bates et al. 2015), while *lmerTest* (Kuznetsova et al. 2019) was used to calculate the estimate (β), standard error (SE), degrees of freedom (df), t-value as coefficient divided by its β (t), and significance value (p). The effects of the fitted models were visualized in plots for a better interpretation of each model by applying the *ggplot2* package (Wickham 2016). In the plots, the y-axis indicates the *dependent variable* (DV) and the x-axis the predictor. The DVs were included in several LMMs as shown in Table 2. As random variables, participant and item (i.e., the video) were always included unless stated otherwise. Although video length and word count of the texts were controlled for and participants were selected according to similar experience, there are inherent differences in behavior when dealing with the video and texts that could not be controlled.

The unit for this initial analysis in this chapter is the entire session and the respective AOIs, with the DVs describing temporal, technical, and cognitive effort for the complete sessions and not individual segments or words. Predictors, or fixed effects, were always condition and/or status as the aim was to investigate the effect of support and participant experience on effort (see research questions in Sect. 3). The categorical predictor condition includes the three PE conditions PE+V+S, P+V, and P+S (see Table 1). The categorical predictor status consists of the two groups with different levels of experience: translation students (S) and professional subtitlers (P) as described in Sect. 5.1. The model fit was tested by checking the normal distribution of residuals and collinearity was assessed by inspecting variance inflation factors for the predictors. In the presented models, all values were relatively low (<2), indicating that collinearity between predictors was not a problem.

Table 2 Overview of models with DVs that describe effort on three different levels including brief definition and unit

Effort level	Model	Dependent Variable
Temporal	LMM-1	Duration (min): first click in Translog-II to "stop logging"
Technical	LMM-2	Keystrokes: Total count of insertions and deletions per session
Cognitive	LMM-3	VisitDur PTT (a) and TT (b): Average visit duration on AOI PTT and TT (average in ms)
Cognitive	LMM-4	VisitCount PTT (a) and TT (b): how often participants entered AOI PTT and TT (count)
Cognitive	LMM-5	TrtS and TrtT: Total reading time on PTT (a) and TT (b) sum of all fixation durations on AOI PTT and TT (in min)
Cognitive	LMM-6a	FactorVideo: Video replay duration (video-AOI active) expressed as factor of video duration
Cognitive	LMM-6b	RelativeVideo: Video replay duration divided (%) by the overall session duration

The DVs are categorized according to the level of effort they describe and they were log-transformed in the models unless stated otherwise. Temporal effort is indicated by the time it took participants to complete each session from the first click into the Translog window until they clicked "stop logging." Technical effort is linked to the number of insertions and deletions (total number of keystrokes). Cognitive effort is described with the attention towards the AOIs on screen (see Fig. 1) measured in average visit duration, visit count, and reading time. The AOIs of interest include the PTT, TT, and video. Visit duration is defined as the sum of all fixations and saccades in an AOI during a single visit, which is the time span from entering an AOI until leaving the AOI. A new visit is counted when the AOI is revisited. These measures can be accumulated and expressed as total visit count or average visit duration per visit. The total reading time (the sum of fixation durations in a session) is described by TrtS on the English PTT, and TrtT on the German TT.

6 Results

In this section, the results are presented in several LMMs (see Table 2) which are later discussed in context. The results sections describe the three levels temporal effort (LMM-1), technical effort (LMM-2), and cognitive effort and visual attention (LMM-3 to LMM-7). Within each level only significant effects of condition and status or interaction with each other are discussed. An overview of all significant results per LMM is given in Table 3.

Contrasts in the interaction column are expressed as follows: before the colon is the condition/status of interest and after the colon follows the part that is contrasted

Table 3 Results per model with DVs and significant negative (−) or positive (+) effects for status and condition and their interactions with PE+V+S or professional subtitlers (P) as reference level

Model	DV	Status	Condition	Interaction contrasts
LMM-1	Duration	No	(−) PE+V	(−) PE+S:S
LMM-2	Keystrokes	(−) students	(−) PE+V	(−) PE+S:S
				(−) P:PE+V
LMM-3a	VisitDur PTT	No	(+) PE+S	(+) S:PE+S
LMM-3b	VisitDur TT	No	(−) PE+V	(+) PE+V:S
LMM-4a	VisitCount PTT	No	(−) PE+S	(−) S:PE+S
LMM-4b	VisitCount TT	No	(−) PE+V	No
LMM-5a	TrtS	No	(−) PE+S	No
LMM-5b	TrtT	No	(+) PE+V	No
LMM-6a	FactorVideo	(−) students	(+) PE+V	No
LMM-6b	RelativeVideo	No	(+) PE+V	No

either to the first condition PE+V+S, or status professional subtitlers, respectively. Overall, in these summative analyses, data points are comparably low in number per condition and task. Therefore, in future work, these results will have to be tested in larger setups with more repetitions or on smaller units of analysis such as segments to draw definite conclusions. The fact that significant effects are already found in such low numbers, however, seems promising.

6.1 Temporal Effort

The first variable of interest is the session duration, i.e. the time it took participants to complete the session. For a general overview on session durations across participants and conditions, the average durations per condition and status are listed in Table 4. After three outliers were excluded by a standard deviation of more than 2.5 times the median per condition, the average time spent per PE session was 9 minutes and 47 seconds. Differences per status varied only slightly and will be examined closer in the LMM-1. The shortest recorded session was a PE condition without video (PE+S) with 4 minutes and 10 seconds. The longest session was a monolingual PE condition with video only (PE+V) that took the participant over 14 minutes and 26 seconds with the longest sessions in the other two conditions not far behind.

For LMM-1, while there was no significant effect for status, there was a significant negative effect of the condition PE+V on the session duration compared to PE+V+S ($\beta = -0.2$, $SE = 0.06$, $df = 43$, $t = -4.4$, $p < 0.001$). This means that irrespective of the group, participants were faster during PE of the German TT when only the Swedish video was available but no English reference PTT. It was

Table 4 Overview of session durations per PE condition and status: professional subtitlers (P) and students (S) in minutes

Condition	Mean (P|S)	Min.	Max.	SD	N (P|S)
PE+S	09:45 (10:37|08:54)	04:10	14:13	02:13	22 (11|11)
PE+V	08:35 (08:52|08:17)	05:30	14:21	02:11	24 (12|12)
PE+V+S	11:00 (10:41|11:13)	07:21	14:26	02:00	25 (12|13)
Total	09:47 (10:03|09:32)	04:10	14:26	02:19	71 (35|36)

surprising that no significant effect was found for status as in condition PE+V the source context could only be inferred from the Swedish video. With professional subtitlers being more used to decoding polysemiotic channels, the assumption was that they would be more efficient regarding the completion time.

In the condition PE+V, since the video was in Swedish, participants had to work mainly with the German TT. It is possible that irrespective of the status, participants did not invest much time in checking the TT with the video. There may be context information in the visual part of the video and in the intonation of the speakers voice, but no verbal-linguistic content information could be taken from the spoken dialogue. This condition, hence, can be compared to monolingual PE which according to Nitzke (2019, 108) "can be very problematic due to the missing source, as content mistakes might remain unnoticed" despite having access to the AV content. Therefore, a closer look on the TT quality, outside the scope of this chapter, will have to show whether the saving in temporal effort also leads to differences in quality.

The finding that status did not have a significant effect in general on the temporal effort in the PE conditions reflects the findings of other studies where no significant differences in temporal effort between students and professional translators was found during PE, monolingual PE, and translation from scratch during the PE of newspaper texts (e.g. Nitzke 2019; Moorkens et al. 2015; De Almeida 2013). This might be due to a lack of experience in PE in both groups, or the attitude towards MT, which was not recorded in the present study.

With regard to an interaction between condition and status on the session duration as visualized in Fig. 2, it was surprising that the effect was significant and negative for students working without the video (PE+S) compared to professionals and condition PE+V+S ($\beta = -0.2$, $SE = 0.08$, $df = 43$, $t = -3.6$, $p < 0.05$). This indicates that, while professional subtitlers were only significantly faster in the condition without English PTT (PE+V) as they could not really check the content of the source dialogue, students were also significantly faster in the condition without Swedish video (PE+S). Condition PE+S can be compared to regular written PE which is closer to a student's written translation training. Being more used to drawing information from the AV content than translation students, professional subtitlers worked slightly slower when they were deprived of the AV content. In condition PE+S, as professional subtitlers were not able to check visual features or information on the scenes, they possibly heavily relied on the English reference script which slowed them down.

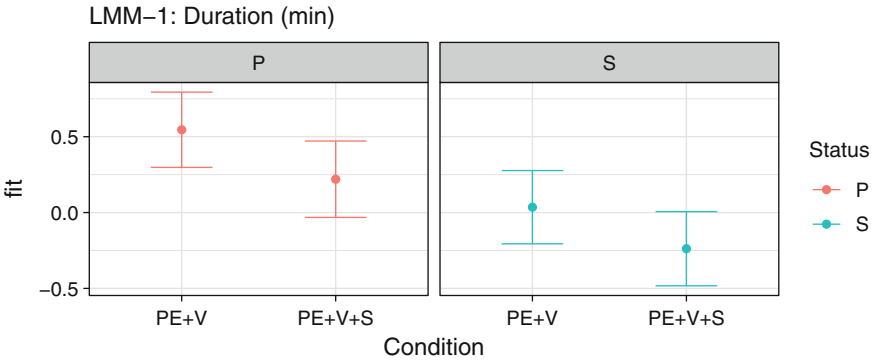


Fig. 2 Effects plot for LMM-1 showing the interaction effect between condition and status on the fitted session duration (y-axis)

Table 5 Descriptive statistic of keystrokes per PE condition and status: professional subtitlers (P) and students (S)

Condition	Mean (P|S)	Min.	Max.	SD	N (P|S)
PE+V+S	531 (594|472)	128	936	233	25 (12|13)
PE+V	356 (413|299)	104	851	185	24 (12|12)
PE+S	462 (597|327)	49	1065	229	22 (11|11)
Total	450 (533|370)	49	1065	226	71 (35|36)

6.2 Technical Effort

The technical effort in this study is defined by the amount of interaction participants showed in a session. This includes direct keystroke measures (insertions, deletions, and both counted together as keystrokes), but also measures such as units of continuous typing or overall typing duration. Descriptive statistics for keystrokes per condition and status are given in Table 5. Overall, on average 450 insertions and deletions were made irrespective of status and condition. Professional subtitlers performed far more keystrokes (533) than translation students (370), and on average the most keystrokes were recorded in session PE+V+S.

In LMM-2, as suggested by the descriptive statistics in Table 5, the number of keystrokes[7] is significantly affected by both condition and status. Similar to duration, the PE condition without English reference PTT (PE+V) has a negative effect ($\beta = -0.4$, $SE = 0.1$, $df = 43$, $t = -3.6$, $p < 0.01$) on the keystroke count compared to the condition with video and PTT (PE+V+S). Thus, participants were not only faster, but they also used fewer keystrokes during PE of the German TT assisted only by the Swedish video.

In contrast to the temporal effect, there was a significant effect observed for status, in that students inserted or deleted significantly fewer tokens than professional subtitlers in all conditions ($\beta = -0.3$, $SE = 0.1$, $df = 23$, $t = -2.6$,

[7] This DV was not log-transformed.

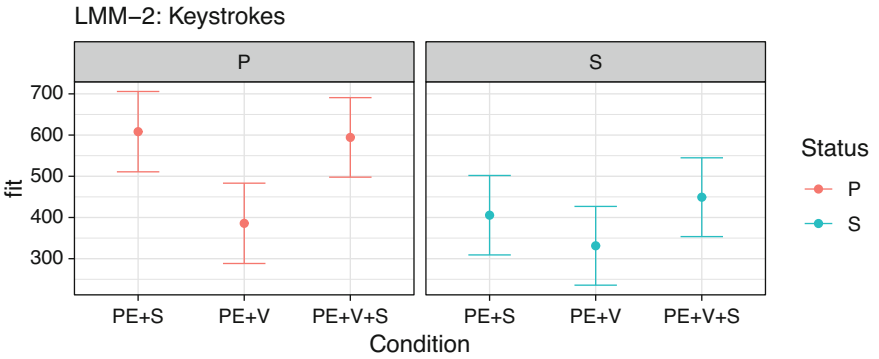


Fig. 3 Effects plot for LMM-2 showing the interaction effect between condition and status on keystrokes (y-axis)

$p < 0.05$). This indicates less technical effort, which does not automatically mean that they were more efficient as they were not significantly faster than professional subtitlers (see LMM-1).

The interaction effect of condition and status on keystrokes is visualized in Fig. 3. Students used significantly fewer keystrokes in the condition without video (PE+S) than professional subtitlers ($\beta = -0.5$, $SE = 0.2$, $df = 54$, $t = -3$, $p < 0.05$). While professionals made significantly fewer keystrokes in the video-only condition compared to working with both video and English reference PTT ($\beta = -0.5$, $SE = 0.1$, $df = 42$, $t = -3.5$, $p < 0.05$), this effect was not significant for students.

The results suggest that in general, fewer revisions were performed during monolingual PE of the German TT with only the Swedish video as reference (PE+V). This effect is particularly significant for professional subtitlers, as they cannot check the linguistic properties in the polysemiotic Swedish ST. For students this effect was less significant, as they used fewer keystrokes in general than the professional subtitlers. This could be attributed to students' lack of experience with this text type and lower expectations regarding quality control. For technical effort, insertions and deletions were statistically tested in separate models, demonstrating similar effects of condition and status on keystrokes combined.

6.3 Cognitive Effort and Visual Attention

In this study, cognitive effort during the PE tasks was measured with gaze data, as cognitive effort is linked to visual attention (e.g., Hvelplund 2011; Vieira 2014). More precisely, visual attention was measured in the average visit duration (LMM-3), but also total visit count (LMM-4) as well as total reading time on PTT (TrtS) and TT (TrtT) in LMM-5. In addition, visual attention directed to the video was measured and tested (LMM-6). For a description of the measures, see Sect. 5.3.

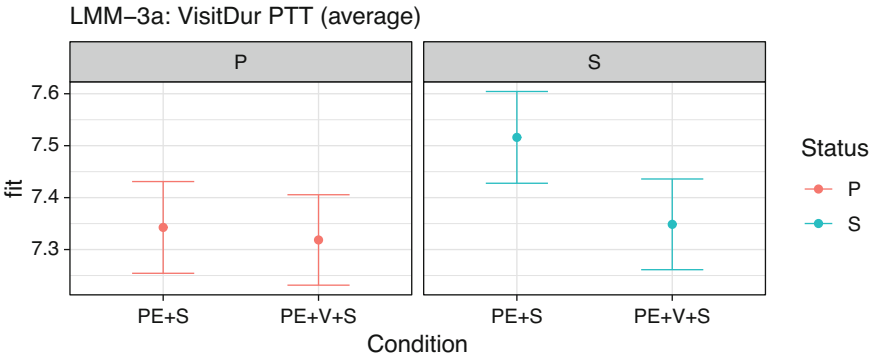


Fig. 4 Effects plot for LMM-3a showing the interaction effect of condition and status on average visit duration on English PTT (y-axis)

LMM-3 Average Visit Duration: The first DV is the average visit duration on the English PTT (LMM-3a), and on the German TT that was post-edited (LMM-3b).

In LMM-3a the condition with English PTT as reference (PE+S) had a significant positive effect on average visit duration on PTT (46 data points: $\beta = 0.09$, $SE = 0.04$, $df = 21$, $t = 2.3$, $p < 0.05$). This means that on average participants spent more time per visit in the PTT compared to when they had video and PTT as reference (PE+V+S).

As shown in Fig. 4, this effect interacted with status in a way that was only significant for students, indicating that for professional subtitlers the availability of the Swedish video did not have a significant impact on the average visit duration on the English PTT. The significant interaction effect for students suggests that the absence of the video led them to spend on average more time per visit in the reference PTT (46 data points: $\beta = 0.2$, $SE = 0.05$, $df = 19$, $t = 3.1$, $p < 0.05$).

In LMM-3b, the DV is average visit duration on TT (the German transcript that was being post-edited). Here, the condition PE+V had a highly significant positive effect for both groups (72 data points: $\beta = 0.4$, $SE = 0.06$, $df = 44$, $t = 6.3$, $p < 0.001$). This means that, on average, participants spent more time per visit in the TT when only the video was present compared to when they also had the English reference PTT available.

The effect of status was not significant, but there was an interaction as shown in Fig. 5. While for both participant groups the effect for the PE+V condition was positive and significant, it was larger and more significant for students (72 data points: $\beta = 0.7$, $SE = 0.07$, $df = 42$, $t = 8.4$, $p < 0.001$) than for professional subtitlers (72 data points: $\beta = 0.3$, $SE = 0.07$, $df = 41$, $t = 3.6$, $p < 0.01$). Translation students thus seem to adjust their attention allocation depending on the task, while for professional subtitlers this is not as evident.

LMM-4 Total Visit Count: The total number of visits per session on AOI PTT and TT was tested as DV in LMM-4a and LMM-4b with the following significant effects. In LMM-4a, the condition without access to the Swedish video (PE+S) had

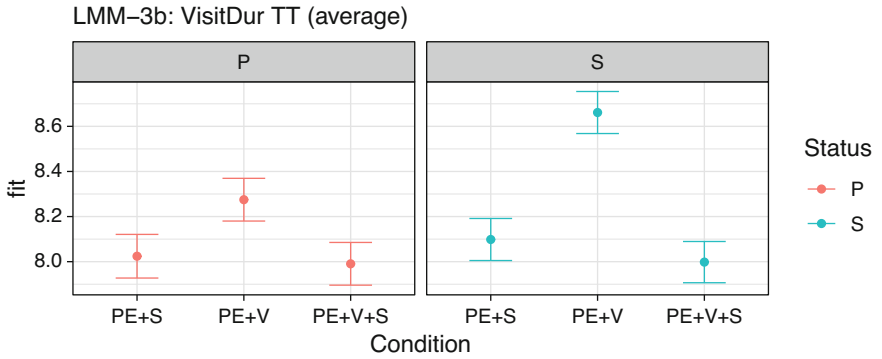


Fig. 5 Effects plot for LMM-3b showing the interaction effect of condition and status on fitted average visit duration on German TT (y-axis)

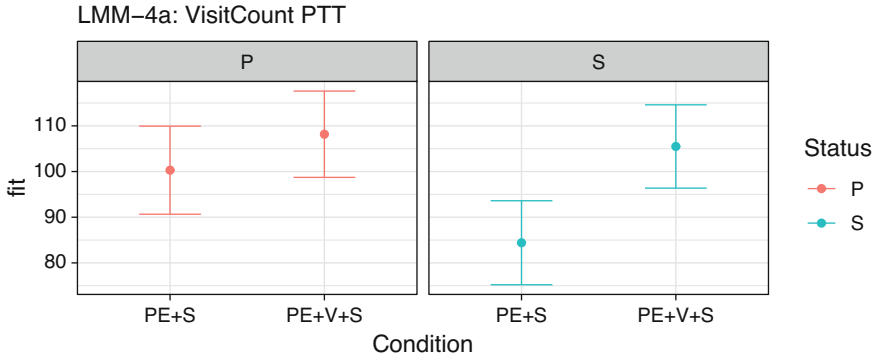


Fig. 6 Effects plot for LMM-4a showing the interaction effect of condition and status on total visit count on the English PTT

a significant and negative effect on the DV total visit count (not log-transformed) on English PTT (47 data points: $\beta = -16$, $SE = 4$, $df = 20$, $t = -4$, $p < 0.001$). Participants visited the English reference PTT less often when they had no video available, suggesting that the script was read more consistently irrespective of the experience.

Status had no significant effect and the interaction was only marginally significant as visualized in Fig. 6. The marginally significant interaction for students and condition PE+S without the Swedish video was negative (47 data points: $\beta = -21$, $SE = 4$, $df = 18.5$, $t = -4.8$, $p < 0.001$) indicating that students switched their attention less often away from the PTT than professional subtitlers when no video was available.

In LMM-4b, regarding moving the attention away from the MT output in TT, the effect of condition on VisitCount TT was significant and negative for condition PE+V (47 data points: $\beta = -0.3$, $SE = 0.8$, $df = 42$, $t = -3.9$, $p < 0.001$).

As there was neither an effect of status nor an interaction of status and condition for VisitCount TT, both participant groups moved their attention less often away from the TT during monolingual PE without the English reference PTT (PE+V) compared to the condition where they had both the English PTT and Swedish video as reference.

LMM-5 Total Reading Time: Effects on total reading time, both on the English reference PTT (TrtS) and on the TT with the German MT output (TrtT), were observed in LMM-5a, and LMM-5b respectively.

In LMM-5a, the total reading time on the English PTT was only marginally significant and shorter in condition PE+S, i.e. without access to the video ($\beta = -0.2, SE = 0.1, df = 21, t = -2, p < 0.07$). This suggests that when participants had the video as support, they also spent more time reading the English reference PTT, possibly checking the content against the video which they spent less time with, when no video was available. While this effect was to be expected, it was surprising that no significant effect or interaction was found for status on reading time on the English PTT.

In LMM-5b, the total reading time on the MT output to be post-edited in TT was positively and highly significantly affected in the condition PE+V without the PTT (72 data points: $\beta = 1.4, SE = 0.3, df = 44, t = 4.4, p < 0.001$). This suggests, as mentioned earlier, that during monolingual PE the focus is more on the TT and the Swedish video is much less consulted than when participants have an English reference script from which linguistic information can be taken to post-edit the German MT in the TT. This means that when no English reference PTT was available, both groups spent significantly more time reading the MT output. For TrtT, there was no significant effect for status and no interaction for the condition PE+S without the video.

LMM-6 Video Replay: In addition to reading and visit times on English PTT and German TT, it was investigated how much time participants spent replaying and consulting the video. This was tested in LMM-6a and LMM-6b.

The first DV in LMM-6a is the factor of video replay and expressed as the total time a video was replayed and fixated divided by the duration of the video. This factor is significantly and positively affected by the condition PE+V, i.e. without the English PTT (48 data points: $\beta = 0.3, SE = 0.1, df = 20, t = -2.8, p < 0.01$). As can be seen in Fig. 7, overall, the Swedish video was replayed and fixated longer when no English PTT was available compared to when both English PTT and Swedish video were available. This indicates that although participants were faster in condition PE+V (LMM-1), they spent more time consulting the video while performing fewer revisions as indicated by the fewer keystrokes (LMM-2).

For status the effect was negative and marginally significant for students (48 data points: $\beta = -0.5, SE = 0.3, df = 23, t = -1.8, p < 0.09$). This suggests that students spent less time replaying the video than professional subtitlers in both conditions. The effect might be linked to the experience of the professional subtitlers knowing that despite the unintelligible audio, there is still information to be found in the video that can assist in the otherwise monolingual PE task. Students not only

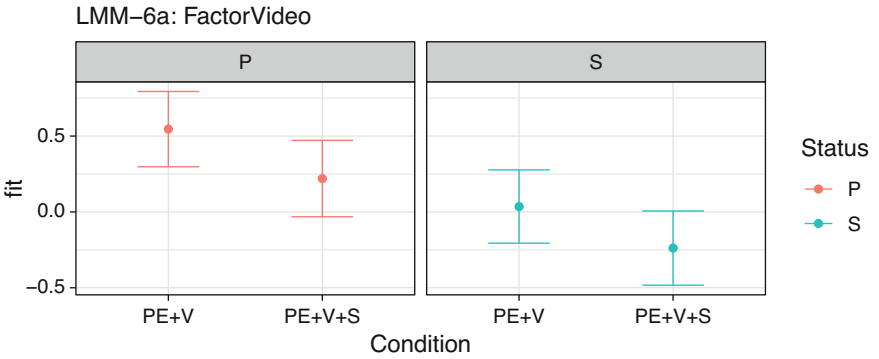


Fig. 7 Effects plot for LMM-6a showing the effect of condition video replay time as a factor of the video duration

performed fewer revisions, they also made less use of the video resource when it was available and rather relied more on the monolingual German TT.

In LMM-6b, the relative time fixating the video while replaying compared to the overall session time was tested. Condition PE+V again had a positive and significant effect on the relative video fixation time (48 data points: $\beta = 0.5$, $SE = 0.1$, $df = 21$, $t = 3.8$, $p < 0.001$), but there was no effect for status.

Thus, although the actual replay time differed in the two groups, this effect is not present, when the replay time is compared to the overall time participants spent with PE. This suggests that although professional subtitlers replayed the Swedish video more than students in the video-only condition, they did not necessarily take longer and generally worked more efficiently.

7 Discussion

This chapter presents a complex and innovative study design that applies an established methodology from TPR to the field of audiovisual translation. In the presented study, two concepts of modern subtitling—pivot subtitling and PE of NMT—are applied to subprocesses involved in interlingual subtitling: transcription and translation of audiovisual texts. In particular, the study attempts to answer four research questions with regard to reference materials in the indirect PE processes of Swedish video transcripts via English as pivot language touching on temporal (RQ1), technical (RQ2), and cognitive effort (RQ3) as proposed by Krings (2001) as well as subtitling experience (RQ4).

The results can be broadly summarized as follows. Regarding the different stages of support offered in the three PE conditions, it was found that the condition without an English reference transcript of the Swedish video had a significant and negative impact on several measures for both participant groups. Being able to draw reference

information only from the Swedish video in the basically monolingual PE condition (PE+S) caused participants to complete the task faster (RQ1) while also performing fewer insertions and deletions (RQ2).

Related to RQ2 and RQ3, participants spend less time typing in fewer production units and had a shorter average visit duration on the TT and fewer visits on TT compared to PE with both the video and English reference script, indicating that they post-edited in shorter chunks. The total reading time on the TT was longer and more time was spent on replaying and watching the video when just the video was available as reference. This does not seem surprising in a task similar to monolingual PE where the majority of attention is on reading the TT.

The condition without access to the Swedish video impacted the average visit duration on the English reference transcript in that participants mainly concentrated on comparing the two written representations similar to regular written translation. At the same time, during this condition, the English reference transcript was visited less often and overall less time was spent reading the English reference compared to when the Swedish video was also available. This suggests that the video was used to gather additional context information together with the English reference transcript, and attention was divided between video, English reference transcript, and German TT.

It is evident that having more resources necessarily requires additional cognitive resources of the participants switching between purely written and audiovisual content just like in the task of translating for dubbing. While this might lead to higher cognitive effort and more technical effort, it might still save time and reduce temporal effort while maintaining satisfactory quality. This is what will have to be investigated next with the recorded data.

Regarding experience (RQ4) and the three conditions with varying reference material, i.e. English pivot TT and/or Swedish video, some interesting findings were made when considering interactions. At first view, student translators and professional subtitlers do not show many differences in terms of the temporal, technical, and cognitive effort during the PE of German NMT transcripts from Swedish video extracts via English as pivot language. At least in terms of temporal effort, continuous typing, and reading of the English reference transcript and target transcripts, no main effects of status (translation students and professional subtitlers) were found.

The amount of support via reference material (Swedish video and English reference transcript), however, seems to impact the effort of translation students and professional subtitlers in different ways. With a closer look on effort regarding experience, the first presented results in this chapter show that students generally performed significantly fewer edits during PE and made less use of the video, which was to be expected as students have less experience with the translation or PE of spoken language and in the handling of audiovisual material. Despite the fact that the video audio was in Swedish—language participants did not understand—the video still contained relevant content and context information that can be useful especially in PE via a pivot language as in the case of the present study.

Results show that professional subtitlers generally performed more revisions and also made more use of the video in doing so. At the same time, this did not slow them down, as there was no significant difference in temporal effort between the two groups. Although final TT quality was not considered in this first coarse-grained analysis, it can be assumed that professional subtitlers worked more efficiently as they are used to this polysemiotic text type of fictional videos and possibly have higher TT quality expectations.

Students in particular seemed to perform not so differently in PE without the video, or with the video and English reference script. This is not surprising, as they are more used to translation of written text, but it seems as if students did not see much use in checking their post-edited translation with the context of the video which professional subtitlers obviously spend more time on, irrespective of whether they had an English reference script or not.

While professional subtitlers were overall no faster or slower than the subtitling students, they performed more edits and spent more time replaying the video, which can be expected to have a positive effect on TT quality. These findings may not tell us much about efficiency, but they give us a first idea of how translation students and professional subtitlers completed the tasks in the different conditions. A look at the TT quality and as well as the amount of research the translators performed online will shed more light on whether the time saving results in decreased quality. Within the scope of this chapter, the focus was only on effort and experience.

8 Conclusion and Outlook

In the next step, the data will be further analyzed regarding TT quality, but also regarding the specific allocation of cognitive resources similar to the studies performed by Hvelplund (2016, 2017). The differences in average visit duration on the English reference transcript between translation students and profession subtitlers might be linked to what Hvelplund (2016) suggests regarding flexibility of attention allocation. When faced with an additional video, the average visit duration is significantly lower for students than for professional subtitlers. Here, a qualitative analysis of the process might shed light on the strategies applied, e.g. switching between replaying the Swedish video and reading the English PTT compared to only switching between the English PTT and the German TT.

Interlingual transcription processes as an intermediate process of subtitling can be compared to translation for dubbing, and the methodology applied in this study makes it possible to analyze the process regarding attention allocation, but also more detailed analyses on the segment level will be possible. In addition to attention allocation, a qualitative analysis of the recorded processes is necessary to better understand the differences between the translation students and professional subtitlers regarding strategies such as online revision, which was also recorded, or phases of the process similar to orientation, drafting, and revision in written translation (e.g. Jakobsen 2002).

The aim of this chapter was to present the methodology and to present initial findings providing first insights on effort in AVT with NMT. The findings complement earlier findings in the project, namely PE in transcription processes—even if it is indirectly via a pivot English—is significantly faster than interlingual transcription processes from scratch or with the assistance of a correct English transcript (cf. Hansen-Schirra et al. 2020).

Just like several studies with technical texts and creative texts have shown in written translation (see Sect. 2), PE of NMT also has the potential to positively impact the production of interlingual transcripts that can serve as an additional and valuable support in subtitling. Although the present analysis has its limitations as it does not go beyond the level of entire session and does not look at the target quality, it provides first insights into the processes involved. The entirety of the analyses and more detailed results and discussions of the key-logging and eye tracking data also in reference to smaller units such as segments will be published in the author's dissertation (Tardel forthcoming).

Acknowledgments This project received funding by the EU and data was collected within the framework of the COMPASS project (CNECT 2017/3135124) from 2018–2019. The immense load of data processing would not have been possible without the assistance of my colleague Silke Gutermuth as she played a major role in helping to draw hundreds of AOIs.

References

Armstrong S, Caffrey C, Flanagan M (2006) Improving the quality of automated DVD subtitles via example-based machine translation. In: Proceedings of Translating and the Computer. Aslib, London

Artegiani I, Kapsaskis D (2014) Template files: asset or anathema? A qualitative analysis of the subtitles of *The Sopranos*. Perspectives 22(3):419–436. https://doi.org/10.1080/0907676X.2013.833642

Assis Rosa A, Pieta H, Bueno Maia R (2017) Theoretical, methodological and terminological issues regarding indirect translation: an overview. Transl Stud 10(2):113–132. https://doi.org/10.1080/14781700.2017.1285247

Baayen RH (2008) Analyzing linguistic data: a practical introduction to statistics using R. Processing 2(3):353. https://doi.org/10.1558/sols.v2i3.471. arXiv: 1011.1669v3

Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1). https://doi.org/10.18637/jss.v067.i01

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus Phrase-Based Machine Translation Quality: a Case Study. arXiv:160804631 [cs]

Beuchert K (2017) The web of subtitling: a subtitling process model based on a mixed methods study of the Danish subtitling industry and the subtitling processes of five Danish subtitlers. PhD Thesis, Department of Management, Aarhus BSS, Aarhus University

Bywood L, Georgakopoulou P, Etchegoyhen T (2017) Embracing the threat: machine translation as a solution for subtitling. Perspectives 25(3):492–508

Carl M (2012) Translog-II: a program for recording user activity data for empirical translation process research. In: The 8th International Conference on Language Resources and Evaluation, International Journal of Computational Linguistics and Applications, Istanbul, 1, vol 3, pp 153–162. https://www.ijcla.org/2012-1/153-162-paper.pdf

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research – exploring the CRITT TPR-DB. Springer, London, pp 13–56

Castilho S, Moorkens J, Gaspari F, Sennrich R, Sosoni V, Georgakopoulou P, Lohar P, Way A, Miceli Barone AV, Gialama M (2017) A comparative quality evaluation of PBSMT and NMT using professional translators. In: Kurohashi S, Fung P (eds) Proceedings of MT Summit XVI, vol 1. AAMT, Nagoya

Castilho S, Moorkens J, Gaspari F, Sennrich R, Way A, Georgakopoulou P (2018) Evaluating MT for massive open online courses: a multifaceted comparison between PBSMT and NMT systems. Mach Transl 32(3):255–278. https://doi.org/10.1007/s10590-018-9221-y

Díaz-Cintas J, Remael A (2014) Audiovisual translation: subtitling. Routledge, London

De Almeida G (2013) Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two romance languages. PhD thesis, Dublin City University

De Sousa SC, Aziz W, Specia L (2011) Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In: Proceedings of the international conference recent advances in natural language processing, 2011, pp 97–103

Doherty S, Kruger JL (2018) The development of eye tracking in empirical research on subtitling and captioning. In: Seeing into screens: eye tracking and the moving image. Bloomsbury Academic, New York, pp 46–64

Etchegoyhen T, Bywood L, Fishel M, Georgakopoulou P, Jiang J, Loenhout GV, Pozo AD, Maucec MS, Turner A, Volk M (2014) Machine translation for subtitling: a large-scale evaluation. In: Proceedings of the ninth international conference on language resources and evaluation (LREC 2014), pp 3–6

Flanagan M (2009) Using example-based machine translation to translate DVD subtitles. In: Proceedings of the 3rd international workshop on example-based machine translation, Dublin. Citeseer, pp 85–92

Georgakopoulou P (2019) Template files: the holy grail of subtitling. J Audiov Transl 2(2):137–160

Gottlieb H (1994) Subtitling: people translating people. In: Teaching translation and interpreting 2. John Benjamins, Amsterdam, p 261

Hansen-Schirra S, Tardel A, Gutermuth S, Schaeffer M, Denkel V, Hagmann-Schlatterbeck M (2020) Attention distribution and monitoring in the intralingual subtitling process. In: Proceedings of the 9th AIETI congress 2019: translatum nostrum. La Traducción y la Interpretación en el Ámbito Especializado, Comares

Hvelplund KT (2011) Allocation of cognitive resources in translation. An eye-tracking and key-logging study. Copenhagen Business School (CBS), Frederiksberg. 10.2011. Issue: 10

Hvelplund KT (2016) Cognitive efficiency in translation. In: Reembedding translation process research, pp 149–170. https://doi.org/10.1075/btl.128.08hve

Hvelplund KT (2017) Eye tracking and the process of dubbing translation. In: Díaz Cintas J, Nikoli K (eds) Fast-forwarding with audiovisual translation. Multilingual Matters, Bristol, Blue Ridge Summit, pp 110–124. https://doi.org/10.21832/9781783099375-010

Ivaska L (2020) A mixed-methods approach to indirect translation : a case study of the Finnish translations of modern Greek prose 1952–2004. PhD thesis, University of Turku, Turku. https://www.utupub.fi/handle/10024/150755

Jakobsen AL (2002) Translation drafting by professional translators and by translation students. In: Empirical translation studies: process and product, vol 27. Copenhagen studies in language, Samfunslitteratur, Frederiksberg, pp 191–204

Jakobsen AL, Mees IM, Mees IM (eds) (2009) Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen. In: Copenhagen studies in language, vol 38. Samfundslitteratur, Frederiksberg

Jia Y, Carl M, Wang X (2019) Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. Mach Transl 33(1–2):9–29. Springer

Just MA, Carpenter P (1976) Eye processing and cognitive processing. Cogn Psychol 8(4): 441–480

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull Math Linguist 108(1):121–132

Künzli A (2017) Die Untertitelung–von der Produktion zur Rezeption, vol 90. Frank & Timme GmbH

Kordoni V, Birch L, Buliga I, Cholakov K, Egg M, Gaspari F, Georgakopoulou Y, Gialama M, Hendrickx I, Jermol M, others (2016) TraMOOC (translation for massive open online courses): providing reliable MT for MOOCs. In: Annual conference of the European association for machine translation, p 396

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. Kent State University Press, Kent

Kuznetsova A, Brockhoff PB, Christensen RHB, Jensen SP (2019) lmerTest: tests in linear mixed effects models. https://CRAN.R-project.org/package=lmerTest

Liu CH, Silva CC, Wang L, Way A (2018) Pivot machine translation using Chinese as pivot language. In: CWMT 2018: proceedings of the 14th China workshop on machine translation. Springer, Wuyishan, pp 74–85

Matamala A, Romero-Fresco P, Daniluk L (2017) The use of respeaking for the transcription of non-fictional genres: an exploratory study. InTRAlinea: Online Transl J 19. http://www.intralinea.org/archive/article/2262

Matusov E, Wilken P, Georgakopoulou Y (2019) Customizing neural machine translation for subtitling. In: Proceedings of the fourth conference on machine translation (Volume 1: Research Papers). Association for Computational Linguistics, Florence, pp 82–93. https://doi.org/10.18653/v1/W19-5209

Melero M, Oliver A, Badia T (2006) Automatic multilingual subtitling in the eTITLE project. Proc Transl Comput 28:1–18

Moorkens J (2018) Eye tracking as a measure of cognitive effort for post-editing of machine translation. Eye Tracking Multidiscip Stud Transl 143:55

Moorkens J, O'brien S, Da Silva IA, de Lima Fonseca NB, Alves F (2015) Correlations of perceived post-editing effort with measurements of actual effort. Mach Transl 29(3–4):267–284

Moorkens J, Toral A, Castilho S, Way A (2018) Translators' perceptions of literary post-editing using statistical and neural machine translation. Transl Spaces 7(2):240–262. https://doi.org/10.1075/ts.18014.moo

Nikolić K (2015) The pros and cons of using templates in subtitling. In: Piñero RB, Cintas JD (eds) Audiovisual translation in a global context. Palgrave Macmillan, London, pp 192–202. https://doi.org/10.1057/9781137552891_11

Nitzke J (2019) Problem solving activities in post-editing and translation from scratch. Language Science Press, Berlin, oCLC: 1147275878

Orrego-Carmona D, Dutka \, Szarkowska A (2018) Using translation process research to explore the creation of subtitles: an eye-tracking study comparing professional and trainee subtitlers. JoSTrans: J Spec Transl 30:150–180

Piperidis S, Demiros I, Prokopidis P (2004) Multimodal multilingual information processing for automatic subtitle generation: resources, methods and system architecture. In: Proceedings of the workshop on e-tools and translation of the 5th international languages and the media conference and exhibition, Berlin, pp 5–8

Silvestre Cerdà JA, Del Agua Teba MA, Garcés Díaz-Munío GV, Gascó Mora G, Giménez Pastor A, Martínez-Villaronga AA, Pérez González de Martos AM, Sánchez-Cortina I, Serrano Martínez-Santos N, Spencer RN, others (2012) Translectures. In: IberSPEECH 2012-VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012, pp 345–351

Tardel A (2020) Effort in semi-automatized subtitling processes: speech recognition and experience during transcription. Journal of Audiovisual Translation 3(2). https://doi.org/10.47476/jat.v3i2.2020.131

Tardel A (forthcoming) Automatization in subtitling processes. PhD Thesis, English Linguistics and Translation Studies, Germernsheim, University of Mainz

Tardel A, Hansen-Schirra S, Schaeffer M, Gutermuth S, Denkel V, Hagmann-Schlatterbeck M (2020, forthcoming) Attention distribution and monitoring in the intralingual subtitling process. In: Proceedings of ICTIC 2019

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 1, Long Papers. Association for Computational Linguistics, Valencia, pp 1063–1073. https://www.aclweb.org/anthology/E17-1100

Toral A, Wieling M, Way A (2018) Post-editing effort of a novel with statistical and neural machine translation. Front Digit Hum 5:9

Toury G (1995/2012) Descriptive translation studies – and beyond, Benjamins translation library, vol 100, 2nd edn. John Benjamins Publishing, Amsterdam/Philadelphia

Valor Miró JD, Spencer RN, Pérez González de Martos A, Garcés Díaz-Munío G, Turró C, Civera J, Juan A (2014) Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. Open Learn: J Open Dist e-Learn 29(1):72–85

Vermeulen A (2011) The impact of pivot translation on the quality of subtitling. Int J Transl 23(2):119–134

Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Mach Transl 28(3–4):187–216. https://doi.org/10.1007/s10590-014-9156-x

Volk M, Sennrich R, Hardmeier C, Tidström F (2010) Machine translation of tv subtitles for large scale production. In: JEC 2010, 4 Nov 2010, Denver. Association for Machine Translation in the Americas, pp 53–62

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York. https://ggplot2.tidyverse.org

Part II
Translation and Entropy

Information and Entropy Measures of Rendered Literal Translation

Michael Carl

Abstract The definition and effect of translation literality (word-for-word translation, as opposed to sense-for-sense translation) have been a topic in translation studies for very many years. Schaeffer and Carl (2014) introduced measures to quantify the literality of observed translations based on their distance to a hypothetical absolutely literal translation assuming monotonicity (identical word order), compositionality (one-to-one translation), and entrenchment (lack of translation variation). This chapter introduces measures to assess the *rendered translation literality*, which reformulates the three literality criteria in an information-theoretic framework. We introduce rendered literality measures on the word and segment levels as the *joint self-information* and the *joint entropy* of the three literality criteria. These literality measures respect the additional requirement for translations to be grammatical according to target language (TL) constraints (i.e., rendered) and rank sets of alternative translations according to their rendered literality. We evaluate the predictive potential of rendered translation literality on translation duration within the multiLing dataset for different languages and different translation production modes. Our data show that the most probable translations respecting TL constraints are also most monotone, compositional, and entrenched translations. We conclude that the information-theoretic literality measures are powerful new predictors for behavioral measures, such as word and segment translation duration.

Keywords Literal translation · Translation entropy · Translation self-information

1 Introduction

The term *literal translation* has been used in many ways but its interest in translation studies is related to statements such as "translators tend to proceed from more literal versions to less literal ones" (Chesterman 2011, 26). To assess such statements, it is

M. Carl (✉)
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_5

necessary to develop literality measures. Halverson (2015) points to some confusion with respect to the usage and meaning of *literal translation*. She discusses the history of the term in some detail and shows that it has sometimes been used to refer to qualities of the translation product and sometimes to describe strategies during the translation process. In an attempt to disentangle its meaning, Halverson (2015, 2019) makes a distinction between "literal translation" and "default translation" and refers to literal translation as "particular patterns of interlingual or intertextual correspondence," while "default translation [is] a particularly unconstrained and immediate production mode" (Halverson 2015, 320). Literal and default translation are related in some ways. It has been claimed that the first translational response is more literal than the revised versions of it (see also Carl this volume-a, Chap. 9) and that literal translations are easier and faster to produce, but with a clear definition of what *literal translation* might mean, those claims are difficult to quantify and validate. In line with Halverson's distinction, this chapter introduces several measures to quantify translation literality as a property of the relation between the source and the target language systems (*langue*) and the translation product performance (*parole*) using an information-theoretic framework. Chapter 9 takes up the discussion on default translation and relates it to the literality measures introduced here.

Schaeffer and Carl (2014) are the first—to the best of our knowledge—to provide a formal definition of *translation literality* that relies on cross-linguistic similarity of the semantic and syntactic relations in the translation product. They suggest measures to quantify the two components of translation literality (i.e., cross-lingual semantic and syntactic similarity) and have shown the relevance of these measures in numerous studies. The *word translation entropy* (*HTra*;[1] Carl and Schaeffer 2014) is taken to index cross-linguistic semantic similarity, and Heilmann and Llorca-Bofí (this volume, Chap. 8) identify *HTra* as a measure of translation selection pressure, where greater cross-linguistic semantic similarity exerts higher translational selection pressure. Higher translational selection pressure results in more entrenched translation solutions with fewer alternative translations. HTra has been shown to be predictive for many behavioral observations, including translation duration and gaze behavior, and correlates significantly across different languages and translation modes (e.g., Ogawa et al. this volume, Chap. 6). The syntactic similarity is measured in terms of relative cross-lingual word reordering (i.e., distortion; Brown et al. 1993) of translation equivalent expressions. The *word distortion entropy* (HCross, i.e., word alignment crossing) has been shown to strongly correlate with word translation entropy (e.g., Carl et al. 2019) indicating that translations involving more heavy reordering of the translated segments also show weaker cross-linguistic semantic similarity as compared to segments that are translated in a more monotonous fashion adhering to the source language (SL) word order.

[1] "H" is often used to refer to entropy.

These entropy measures indicate to what extent the source language (SL) and target language (TL) *systems* are related, i.e., it measures the possible "heterogeneity of the translation solutions of a source text (ST) token" (Heilmann and Llorca-Bofí this volume, Chap. 8). *HTra* and *HCross* are properties of ST tokens; more heterogeneous alternative translations lead to higher entropy values indicating translational discrepancy in the SL and TL systems, while less translational variation is an indicator of higher cross-linguistic similarity. *HTra* quantifies whether an expression occupies "as nearly as possible the 'same' place in the 'economy' of the TL as the given SL category occupies in the SL" (Catford 1965, 26) or whether "structural and semantic shifts [are required] which destroy formal correspondence altogether" (Ivir 1981, 58). In this latter case, we are likely to see more translation variation, higher heterogeneity of translation solutions, and thus higher translation entropy values. The word translation entropy and word distortion entropy as conceived in Schaeffer and Carl (2014) thus provide system-based translation heterogeneity measures quantifying the possibilities for different translation renderings. These measures may be asymmetric, and since we look from the source to the target, these entropy values are ST specific.

However, it might be interesting to assess whether one (observed) translation is more literal than another alternative, a distinction that cannot be obtained with entropy values. Heilmann and Llorca-Bofí (this volume, Chap. 8) therefore suggest measuring the *word translation (self-)information*, which quantifies the literality of individual translations "within the possibilities of the TL." The self-information allows different translations to be ranked according to their literality. We thus introduce the *word translation information* and the *word distortion information*[2] which are, just as the entropy values, indicative of cross-linguistic semantic and syntactic relatedness, respectively, but for each translation individually, and are thus properties of the target text (TT).

In addition, we introduce rendered literality measures of alternative translations on a token level as well as a segment level. On the token level, the *joint ST-TT-alignment crossing information* (ISTC) measures a literality score for each target word, while the *joint ST-TT-alignment crossing entropy* (HSTC) measures a literality score for each source word. ISTC indicates how entrenched a given translation is, while HSTC indicates how heterogeneous a set of translation is expected to be. The *ISTC* measure is also used on the segment level to compute the average rendered literality score for each translation of a source segment. Another total literality score (*HTot*) measures the expected translation heterogeneity for the source segment. First, we lay out the basic assumptions underlying these literality measures in Sect. 2. Then we present the actual implementation of the measures on the word level in Sect. 3. We present some examples to illustrate the components of the measures on a word level in Sect. 4. In Sect. 5, we develop translation information and entropy measure on a segment level and show how it predicts translation behavior.

[2]The distortion is measured in terms of alignment crossings for which we compute *Cross* values.

2 Rendered Literal Translation

The similarity of cross-linguistic semantics and syntax in translation has often been referred to as *literal translation*. Chesterman (2011, 30) cites Dimitrova (2005, 53) suggesting literal translation to be "a TT fragment which is structurally and semantically modeled upon the ST fragment while respecting the TL grammatical constraints." Schaeffer and Carl (2014), Schaeffer et al. (2016), and Carl and Schaeffer (2017b, c) formalize this definition with three criteria:

1. Word order is identical in the ST and TT.
2. ST and TT items correspond one-to-one.
3. Each ST word has only one possible translated form in a given context.

Carl and Schaffer (2017c) call a translation *absolutely literal*[3] if it fulfills all three criteria, i.e., if all ST and TT words can be one-to-one aligned, the source and the target are produced in the same word order, and there is only one (dominant) translation in the given context for each alignment group. An absolutely literal translation, under this definition, equals a deterministic substitution of lexical items; there is no knowledge added during the translation process, as all translational choices are determined by the ST (and a deterministic transfer lexicon). However, absolutely literal translations may not (or only in very rare cases) respect TL grammatical requirements. A "rendered" literal translation obeying the TL constraints will thus violate one or more literality criteria. While all three literality criteria are, in theory, independent, in practice, there is a significant correlation between variation in cross-linguistic word-order, segmentation, and lexical choice. It is therefore likely that all three literality criteria co-vary to some extent, but we suspect that the most frequent translational choices will—in general—be those that reflect most closely the semantic and syntactic properties of the SL while at the same time abiding by and rendering the possibilities of the TL.

Literality criterion 2 is grounded in the compositionality assumption of translation, i.e., whether words can be mapped from the source to the target language and reproduce meanings of complex expression in a similar segmentation for which the source language allows. A *principle of translation compositionality* has been spelled out in Rosetta (1994, 17), according to which "Two expressions are each other's translations if they are built up from parts which are each other's translation, by means of translation-equivalent rules." In line with the definition, we can say that a translation is absolutely compositional if a one-to-one translation-equivalence can be established. Translation compositionality should be distinguished from monolingual compositionality. Pustejovsky (2012) gives examples of monolingual co-composition and co-specification. While, for instance, in *Mary waxed the car clean*, the meaning of the predicate (*clean*) changes the meaning of the verb (*waxed*), an absolute compositional translation, for instance, into German as *Mary*

[3]Carl and Schaeffer (2014) previously referred to this as *ideal literal translation*.

wachst das Auto sauber may be possible where each word can be aligned in a one-to-one fashion.[4] Pustejovsky (2012) gives other examples of co-composition, including adjectives which may change their meaning "depending on the word they characterize: the color of 'white wine' is actually yellow-ish" (ibid.), or similarly, Portuguese "vinho verde" (*green wine*) is actually young, typically white wine which originates in the green northern part of Portugal. Co-composition may be differently realized in different languages, for instance, "black tea" which is realized as "red tea" in Chinese but is actually brown-ish. Similarly, magnifiers such as "heavy rain" and "bad accident" are, respectively, realized as "strong rain" (*starker Regen*) and "heavy accident" (*schwerer Unfall*) in German. According to the "principle of translation compositionality," we could presume a translation-equivalent rule attached to the noun "tea" which states that the modifier for dark-colored tee is realized as "black" in English and "red" in Chinese and that "rain" may be *heavy* in English and the equivalent of *strong* in German, etc. This opens the possibility to align *black* with *red* in the context of *tea*, and *strong* with *heavy* in the context of *rain*. If we do not assume such translation-equivalent rules, we might need to assume that *black tea* and *strong rain* are compound expressions that need to be translated non-compositionally.[5] Pustejovsky (2012) maintains there are many co-compositional constructions in languages. We hold that, as long as a complex meaning can be compositionality reproduced in the target language, a maximum compositional fragmentation of the translation and an arbitrary number of translation-equivalent rules may be most appropriate to assess the compositionality and literality of translations. While it may be difficult in some cases to decide exactly which translation-equivalent rules to presume and to decide the boundaries of basic translation equivalent expressions, once we know the fragmentation of a translation into compositional alignment groups, we can count the number of source and target items in the alignment group to quantify the literality criterion (2). As we will show below (Table 2), most frequently, alignment groups consist of two elements, one token in the source and one token in the target—between 22% for Japanese and 60% for Spanish—making one-to-one translation alignments the most likely correspondence across languages.

Criterion 1 measures word order differences (monotonicity) of the alignment groups in the source and target texts. Various ways of measuring syntactic equivalence have been suggested, some of which are addressed in detail in Vanroy et al. (this volume, Chap. 10). In the CRITT TPR-DB,[6] we model changes in word order (i.e., reordering of alignment groups) by means of local distortion. We measure local distortion by means of a *Cross* measure which counts the number of words on the target side between the ending of two successive alignment groups. Note

[4]However, there may be better and more idiosyncratic translations.

[5]A similar argument can be made with terms such as "White House" (in Washington) or idioms (such as "show red card") which translates literally into many languages and can thus be aligned word-by-word.

[6]CRITT TPR DB https://sites.google.com/site/centretranslationinnovation/tpr-db

that this measure includes two factors: the number of tokens on the target side of the alignment group and a possible shift of the translation to the right or left. For instance, assume *black tea* was considered a non-compositional translation and thus consists of two TL words. In addition, assume *black tea* was shifted two words to the right, as in a translation from *Black tea I like* to *I like black tea*, for instance. The translation distortion (its associated *Cross* value) would amount to 4, counting the number of words between the end of the preceding translation (i.e., the beginning of the segment) and the end of the *black tea* translation. If the segments were translated the other way around (from *I like black tea* to *Black tea I like*), the *Cross* value for *black tea* would be −4. A monotonous translation may consist of sequences of phrasal alignment groups where all *Cross* values ≥1. A translation with identical ST and TT word order, such as in *Mary waxed the car clean* ←→ *Mary wachst das Auto sauber* where all *Cross* values amount to 1 is monotone and compositional. As shown in Table 3, a *Cross* value of 1 is indeed the most frequent observation—between 12% for Japanese and 71% for Danish—making segments of compositional, monotonous translation the most likely translational correspondence across languages.[7] Note that the *Cross* measure accounts for distortion on the target side, and it does consider the possible variation of ST words in an alignment group.

Literality criterion 3 assesses to what extent the available translation solutions are entrenched, i.e., it measures how much an ST expression exerts a selection pressure for the production of a particular translation solution. Tokowicz et al. (2002, 442) showed that "the more translations a word has, the lower the semantic similarity of the translation pair." This suggests that, in the extreme case, if there was an exactly meaning-identical token in the TL, we would expect no translational variation, as all translators will probably produce that same translation—the translation would thus be absolutely entrenched. If, however, the TL only provides solutions with partially or sparsely overlapping meanings, we are likely to see more different and less entrenched translations. Schaeffer and Carl (2014) and Schaeffer et al. (2016) quantify literality criteria 3 as *word translation entropy* (*HTra*), which measures observed translational choices, to indicate cross-linguistic similarity of meaning in alignment groups. The current implementation of this criterion in the TPR-DB assesses this property for each ST token and has proven to be a powerful predictor for various behavioral measures, as reported in several chapters in this volume.

Note, however, that literality criteria 1–3 do not address Chesterman's requirement for a literal translation to respect "the TL grammatical constraints." Criteria 1 and 2 potentially measure to what extent a TT fragment is structurally modeled upon the ST (segmentation and word-order) and thus provide measures for syntactic translation difficulty (see Varnroy et al. this volume, Chap. 10). Literality criterion 3 provides an indicator of the cross-linguistic semantic relatedness of an ST expression and thus how difficult it might be to find a meaning-equivalent translation. *HTra*

[7] In Table 3, we subtract the length of the target group from the *Cross* value, so as to count only the shift to the left or to the right.

is an information-theoretic measure that quantifies the variation in a set of alternative translations and accounts for criterion 3 under the assumption that more entrenched translations also represent a more literal meaning equivalence. In this chapter, we extend this information-theoretic framework to include all three literality criteria. As with the word translation entropy (criterion 3), which indexes the semantic heterogeneity of TL, we assume that a word order entropy (criterion 1) and a segmentation entropy (criterion 2) are indicators of syntactic heterogeneity. We compute probabilities for all three indicators individually and the joint probability for all three literality criteria and compute the joint self-information and joint entropy.

We extend the information-theoretic view and include all three literality criteria as joint probabilities in the following way. Given a source word (e.g., *black*) and a set of alternative translations, the joint probabilities include (1) the probability of the source group, e.g., the probability of *black* to be grouped with *tea*; (2) the probability of the translation group, e.g., the probability of the Chinese translation equivalents for *red* or *red tea* or *tea* or still something different; and (3) the probability of a relative shift of the translation in the target segment. We develop translation literality measures based on the joint probability and compute the *joint ST-TT-alignment crossing information* (*ISTC*) and the *joint ST-TT-alignment crossing entropy* (*HSTC)*. As a prerequisite, we also introduce the *word translation information* (*ITra*) and the *word distortion information* (*ICross*).

3 Translation Literality Measures

Each source token (w) is associated with a set of alignments that consists of a group of source tokens (s), a group of target tokens (t), and a corresponding distortion index (c). The set $A : \{\{s_1, t_1, c_1, \}, \{s_2, t_2, c_2\}, \{s_3, t_3, c_3\} \ldots \{s_n, t_n, c_n\}\}$ of word alignment and distortion parameters represents n alternative translations for w in the given context (cf. Sect. 4). The *word translation entropy* (Carl and Schaeffer 2014, Carl et al. 2016, Carl and Schaeffer 2017a, b, c) quantifies the variation of observed translations $t_k \in A$ in a corpus of alternative translations (A). *HTra* computes for a source token w, the sum over the product of the translation expectation $p(t_k|w)$, and the word translation information (*ITra*) as shown in Eqs. (1) and (2):

$$ITra\,(w, t) = -\log_2\,(p\,(t \mid w)) \tag{1}$$

$$HTra\,(w, A) = \sum\nolimits_{t_k \in A} p\,(t_k|w) \times ITra\,(w, t_k) \tag{2}$$

Source text words with low *HTra* values have only a few target-language equivalents, which represent entrenched translation solutions and which make the translation relatively predictable. These translations are likely to have stronger interlingual connections. Higher *HTra* values imply a larger set of possible trans-

lations from which a translator can choose. The *word translation information* (*ITra*) indicates the likelihood of a translation.

$$ICross\,(w, c) = -\log_2\,(p\,(c \mid w)) \tag{3}$$

$$HCross\,(w, A) = \sum\nolimits_{c_k \in A} p\,(c_k \mid w) \times ICross\,(w, c_k) \tag{4}$$

Each of the n translations of source token w is associated with *Cross* a value $c_k \in A$ indicating the local distortion of that translation in the target text. Analogous to *HTra*, the *word distortion entropy*, *HCross*, in Eq. (4) quantifies the variation of local distortion between successive alignment groups. If *HTra* indexes translations possibilities for the source w, *HCross* indexes word position (reordering) possibilities in the target text from which a translator chooses. *HCross* thus represents the possibilities for word-order similarity of the source and the target. As with the word translation information (*ITra*), we also compute a word distortion information (*ICross*) in Eq. (3), which measures the self-information of observed word distortion.

$$ISTC\,(w, s, t, c) = -\log_2\,(p\,(s, t, c \mid w)) \tag{5}$$

We introduce two literality measures: the *ISTC*, Eq. (5), and the *HSTC*, Eq. (6). As discussed above, *ISTC* measures the self-information of a particular translation given a set of alternatives, i.e., it indicates to what extent a specific translation is semantically and syntactically similar to the source as compared to the alternative translations. Translation literality and information are reverse proportional, as low joint probabilities have a high amount of information.

$$HSTC\,(w, A) = \sum\nolimits_{\{s_k, t_k, c_k\} \in A} p\,(s_k, t_k, c_k \mid w) \times ISTC\,(w, s_k, t_k, c_k) \tag{6}$$

The *HSTC* measure indicates the degree to which a source token allows for literal translations, i.e., it measures the variety of translation solutions in a given context. It is the sum over all alternative translations taking into account the translation expectation and its self-information.

4 Literal Word Translation across Languages

Table 1 shows an excerpt from the multiLing dataset.[8] It shows a summary table of 152 translation alignment groups for "Yesterday" (word Nr. 23 of Text 1) in the context of the sentence "Yesterday**,** he was found guilty of four counts of murder

[8]See Appendix 1 for a description of the dataset used in this and several other studies in this volume.

Table 1 Translations of "Yesterday" with their probabilities, self-information, and entropy literality measures into six languages

TL	TGroup	ProbT	ITra	HTra	Cross	ProbC	ICross	HCross	ISTC	HSTC
da	i går	0.792	0.337	0.738	2	0.625	0.678	1.33	0.678	1.33
					4	0.167	2.580		2.580	
	igår	0.208	2.260		1	0.208	2.260		2.260	
de	Gestern	1.000	0.000	0	1	0.773	0.372	1.08	0.372	1.08
					3	0.046	4.460		4.460	
					7	0.136	2.870		2.870	
					13	0.046	4.460		4.460	
es	Ayer	0.903	0.147	0.612	1	0.839	0.254	1.01	0.254	1.01
					3	0.032	4.950		4.950	
					7	0.032	4.950		4.950	
	Ayer fue hallado	0.032	4.950		4	0.032	4.950		4.950	
	después	0.032	4.950		2	0.032	4.950		4.950	
	En el día de Ayer	0.032	4.950		5	0.032	4.950		4.950	
hi	–	0.133	2.910	0.906	0	0.133	2.910	2.87	2.910	2.87
	कल	0.800	0.322		1	0.333	1.590		1.590	
					3	0.133	2.910		2.910	
					4	0.067	3.910		3.910	
					5	0.067	3.910		3.910	
					7	0.067	3.910		3.910	
					8	0.067	3.910		3.910	
					15	0.067	3.910		3.910	
		0.067	3.910		17	0.067	3.910		3.910	
ja	、	0.026	5.290	0.343	5	0.026	5.290	1.91	5.290	2
	昨日	0.949	0.076		1	0.641	0.642		0.642	
					4	0.026	5.290		5.290	
					6	0.026	5.290		5.290	
					7	0.051	4.280		4.280	
					8	0.051	4.280		4.280	
					9	0.128	2.960		3.280	
					10	0.026	5.290		5.290	
					11	0.026	5.290		5.290	
	昨日_付_で	0.026	5.290		9	0.128	2.960		5.290	
zh	–	0.048	4.390	1.61	0	0.048	4.390	1.41	4.390	2.16
	于_昨天	0.048	4.390		8	0.095	3.390		4.390	
	昨天	0.571	0.807		1	0.714	0.485		1.070	
					6	0.095	3.390		3.390	
	昨天_,	0.048	4.390		2	0.048	4.390		4.390	
	昨日	0.286	1.810		1	0.714	0.485		2.070	
					8	0.095	3.390		4.390	

following a long trial" into the six target languages, Danish (24), German (22), Spanish (31), Hindi (15), Japanese (39), and Chinese (21).[9] The word *yesterday* has a relatively low *HTra* value with strong selection pressure across the six languages. The first columns in the Table show the target language and the realized translation (*TGroup*), the translation probability (*ProbT*), translation information (*ITra*), and translation entropy (*HTra*). It shows that, for instance, all 22 German translators produced the same German translation *Gestern*, for *yesterday*. The corresponding translation probability is thus $ProbT = 1$, while *ITra* and *HTra* values are 0. An *ITra* and *HTra* value of 0 represents a deterministic translation in which all translators produce the same target token, word, or phrase. Two different translation versions were produced for Danish; three for Japanese, Spanish, and Hindi; and five for Chinese. The two Danish translations *i går* and *igår* are two valid spelling variations which have two different translation probabilities (0.792 and 0.208, respectively), and thus each of the translations has also a different translation self-information, while the word translation entropy ($HTra = 0.738$) relates to the entire set of observed translations for the source expression and thus indicates its translation heterogeneity.

A similar translation self-information (*ITra*) in alignment groups with different *HTra* values may represent different translational processes. For instance, while the Chinese *HTra* value (1.61) is substantially higher than for Danish, some of the translations may have similar or even lower *ITra* values. Choosing a translation with (say) probability $ProbT = 0.2$ out of a set with two alternative solutions may be quite different from choosing a translation with the same probability out of a set with, for instance, eight different options. It might be worth investigating whether the ratio of *HTra* and *ITra* values may shed additional light on translational processes. However, instances with very low probabilities (i.e., high *ITra*) are likely to represent alignment or translation errors, such as some of the Spanish and Japanese examples in Table 1.

The "Cross" field indicates the *Cross* value that is associated with the alignment group and indicates the relative distortion of the translation. In many cases, there are several different *Cross* values for one translational solution attesting different reordering choices. For instance, the Spanish translation "ayer" was shifted 1, 3, and 7 positions in the target, as can be verified in Appendix 2, but the most probable translation ($ProbC = 0.839$, $ICross = 0.254$) is the one close to the ST word order with $Cross = 1$. Hindi and Japanese provide extreme cases in which the same translation can be shifted to many different positions in the target, presumably due to freer word order than the other languages allow. However, also in these cases, the more literal solution (choosing the translation position with lower distortion distances) is more likely than translation solutions with larger reordering crossings. Given that, in this example, there is more freedom in positioning the translation in the target segment than in the actual lexical choice. *HCross* values are higher than *HTra* values.

[9]The number of alternative translations is in parenthesis.

In most cases, the source token "Yesterday" was in a single one-to-one alignment group with Spanish translation equivalent. However, in some cases, it was also grouped with the following comma "Yesterday," and in two instances, the ST group consists of six tokens, "Yesterday, he was of found"[10] and one occurrence of the discontinuous group "Yesterday, following'" (shaded in grey in Table 1). These different source groups lead to differences in the values for *ICross* and *ISTC* as well as *HCross* and *HSTC* as for Japanese and Chinese. As the joint probability is always less or equal to each of the individual probabilities, and since in most cases *ProbC* is smaller than or equal to *ProbT*, and only in the cases where different ST tokens are the same TT group, there are differences between *ICross* and *ISTC* and between *HCross* and *HSTC*.

4.1 Size of Alignment Groups

Table 2 shows the probabilities of the number of tokens in alignment groups (AGs) for the six languages in the multiLing dataset. An AG with a size of $AGnbr = 1$ indicates that the source token was not aligned (it only consists of the source token), and $AGnbr = 2$ indicates that the source and the target consist each of one token. For $AGnbr > 2$, the tokens are distributed in some way over the source and the target side of the alignment group, summing up to the total. For instance, the word *black* might be associated with an alignment group *black tea* $\longleftrightarrow$ *red tea* which amounts to $AGnbr = 4$, or *black* $\longleftrightarrow$ *red* which amounts to $AGnbr = 2$. Table 2 shows the percentage of occurrences with *AGnbr* 1–9. It shows that for all languages, a compositional one-to-one alignment is the most probable, with around 60% of the cases for Spanish (es) and German (de) and 57% for Danish (da). One-to-one compositional translation is also the most likely translation strategy for Chinese (zh), Hindi (hi), and Japanese (ja), although to a lesser extent. Notice that for Japanese, the tail is quite long with almost 1% of instances with over 20 tokens per AG, which may be due to the mecab[11] tokenizer used.

Table 2 Size of alignment groups per language as percentages

AGnbr	1	2	3	4	5	6	7	8	9
da	1.47	56.78	21.56	8.25	5.61	2.89	1.17	0.94	0.59
de	3.04	59.74	18.82	8.27	3.81	1.51	1.37	1	0.5
es	0.42	61.38	19.01	7.61	3.76	3.29	1.83	0.81	0.93
hi	3.26	31.77	15.31	11.49	11.78	8.12	5.83	3.48	2.35
ja	0.89	22.02	16.92	14.84	14.17	8.79	7.3	4.55	3.16
zh	2.68	36.42	27.73	14.75	7.91	4.28	2.36	1.77	0.72

[10] This is most likely an erroneous alignment.

[11] https://pypi.org/project/JapaneseTokenizer/

Table 3 Translation distortion, shift to the left or right context, per language as percentages

Shift	−4	−3	−2	−1	1	2	3	4	5
da	0.4	0.65	0.95	1.63	71.22	11.07	5.27	2.82	1.17
de	1.32	2.11	2.3	3.12	51.95	10.39	5.57	3.16	2.77
es	0.54	1.04	3.2	4.95	54.73	15.34	9.23	4.6	2.54
hi	3.41	6.55	8.78	9.87	22.16	5.71	4.72	3.47	3.48
ja	5.6	8.01	9.42	11.28	12.82	6.17	4.2	3	2.16
zh	2.13	2.39	3.38	5.79	37.63	12.72	7.1	4.69	3.45

4.2 Distortion Probabilities

Table 3 shows probabilities for translation distortion values (Shift). This Shift value takes only into account the translation shifting part in the *Cross*[12] values, ignoring the length of the target group. A Shift value of 1 indicates a monotonous translation, in which translations are produced sequentially. Positive Shift values represent a displacement of the translation into the right context, while a negative value represents a displacement to the left by the indicated number. Zero values represent translation omissions (no link of the ST word into the target). The percentage of omissions is identical to the percentage of *AGnbr = 1* in Table 2. As Table 3 shows, most translations are monotonous for all six languages: more than 70% of Danish translations are produced sequentially in the same order as the English source. Again, the three Asian languages show a different pattern as compared to the three European languages.

4.3 HTra across Languages

In order to compare *HTra* values across different languages, we normalize entropy values (*HTraN*) by the *log(n)* of observations, according to Eq. (7).[13] As shown in Table 1, the translations of *yesterday* (word Nr. 23 of Text 1) have a low *HTra* value into the six languages. Table 4 reproduces the *HTra* values from Table 1 in their normalized version and shows their total *HTraN* value. Of the 152 total translations of *Yesterday* into the six languages, there are together only 19 different translations, which amounts to a total *HTraN* value of 0.411. Another low HTra word, i.e., with relatively entrenched although slightly flatter translation distribution, is English *patients* with a total *HTraN* of 0.441. The word *put* (in Text 1, word Nr. 108), in contrast, has 107 different translations into the 6 target languages (from theoretically 152 possible different translations) with relatively high *HTra* values

[12] Shift was computed as follows: if (Cross >0) {Shift = Cross – TAGnbr + 1} else if (Cross <0) {Shift = Cross + TAGnbr – 1}.

[13] A normalization is required, since the number of translation *n* varies for different texts and languages between 22 and 39 (see Appendix 1).

Table 4 Average normalized *HTra* values across words and languages

HTraN per language and word								
Text-Id	ST word	da	de	es	hi	ja	zh	*Total HTraN*
1–23	*Patients*	0.055	0.229	0.041	0.768	0.065	0.225	0.441
1–46	*Yesterday*	0.161	0.0	0.123	0.232	0.065	0.365	0.411
	...							
1–108	*Put*	0.848	0.939	0.680	0.885	0.944	0.631	0.885
	Total *HTraN*	0.651	0.651	0.627	0.762	0.715	0.691	0.7152

for all 6 languages and a total value of *HTraN (put)* = *0.885*.

$$HTraN(s) = HTra(s)/\log_2(n) \tag{7}$$

Table 4 also shows the marginal total *HTraN* values for each language. It suggests that European languages (Danish, German, and Spanish) with average *HTraN* values of 0.651, 0.651, and 0.627 are closer to English than the Asian ones (Hindi, Japanese, and Chinese) with *HTraN* values of 0.762, 0.715, and 0.691, respectively.

Punctuation marks and numbers have the lowest *HTra* values across all languages, which occasionally are also translated as conjunctions such as "and" or "or." See Ogawa et al. (this volume, Chap. 6) for a more detailed analysis of HTra correlations across different languages.

4.4 Correlation of Translation Information, Entropy, and Literality

Figure 1 shows a correlation matrix for all 120,083 word translations in the multiLing data set. It plots the six information and entropy measures introduced in Sect. 3. As already discussed in Table 1, *ISTC* strongly correlates with the word-order distortion information *ICross* ($\rho = 0.80$) and with word translation information *ITra* ($\rho = 0.82$), while *ICross* and *ITra* correlate moderately ($\rho = 0.59$). A similar pattern applies to the entropy values (*HCross*, *HSTC*, and *HTra*), where we see a very strong correlation between *HSTC* and *HTra*, as well as between *HSTC* and *HCross*, and a slightly weaker, significant correlation between *HTra* and *HCross* ($\rho = 0.69$). These rather strong correlations imply that unusual lexical translation choices are likely to come along also with more unusual reordering of the translation, and vice versa.

All three information measures also correlate significantly and moderately (ρ values between *0.43* and *0.60*) with the number of words in the alignment group (*AGnbr*), which is the sum of words in the source and the target side of the alignment group. In particular, the high correlation of *AGnbr* with *ITra* and *Htra*

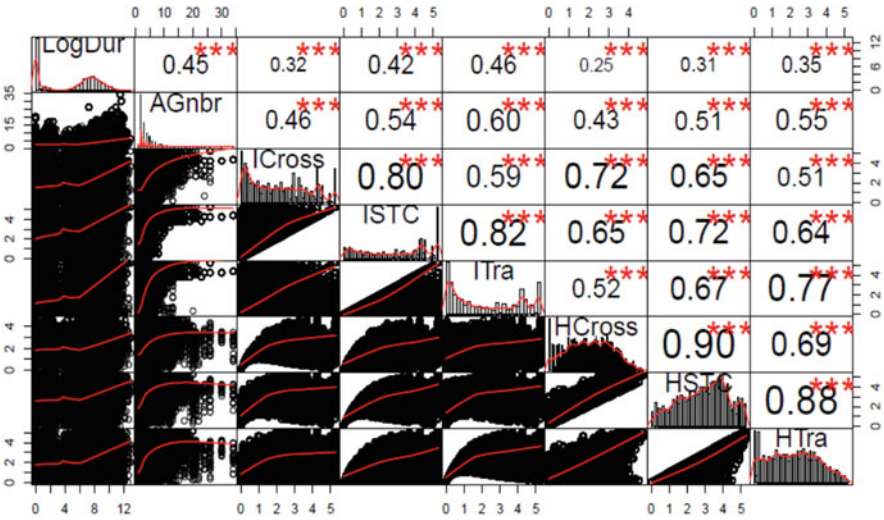


Fig. 1 Correlation matrix (Spearman ρ) of information and entropy values for the multiLing data

suggests that less compositional translations segments are also more semantically and syntactically distant from their source (see also Carl this volume-b, Chap. 14).

All measures correlate significantly but to a different degree with translation production duration (*LogDur*). About 50% of the 120,083 data points (i.e., word translations) were produced during post-editing or monolingual post-editing (i.e., PE without source text) tasks, and around 2/3 of those words (39,257 translated words) were not edited at all. Accordingly, there is a substantial amount of zero production duration for those translations, which is visible in the peak in the *LogDur*[14] histogram on the top left in Fig. 1. When taking out those data points, by only considering translations with production duration, e.g., *Dur > =20 ms*, the significance values of the correlations do not change, and only correlation coefficients for the information measures become slightly weaker. For instance, the correlation between *ISTC* and *LogDur* in Fig. 1 is $\rho = 0.42$. Taking out the 39,257 data points with zero production duration reduces this correlation to $\rho = 0.31$.

Among the three information measures, *ITra* has the strongest correlation with production duration, $\rho = 0.46$, and $\rho = 0.38$, when taking out zero durations (not shown in Fig. 1). This indicates that variation of lexical choice (*ITra*) seems to be—in general—a better indicator for translation duration than variation in reordering (*ICross*). It is interesting to see that a quite similar correlation pattern applies to the entropy values (*HCross*, *HTra*, and *HSTC*), as compared to the corresponding information values, although the correlation with duration is weaker. Notice also

[14] *LogDur* was computed as $\log(Dur + 1)$.

that *ISTC* strongly and significantly correlates with *HSTC* ($\rho = 0.73$), as well as *ITra* and *HTra* ($\rho = 0.77$).

The availability of joint syntactic and semantic information makes it also possible to assess their dependency pattern. Two events are statistically independent if the product of their probabilities equals their joint probability. Similarly, it is possible to compute the mutual information as the difference of independent entropy values and their joint entropy. The mutual information is, in fact, identical to the Kullback-Leibler (1951) divergence, which is discussed in some detail in Wei, this volume, Chap. 7. It may, thus, be possible and interesting to investigate patterns of syntactic-semantic dependencies, of information gain, surprisal, and entropy reduction during the translation process, and relate these patterns with behavioral observations such as gazing and typing patterns. We will take this topic up in the discussion section.

4.5 *Effects of Literality Measures on Translation Duration*

We also tested the effect of the two literality measures (*ISTC* and *HSTC*) on *LogNormDur*. *LogNormDur*[15] is the log-transformed production duration, normalized by the number of words in the source alignment group (*SAGnbr*). Figure 2 shows the interaction effects for *ISTC* and *HSTC* on production duration for four translation modes[16] (see Appendix 2 for an explanation of the modes). The *HSTC*

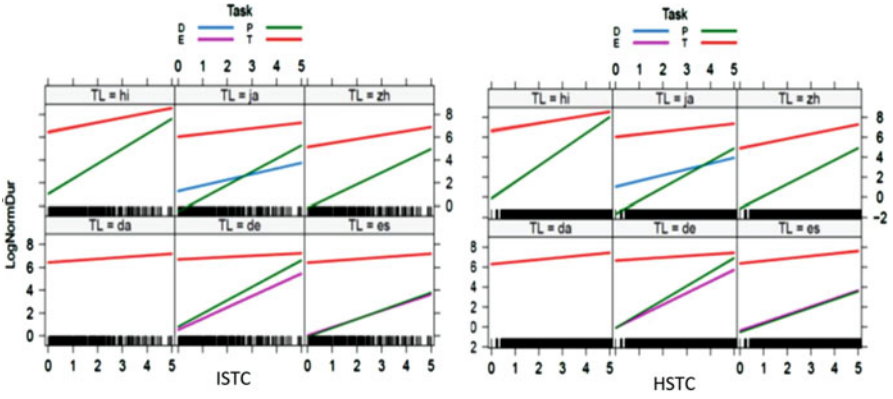


Fig. 2 Effect of word-based literality measures on production duration for four translation modes and six languages

[15] *LogNormDur* = log(*Dur/SAGnbr* + *1*), where *SAGnbr* is the number of words in the source alignment group.

[16] We used linear regression in R with target language (TL) and translation mode (Task) as interaction effects: "LogNormDur ~ ISTC * TL * Task" (Fig. 2, left) and "LogNormDur ~ HSTC * TL * Task" (Fig. 2, right).

and *ISTC* models (overall model fit was $r^2 = 0.45$ and $r^2 = 0.49$, respectively) show a consistent pattern in which the literality measures have the weakest (although significant) effects on from-scratch translation (T) and the strongest effects on post-editing (P) and monolingual post-editing (E), while translation dictation (D, only for Japanese) takes a middle position. It also shows that the effects are steepest for Hindi, while for Spanish, they are much weaker.

5 Segment-Level Literality Measures

We also compute an average information literality measure on a segment level. A segment S consists of a sequence of tokens $(w_1, \ldots, w_m)$, each of which is associated with one (possibly empty) alignment (s,t) and a distortion index (c), so that $S : \{\{w_1, \{s_1, t_1, c_1\}\}, \ldots, \{w_m, \{s_m, t_m, c_m\}\}\}$. We sum over the alignment/distortion information (*ISTC*) to compute an average segment-based *HSTC* value. In contrast to the word-based entropy values (*HSTC*), where the information is factored by the expectation of the word, here we do not know the probabilities of the individual words in the segment. We assume that each word has the same probability and normalize the sum over the joint self-information of the translated words by the length (m) of the source segment. Equations (8)–(10) show how average information values *HCross*, *HTra*, and *HSTC*[17] are computed on the segment level.

$$HTra(S) = 1/m \times \sum\nolimits_{\{w_k,\{t_k\}\}\in W} ITra\,(w_k, t_k) \tag{8}$$

$$HCross(S) = 1/m \times \sum\nolimits_{\{w_k,\{c_k\}\}\in W} ICross\,(w_k, c_k) \tag{9}$$

$$HSTC(S) = 1/m \times \sum\nolimits_{\{w_k,\{s_k,t_k,c_k\}\}\in W} ISTC\,(w_k, s_k, t_k, c_k) \tag{10}$$

Figure 3 shows (on the right) a correlation matrix of the segment-based entropy values and production duration for the segment. As with the entropy values on the word level (Fig. 1), there is a very strong correlation between the three entropy values ($\rho > 0.85$), and there is also a moderate significant correlation of all entropy measures with production duration. The left side shows the interaction effect of *HSTC* on production duration for the six languages and different production modes. The *Dur* value (top left in the correlation matrix) represents the production duration

[17] After completion of this chapter we have realized that the naming of these segment-based measures may lead to an unlucky confusion with metrics of the same name on the word-level. As the segment level values represent average self-information for word translations, alignment crossing and joint source-target-alignment, we decided to rename the segment-level metrics in the CRITT TPR-DB. For future use, we rename the segment-level measures HTra, HCross, HSTC into ITraSeg, ICrossSeg, ISTCSeg respectively.

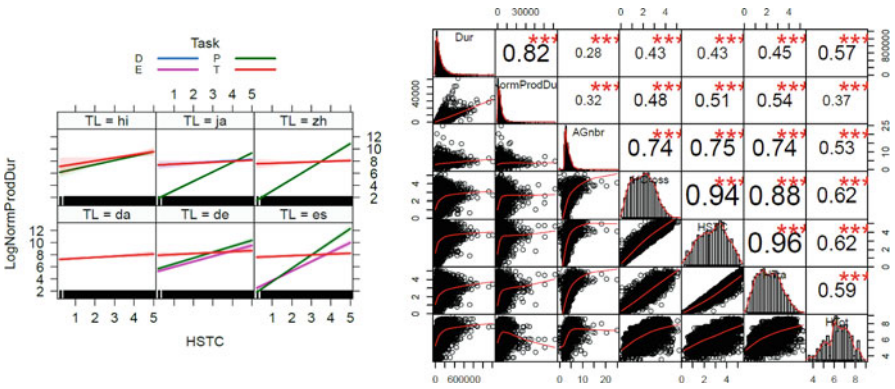


Fig. 3 Effect plot of segment-based literality measure on production duration (left) and correlation matrix for various segment-based entropy scores (right)

for a segment. It includes the performance duration needed for typing (or post-editing, dictating) the segment from the first keystroke to the last keystroke and the preceding processing duration, i.e., the pause (e.g., the gaze time) that precedes the production performance. The *NormProdDur* includes only the typing duration (i.e., the performance) and is normalized by the number of tokens in the ST segment, and the *LogNormProdDur* is its log value of *NormProdDur*.[18] As with the effect plot for literality scores on a word level, also the segment-level literality score on the left side in Fig. 3 shows a stronger effect for post-editing than for from-scratch translation. However, in contrast to the word-level analysis, the literality effects seem to be more pronounced for Spanish, Japanese, and Chinese post-editing,[19] while the effect is weaker for Hindi post-editing. A similar effect can also be observed when using a range of alternative duration variables as a dependent variable including *LogDur*, *LogNormDur*, and *LogProdDur*, where in all cases the effect is stronger for Spanish, Japanese, and Chinese post-editing and weaker for Hindi. Further investigation may be required to assess why those effects are different on a word level and a segment level.

5.1 Segment-Based Total Translation Entropy

Note that the *HSTC* measure on the segment level is—in some respect—similar to *ISTC* on the word level. *HSTC* on the segment level produces a literality score for each translation based on the joint probabilities for each token in that segment. It thus ranks different translation *observations*, just like *ISTC* does on the word level. However, we can also compute a translation entropy score on the segment level that quantifies the heterogeneity of the source segment. Instead of summing *ISTC*

[18] In terms of features in the TPR-DB, *LogNormProdDur = log((Dur – PreGap)/TokS) + 1).*

[19] We did not filter out any segments or check whether there were non-postedited segments (i.e., *Dur = 0*).

values over the joint information for one segment, we compute for each segment the total entropy *HTot* over the joint information of all n alternative translations. Thus, each ST word (w_i) of a source segment S is associated with the set of all n alternative translations and their alignments ($A_i : \{\{s_1, t_1, c_1, \}, \ldots \{s_n, t_n, c_n\}\}$). With the source segment $S : \{\{w_1, A_1\}, \ldots, \{w_m, A_m\}\}$, we can then compute the joint self-information for all $m \times n$ possible alignments of the segment, for each target language separately, as shown in Eq. (11).

$$HTot(S) = \sum_{\{w_k, A_k\} \in S} \sum_{\{s_j, t_j, c_j\} \in A_k} p\left(s_j, t_j, c_j | w_k\right) \times ISTC\left(w_k, s_j, t_j, c_j\right) \tag{11}$$

HTot quantifies the total translation entropy of a source segment, similar to *HSTC* on a word level. A *HTot* value is computed for the translation of each ST segment into the six different languages. Figure 4 shows a correlation matrix between the *HTot* values of 41 source segments in the multiLing data for 5 languages.[20] The graph shows a strong and significant correlation between the segments' *HTot* scores indicating that entropy values for the segments are similar across languages. Ogawa et al. (this volume, Chap. 6) come to a similar conclusion by looking at translations of various word categories across three languages. The figure on the left shows the correlation of the normalized *HTot* values for the 41 segments, sorted by the *HTotN* value of the German segment in ascending order.

The segment with the lowest *HTot* values is the first segment of Text 5 with *HTot* = *3.53* for German and *HTot* = *3.85* for Spanish, reproduced in Table 5. For this segment, a total of 60 different translations were produced by 127 translators into the six languages. Thirty-one Spanish translators produced only seven different translations into Spanish, and the 23 German translators also produced seven different German translations, which are shown in Table 5. The table shows that the average number of words per alignment group (*AGnbr*) is about 2, i.e., one

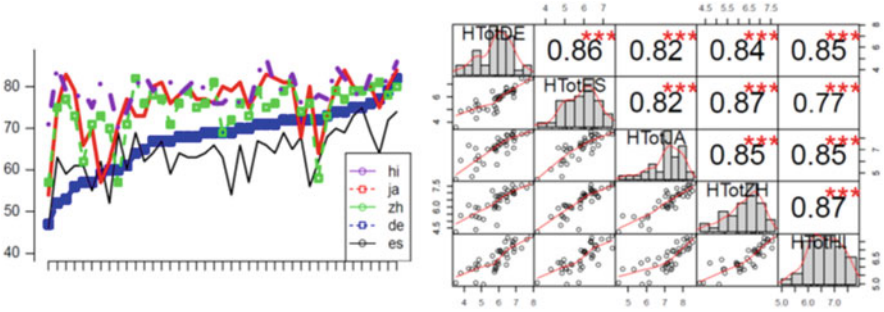


Fig. 4 Correlation between normalized *HTot* scores of 41 segments between five languages (left) and the correlation matrix of non-normalized values (right)

[20]The Danish set only has translations for the 23 segments of the first three texts; it therefore does not appear in the plot.

Table 5 Example of literal translations into German

N	AGnbr	Cross	HTra	HCross	HSTC	*Sociology* is a relatively new academic discipline
15	2.12	1.25	0.16	0.08	0.2	*Die Soziologie* ist eine relativ neue wissenschaftliche Disziplin
2	2	1	0.44	0.65	0.76	*Soziologie* ist eine relativ neue wissenschaftliche Disziplin
2	2.12	1.25	0.75	0.08	0.78	*Die Soziologie* ist eine relativ junge akademische Disziplin
1	2.37	1.25	1.29	0.08	1.33	*Die Soziologie* ist eine relativ junge Wissenschaft
1	2.37	1.25	1.49	0.08	1.52	*Die Soziologie* ist eine relativ junges Wissenschaft
1	2	1	1.24	0.65	1.56	*Soziologie* ist eine relativ neue akademische Fachrichtung
1	2.12	1.25	1.26	1.76	2.1	*Soziologie* ist eine noch recht neue akademische Disziplin

ST word and one TT word representing the most compositional translation. The average absolute *Cross* values are also very low (mostly slightly above 1), indicating an almost entirely monotone translation. It also shows that one translation has an inflection error (*junges Wissenschaft*) and that the most frequent adjective is the non-cognate form, *wissenschaftliche*, rather than the cognate *akademische*, but see also Heilmann and Llorca-Bofí (this volume, Chap. 8). The column **N** indicates the number of translations produced. The 23 translations are unequally distributed between the seven alternative versions: the vast majority of the 15 translators produced the translations in the first row, which also have the lowest *HTra* and *HSTC* values. Note that the 1-to-2 alignment *Sociology* $\longleftrightarrow$ *Die Soziologie* is far more frequent than the more compositional 1-to-1 translation *Sociology* $\longleftrightarrow$ *Soziologie*. However, the former solution is—due to its higher frequency—apparently better with respect to the TL grammatical constraints and thus turns out to be the most *rendered* literal solution, according to the introduced information-theoretic metric. The translations in row 2 and 7 ($AGnbr = 2$, $Cross = 1$) are most literal with respect to monotonicity and compositionality criteria, but low *HCross*, *HTra*, and *HSTC* values in the first row indicate a higher agreement among translators with respect to the literality criteria and thus potentially has better fit with TL grammatical constraints. We take this up in the discussion section.

In contrast to the literal translations in Table 5, the third sentence in text 3 of the multiLing corpus is one of the most nonliteral segments in the dataset. This sentence was translated by 148 different translators into 147 different versions for the six languages. It has high *HTot* values across all languages: Spanish, 7.47; Japanese, 8.34; Danish, 7.17; Hindi, 7.11; Chinese, 7.55; and German, 7.32. Table 6 shows 4 out of the 31 different translations into Spanish. Segments with low segment translation information ($HSTC = 1.49$ and $HSTC = 1.52$) indicate, in fact, a more literal translation than some of the translations in Table 5, with only slightly larger average alignment groups and distortion factors.

Table 6 Four Spanish translations of an English high entropy sentence (*HTot = 7.47*)

His withdrawal comes in the wake of fighting flaring up again in Darfur and is set to embarrass China, which has sought to halt the negative fallout from having close ties to the Sudanese government					
Translation 1	**AGnbr: 2.46**	**Cross: 1.57**	**HTra: 1.19**	**HCross: 0.9**	**HSTC: 1.49**
Su Retiro se produce a raíz de la lucha contra la quema de nuevo en Darfur y pretende avergonzar a China, que ha tratado de frenar las consecuencias negativas de tener estrechos vínculos con el gobierno sudanés					
Translation 2	**AGnbr: 2.33**	**Cross: 1.54**	**HTra: 1.25**	**HCross: 0.91**	**HSTC: 1.52**
Su Retiro se produce a raíz de la lucha contra la quema de nuevo en Darfur y está previsto que avergüence a China, que ha tratado de frenar las consecuencias negativas de mantener estrechos vínculos con el gobierno sudanés					
...					
Translation 30	**AGnbr: 6.49**	**Cross: 3.81**	**HTra: 3.77**	**HCross: 3.13**	**HSTC: 4.18**
Su retirada llega en en el inicio de la lucha que se lleva a cabo en Darfur y su intención es avergonzar a China por su negativa a romper la estrecha relación que mantiene con el gobierno sudanés					
Translation 31	**AGnbr 3.89**	**Cross: 2.89**	**HTra: 3.63**	**HCross: 2.72**	**HSTC: 4.35**
Esto radica en las ganas de que resurja la lucha en Darfur y se ha hecho Para avergonzar a China, la cual ha pronunciado que solucionará su fallo en haber mantenido lazos estrechos con el gobierno de Sudan					

The upper Translations 1 and 2 have low literality values; Translations 30 and 31 have high literality values.

Larger word-order distortion is more likely in the example in Table 6, a sentence with 37 ST words (including punctuation) as compared to eight ST words in the example in Table 5. Table 6 shows instances of nonliteral translation 30 and 31, with high translation information (*HSTC = 4.18* and *HSTC = 4.35*). This nonliterality is the result of larger alignment groups, larger reordering, and usage of less frequent word choice in line with the three literality criteria. Note also that Translation 31 has a higher HSTC value than Translation 30 even though all of the individual literality criteria are lower than for Translation 30. It is thus possible that Translation 31 makes use of more entrenched word (or phrase) translations and more frequent reordering patterns, but in a combination that renders the entire translation less literal than Translation 30, which receives a lower joint entropy score based on less entrenched translations, monotonicity, and compositionality scores.

6 Discussion and Conclusion

We reinterpret translation literality measures in an information-theoretic framework. The chapter introduces various self-information and entropy measures on a word level and a segment level. The *joint ST-TT-alignment crossing entropy* specifies the expected heterogeneity of possible translations for a source expression, whereas

Table 7 Differences in alignment grouping (AGnbr) lead to different literality scores

AGnbr	Cross	HTra	HCross	HSTC	Yesterday, he was found guilty of four counts of murder following a long trial
2.12	1	0.83	0.91	1.25	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
3.56	1.88	2.57	3.01	3.55	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo

the *joint ST-TT-alignment crossing information* measures the literality of individual translations. Both measures significantly correlate: translations have lower levels of self-information if the language systems exert more selection pressure for a particular solution. We show that more frequent translations with high amounts of self-information tend to be less literal, and they are less compositional and less monotone. More compositional translations tend to be realized in a monotonous, word-for-word fashion and with clearly preferred lexical equivalents in the target. A large amount of lexical variation correlates with less predictable word reordering and with a less compositional and more literal translation. We also show that the literality measures of information and entropy on a word level and a segment level are predictive for translation duration.

One major shortcoming of the measures is their dependence on alignment grouping. As automatic alignment methods have not (yet) proven to be sufficiently precise, all translations in the multiLing corpus have been manually aligned (cf. Carl et al. 2016). However, manual alignment grouping may be inconsistent or incoherent, and this inconsistency leads to different translations and distortion probabilities and as a consequence different literality scores of the same sentence. For instance, Appendix 2 shows 31 Spanish translations of "Yesterday, he was found guilty of four counts of murder following a long trial" which was discussed in Sect. 4. The segment has been translated by four translators in exactly the same way, but all translations have different literality scores. Two of the translations are reproduced in Table 7, which shows two quite different strategies of alignment groupings, one with an average of *2.12* tokens per alignment group, the other one with almost twice as many tokens, an average of *3.56* tokens per alignment group. This difference also leads to different information and entropy values for the two segments. This apparent disagreement of translation grouping reminds us of the fundamental disagreement in translation quality rating that has frequently been reported (e.g., Lommel et al. 2014), and for which to date, no general agreed solution has been suggested.

Another question of concern regards how much the results can be generalized if the self-information and entropy values are computed on sets of 15 or 30 alternative translations: to what extent are the word-based *HTra* values and, as a consequence, the *HSTC* and *HTot* representative of the cross-linguistic self-information and entropy that the language systems provide? Will the *HTra* values and the relation between them change if another set of translations is considered? If *HTra* (and

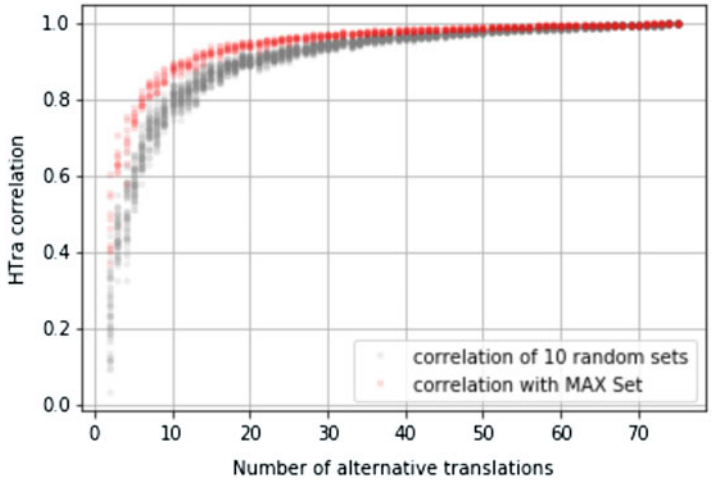


Fig. 5 Convergence of HTra scores towards a "real" population score

HCross) was overly dependent on the selections of our dataset, the results could not be carried over to other sessions.

To assess this question, we investigated the convergence of *HTra* values with a set of 75 alternative English-to-Chinese translations of 160 words (i.e., Text 1 of the multiLing corpus). Assuming that the 75 alternative translations exhaust the translational variation and represent a "real population" of alternative translations, the *HTra* values from this set would also represent the "real" word translation entropy. To assess how quickly subsets would converge toward the "real" *HTra* values, we repeatedly extracted 10 random sets from the "population" with $n : 2..75$ alternative translations, computed their *HTra* values, and ran a Pearson correlation between each of the ten subsets and the population (MAX) *HTra* (red dots in Fig. 5). We also computed the Pearson correlations between all ten random sets, which indicate the expected agreement between *HTra* values of arbitrary generated alternative translations (black dots in Fig. 5). The *HTra* values of the extracted subsets surprisingly quickly converged toward the MAX population *HTra* and a bit less quickly among their mutual *HTra* values. The graph in Fig. 5 shows that with approximately ten alternative translations, we reach a correlation with the real (MAX) *HTra* of more than $r = 0.8$, and with 20 observations (i.e., alternative translations), we are above $r = 0.9$. With our dataset of 20–40 alternative translations, we can thus confidently say that generalizations may be valid beyond our limited dataset. Results that are compatible with such findings are also reported by Hale (2016, 403) who reports that in the context of parsing "even a restriction to three or four syntactic analyses leads to good performance."

While this chapter introduces measures of joint information and joint entropy to assign (non) literality scores to instances of the translation product, it is also possible to apply the framework to the translation process and investigate *translation*

surprisal and *translation entropy reduction* during translation task performance (Hale 2016). In this setting, information and/or entropy values would be calculated before and after each translation step, and their ratio (or difference) would indicate the surprisal or entropy reduction, respectively. Hale (2016, 398) points out that "Surprisal and Entropy Reduction are incremental complexity metrics that predict how difficult each word should be as it is perceived in time ... [and b]oth metrics suppose that greater information value should relate to greater processing difficulty." Wei (this volume, Chap. 7) investigates gazing patterns under an angle of surprisal and entropy reduction. We think there is much scope to extend this framework in future research.

Appendix 1: The *multiLing* Corpus

In this study, we use different parts of the *multiLing* corpus. The *multiLing* corpus is part of the CRITT TPR-DB[1] which consists of six short English texts that have been translated into Danish (da), German (de), Spanish (es), Chinese (zh), Hindi (hi), and Japanese (ja) by several translators under various conditions: from-scratch translation (T), post-editing[2] (P), and monolingual editing[3] (E). The set has been described among others in (Carl et al. 2016), and various subsets have been used in previous studies (e.g., Schaeffer and Carl 2014, Schaeffer et al. 2016, Carl and Schaeffer 2017a, b, and in this volume, among others). Table 8 gives an overview of some of the properties of the six source texts and the number of segments and words in each text, as well as the number of produced translations for each translation mode. The six texts have a total of 847 words and 41 segments, respectively, 160, 153, 146, 110, 139, and 139 words and 11, 7, 5, 5, 6, and 7 segments for texts 1–6, as indicated in rows #W and #S. Table 8 also shows how many translations have been produced in each of the translation modes. There is a total of 889 translations for the six texts into the six languages—the total number for each ST, translation mode, and language is shown in the "total" columns and row—amounting to a total of 6112 translated segments and almost 126,000 target language words. Not all languages have the same number of translations. For instance, English-to-Danish

[1]The *multiLing* corpus is a subset of the TPR-DB, which can be downloaded free of charge from https://sites.google.com/site/centretranslationinnovation/tpr-db

[2]We used mostly Google translate output for post-editing. In the case of Hindi, we also used AnglaBharti (http://tdil-dc.in/tdildcMain/IPR/AnglaMT-II.pdf). There exists no post-editing data for Danish.

[3]In the monolingual editing mode (E), translators are asked to "post-edit" the MT output without access to the source text. This is sometimes also referred to as blind post-editing, and one reviewer suggested that it is neither post-editing nor translation. The "E" data exists only for German, Spanish, and Chinese. The same column shows translation dictation (D), which is only available for English to Japanese.

only has from-scratch translations for three texts, and only Japanese has translation dictation, marked in bold in Table 8.

All translation sessions were recorded using Translog-II (Carl 2019), which logs all keystrokes with a time stamp. In addition, all sessions were recorded using an eye-tracker, and the gaze data were synchronized with the keystroke log and post-processed as described in Carl et al. (2016).

Four of the texts are general news texts, with different degrees of difficulty (average sentence length and word frequency), and two texts (5 and 6) are excerpts from a sociological encyclopedia. The texts were translated without external help[4]—so as to capture gaze data during the translation sessions. A short glossary (with 1–4 term translations) was shown to (most of) the Japanese and Chinese translators before each session. The translations were produced mostly by translation students into their L1 and collected over the past 10 years (2008–2017) for various studies. The translation of each set of six texts took approximately 1.5–2 h, including a short briefing, filling a general personal questionnaire, and calibration of eye-tracker.[5] Due to the general nature of the texts, we believe that the recorded user activity data represent general translation performance, and the observed variation in the process and the product are comparable and representative—with some caution—to similar situations of general language translation.

All texts segments and words were manually aligned using the YAWAT tool (Germann 2008).[6] In a briefing phase, aligners were instructed to produce maximally exhaustive and compositional alignments. That is, the translated segments should be fragmented into minimal alignment groups, which should ensure a most compositional (dynamic) equivalence of the source and target text. The translated segments should also cover a maximum number of words, so that a minimum number of words remain unaligned. Words should only remain unaligned if the content was inserted or missing in the translation.

Appendix 2: Literality Values for Alternative Spanish Translations

Alternative Spanish translations of "Yesterday, he was found guilty of four counts of murder following a long trial" sorted by HSTC value. The *HTot* value is *5.92*.

- **AGnbr**: number of source and target tokens per alignment group
- **Cross**: mean of alignment distortion.
- **HCross**: entropy of *Cross* values, according to Eq. (8).

[4]That is, consultation of dictionaries or Internet search was not allowed.

[5]The metadata can be downloaded from the repository. It provides more detailed information for each of the more than 150 translators.

[6]The aligned versions can be accessed via a browser from the CRITT website: https://sites.google.com/site/centretranslationinnovation/yawat

Table 8 Summary information of the *multiLing* dataset: six source texts with a total of 847 words were translated into six different languages (da, de, es, hi, ja, zh) under four different translation modes (post-editing, translation, monolingual editing, dictation) resulting in 889 translated texts with a total of more than 120,000 target text words in the six languages

Text	Source texts																					
	1	2	3	4	5	6	total	1	2	3	4	5	6	total	1	2	3	4	5	6	total	
#W	160	153	146	110	139	139	847	160	153	146	110	139	139	847	160	153	146	110	139	139	847	
#S	11	7	5	5	6	7	41	11	7	5	5	6	7	41	11	7	5	5	6	7	41	
	Postediting						Total	Translation						Total	Editing/***Dictation***						Total	Total
da	–	–	–	–	–	–	–	24	23	22	–	–	–	69	–	–	–	–	–	–	–	69
de	8	7	8	8	7	8	46	7	8	8	8	8	8	47	7	8	8	7	8	8	46	139
es	10	12	10	12	8	12	64	11	9	11	10	11	8	60	10	9	10	10	10	11	60	184
hi	8	12	8	10	12	11	61	7	7	6	7	6	6	39	–	–	–	–	–	–	–	100
ja	13	12	13	12	13	12	75	13	13	12	13	12	13	76	***13***	***13***	***13***	***12***	***13***	***14***	***78***	229
zh	13	17	15	14	12	10	81	11	12	12	13	15	14	77		1	1	4	2	2	10	168
Total	52	60	54	56	52	53	327	73	72	71	51	52	49	368	17	18	19	21	20	21	116	889

- **HTra**: entropy of translations, according to Eq. (9).
- **HSTC**: joint entropy of ST-TT-alignment cross values, according to Eq. (10).
- **HTot**: for the source sentence is 5.92 and 0.66 for normalized HTot.

AGnbr	Cross	HTra	HCross	HSTC	Yesterday, he was found guilty of four counts of murder following a long trial
2.12	1	0.83	0.91	1.25	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
2.12	1	0.95	0.85	1.33	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
2.12	1	1.08	0.85	1.42	Ayer, fue declarado culpable de cuatro cargos de asesinato tras un juicio largo
2.57	1.25	1.14	1.11	1.66	Ayer, fue declarado culpable de cuatro cargos de asesinato tras un largo juicio
2.88	1.62	1.34	1.29	1.88	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
3	1.44	1.36	1.09	1.9	Ayer, fue hallado culpable de cuatro asesinatos tras un largo juicio
3.06	1.81	1.26	1.41	1.92	Ayer, fue declarado culpable de cuatro cargos de asesinato tras un juicio largo
2.38	1.12	1.23	0.91	2.09	Ayer, fue culpado de cuatro cargos de asesinato tras un juicio largo
2.62	1	1.75	0.85	2.16	Ayer, fue culpado de cuatro homicidios tras un largo juicio
1.88	0.94	2	0.95	2.27	Ayer se le consideró culpable de cuatro casos de asesinato tras un juicio largo
2.94	1.62	1.68	1.83	2.52	Ayer fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
2.5	1.38	1.23	1.77	2.55	Ayer, tras un largo juicio, fue declarado culpable de cuatro asesinatos
3.75	2.06	1.91	1.75	2.55	Ayer, fue hallada culpable de cuatro cargos de asesinato tras un juicio largo
2.38	1	2.13	1.09	2.57	Ayer, se le declaró culpable de cuatro asesinatos siguiendo un largo trayecto
2.81	1.25	1.61	1	2.6	Ayer, fue considerado culpable de cuatro asesinatos tras un juicio largo
2.24	1.31	1.95	1.64	2.73	Ayer, después de un largo juicio, se le reconoció culpable de cuatro imputaciones de asesinato
3.12	1.5	1.86	1.61	2.86	Ayer fue declarado culpable por los cuatro asesinatos tras un largo juicio
3.62	1.94	2.59	1.65	2.93	Ayer fue culpado de cuatro cargos de homicidio tras un largo juicio
2.87	1	1.8	2.36	2.99	Fue hallado Ayer culpable de cuatro cargos de asesinato tras un juicio largo
3.56	1.88	2.19	1.48	3.35	Ayer, fue declarado culpable de cuatro homicidios tras un largo juicio

(continued)

AGnbr	Cross	HTra	HCross	HSTC	Yesterday, he was found guilty of four counts of murder following a long trial
3.69	2.12	3.26	1.83	3.45	Ayer fue declarado culpable en cuatro sentencias de homicidio después de un largo juicio
3.56	1.88	2.57	3.01	3.55	Ayer, fue hallado culpable de cuatro cargos de asesinato tras un juicio largo
2.69	1.88	2.82	2.72	3.6	En el día de Ayer se declaró culpable a Norris de los cuatro asesinatos tras un largo proceso judicial
2.31	1.12	1.91	2.05	3.7	Después de un largo juicio, Ayer fue acusado de cuatro cargos por homicidio
3.37	2.25	3.11	2.3	3.81	Ayer, se dictaminó que era culpable de las cuatro causas por asesinato después de un largo juicio
3.69	3.19	3.34	3.28	4.13	Después de un largo juicio, se le consideró culpable de los cuatro cargos de asesinato

References

Brown PF, Pietra D, Stephen A, Pietra D, Vincent J, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19(2):263–311. https://www.aclweb.org/anthology/J93-2003

Carl M (this volume-a) Micro units and the first translational response universal. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M (this volume-b) Translation norms, translation behavior, and continuous vector space models. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M (2019) Logging and mapping keystroke with Translog-II and the CRITT TPR-DB. Paper presented at 9th Language & Technology Conference. LTC 2019, Poznań, Poland

Carl M, Schaeffer M (2014) Word transition entropy as an Indicator for expected machine translation quality. In: Miller KJ, Specia L, Harris K, Bailey S (eds) Proceedings of the workshop on automatic and manual metrics for operational translation evaluation. MTE 2014. European Language Resources Association, Paris, pp 45–50

Carl M, Schaeffer M (2017a) Sketch of a noisy channel model for the translation process. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 71–116. (Translation and multilingual natural language processing; no. 7)

Carl M, Schaeffer M (2017b) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. Hermes 56:43–57. https://doi.org/10.7146/hjlcb.v0i56.97201

Carl M, Schaeffer M (2017c) Measuring translation literality. In: Jakobsen AL, Mesa-Lao B (eds) Translation in transition: between cognition, computing and technology. John Benjamins Publishing Company, Amsterdam, pp 82–106. (Benjamins Translation Library, Vol. 133)

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, Berlin, pp 13–54

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process. Transl Cogn Behav 2(2):211–232

Catford JC (1965) A linguistic theory of translation: an essay in applied linguistics. Oxford University Press, Oxford

Chesterman A (2011) Reflections on the literal translation hypothesis. In: Alvstad C, Hild A, Tiselius E (eds) Methods and strategies of process research. John Benjamins, Amsterdam, pp 23–35

Dimitrova BE (2005) Expertise and explicitation in the translation process. Benjamins Translation Library 64

Germann U (2008) Yawat. Yet another word alignment tool. In: Proceedings of the ACL-08: HLT Demo Session (companion volume). Association for Computational Linguistics, Stroudsburg, PA, pp 20–23. http://www.aclweb.org/anthology/P08-4006

Hale J (2016) Information-theoretical complexity metrics. Wiley Online Library, Hoboken, NJ. https://doi.org/10.1111/lnc3.12196

Halverson SL (2015) Cognitive translation studies and the merging of empirical paradigms. The case of 'literal translation'. Translation Spaces 4(2):310–340

Halverson SL (2019) A construct for cognitive translation and interpreting studies. Transl Cogn Behav 2(2):187–210

Heilmann A, Llorca-Bofí C (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Ivir V (1981) Formal correspondence Vs. translation equivalence revisited. Poet Today 2(4):51–59

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86. https://doi.org/10.1214/aoms/1177729694. JSTOR 2236703. MR 0039968

Lommel A, Popovic M, Burchardt A (2014) Assessing inter-annotator agreement for translation error annotation. In: MTE: Automatic and Manual Metrics for Operational Translation Evaluation, LREC-2014. Association for Computational Linguistics, Stroudsburg, PA

Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and source-text background information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Pustejovsky J (2012) Co-compositionality in Grammar. In: Hinzen W, Machery E, Werning M (eds) The Oxford handbook of compositionality. https://doi.org/10.1093/oxfordhb/9780199541072.013.0017

Rosetta MT (1994) Compositional translation. In: The Springer international series in engineering and computer science. Springer, Cham

Schaeffer M, Carl M (2014) Measuring the cognitive effort of literal translation processes. In: Germann U, Carl M, Koehn P, Sanchis-Trilles G, Casacuberta F, Hill R, O'Brien S (eds) Proceedings of the workshop on humans and computer-assisted translation (HaCaT). Association for Computational Linguistics, Stroudsburg, PA, pp 29–37

Schaeffer M, Dragsted B, Hvelplund K, Balling L, Carl M (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, Cham, pp 183–210

Tokowicz N, Kroll JF, de Groot AMB, van Hell JG (2002) Number-of-translation norms for Dutch-English translation pairs: a new tool for examining language production. Behav Res Methods Instrum Comput 34:435–451

Vanroy B, De Clercq O, Tezcan AA, Daems J, Macken L (this volume) Metrics of syntactic equivalence to assess translation difficulty. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Wei Y (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

RedBird: Rendering Entropy Data and ST-Based Information into a Rich Discourse on Translation

Investigating Relationships Between MT Output and Human Translation

Haruka Ogawa, Devin Gilbert, and Samar Almazroei

Abstract This study investigates the relationship between machine translation (MT) and human translation (HT) through the lens of word translation entropy, also known as HTra (i.e., a metric that measures how many different translations a given source text word has). We aligned different translations from multiple MT systems (three different target languages: Japanese, Arabic, and Spanish) with the same English source texts (STs) to calculate HTra for each language, and we then compared these values to additional HT data sets of the same STs and languages. We found that MT HTra correlates strongly with HT HTra within and across the languages. We also annotated the ST in terms of word class, figurative expressions, voice, and anaphora in order to examine the relationships these ST features have with HTra. For this same purpose, we normalized all HTra values (nHTra) in order to compare HTra values across all six data sets. We found that these source text features are, in general, associated with HTra in the same manner regardless of target language or the distinction between MT and HT.

Keywords Translation entropy · Machine translation · Human translation · Source text features

1 Introduction

Driven by high-tech advances and the rise of global marketing, veritable waves of automation have reshaped the landscape of many industries, and the language industry is no exception. One of the recent, widespread shifts in the landscape of the language industry is the integration of machine translation (MT) systems

H. Ogawa · D. Gilbert (✉) · S. Almazroei
Department of Modern and Classical Language Studies, Kent State University, Kent, OH, USA
e-mail: hogawa@kent.edu; dgilbe10@kent.edu; salmazr1@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_6

into translation project workflows in parallel with recent advances in MT performance (Schwartz 2018; Jia et al. 2019; Läubli et al. 2020). Consequently, more and more language service providers tend to rely on MT systems coupled with human post-editing to ensure greater productivity and shorter turnaround times (Guerberof Arenas 2009; Moorkens and O'Brien 2015; Koponen et al. 2019).

Our project has come into being against the backdrop of just these changes. Our aim is to better understand the relationship between MT and human translation (HT) by examining associations between the two, focusing specifically on the concept of word translation entropy, known as HTra (Carl et al. 2016b, see Section 2 of this chapter for details). Using three diverse languages, namely Japanese, Spanish, and Arabic, we address two research questions. The first of these inquires as to why there is a correlation between HTra calculated from multiple MT systems and HTra calculated from multiple human subjects, as observed in Almazroei et al. (2019). We aim to provide some possible reasons why there are positive correlations within and across languages (see Lacruz et al., this volume).

Our second research question is centered around an exploratory analysis: what commonalities and differences between the three languages—and between MT and HT—are there when comparing HTra with respect to several semantic/syntactic features of the source text (ST)? Much of our research focus here has to do with part of speech (PoS) on the ST side. We are also interested in the differences in HTra as a representation of so-called linguistic distance, or the relative distance between different languages (see Isphording and Otten 2013) between the source language and the target language. Additionally, we annotated figurative expressions, passive/active voice, and anaphora to investigate whether these categories in the ST are associated with different HTra values.

2 Related Literature

Let us first review the most crucial concept to our study: word translation entropy, or HTra. It is a product-based translation difficulty indicator, a metric of lexico-semantic variation among multiple translations of the same source text. Word translation entropy is adapted from information entropy, often called Shannon entropy (Shannon 1951). Whereas information entropy measures how much information is in a string, HTra is a measure that reflects how many strings, out of a finite number of strings, are non-identical to each other.[1] If a certain number of translators translate the same sentence and then align all of the words in each of their translations to the same source sentence, HTra is a way to measure how varied their translations of any given word are.

[1]This online calculator (https://www.shannonentropy.netmark.pl/) can be a useful resource for understanding how entropy is calculated.

To give a basic example of what HTra represents in the context of a translation study, if a group of 12 Japanese-English translators were given the word 猩々紅冠鳥 to translate, and all 12 of them translated it as "cardinal," then the HTra would be 0. Now, if they all translated it differently—no matter how slight the differences, e.g., "cardinal" and "cardinals" would be counted as non-identical strings—then they would achieve the maximum entropy value for any group of 12 strings: 3.58. HTra also takes into account the frequency (or probability) with which any given string occurs. For example, if 11 translators rendered 猩々紅冠鳥 as "cardinal" while 1 translator wrote "redbird," the HTra value would be 0.41. If, however, 6 translators wrote "cardinal," and the other 6 wrote "redbird," then the HTra value would be 1.0.

Researchers have found that HTra correlates with behavioral measures of cognitive effort in HT, such as production duration and fixation counts (e.g., Carl and Schaeffer 2017; Vanroy et al. 2019). Simply put, this suggests that translators tend to spend more time translating or looking at words that also have higher HTra values. HTra has also been reported to correlate across languages. Schaeffer et al. (2018) examined HTra calculated from translations in six languages (Danish, Spanish, German, Hindi, Chinese, and Japanese), all of which were based on the same STs. They found that the HTra of one language correlates with that of another and that the correlations are stronger when the two languages belong to the same group (i.e., Asian or European language groups; Schaeffer et al. 2018).

Furthermore, Carl and Schaeffer (2017) compared from-scratch translation to post-editing of statistical MT output and showed that the HTra of post-editing correlates with the HTra of from-scratch translation. They also investigated the possible semantic and syntactic properties in the English source texts that might have an effect on the number of alternative translations in German and Spanish. Their findings have shown that some PoS categories, such as superlatives and proper nouns, exhibit less word translation entropy, whereas other categories like verb participles and particles have produced a larger number of translation alternatives in the target texts.

These findings have motivated us to study the relationships between MT and HT with respect to HTra. We are particularly intrigued by the remark by Carl and Toledo Báez that MT systems and humans may "face similar decision-making problems for the same ST words across different languages" (2019, 347). Do MT systems have problems similar to human translators regardless of target language? To find an answer to this question, we decided to examine HTra in three languages that are typologically distinct from one another.

3 Procedure

In preparing MT output that we would be able to compare to HT, we used the multiLing texts from the Translation Process Research Database (TPR-DB) housed by the Center for Research and Innovation in Translation and Translation

Technology (CRITT). There are six English source texts in multiLing, comprising a total of 847 ST tokens and 40 segments. Four of them are news articles, and two of them are encyclopedic texts dealing with sociology. Each text was translated using commercially available MT systems: 13 different systems for Japanese, 12 for Arabic, and 9 for Spanish.[2] We threw out the data from three potential MT systems that were originally used for Spanish, along with another three for Japanese, because the output quality was too low.[3] This resulted in 34 different TT versions of each multiLing text.

After obtaining the MT output, the tokens in each machine-translated target text were aligned compositionally to the tokens in the corresponding English source text using Yawat (Germann 2008). This means that the tokens in each multiLing source text were aligned 34 different times. Tokens were aligned with the aim of breaking phrases down to the smallest units possible, with consistency being key in order for the HTra metric to only reflect output variation and not differences in alignment. For example, if an MT system translates the news story headline "Killer nurse receives four life sentences" as "La enfermera del asesino recibe cuatro condenas a cadena perpétua," "Killer" would be aligned with "del asesino," "nurse" with "La enfermera," "receives" with "recibe," "four" with "cuatro," "life" with "a cadena perpétua," and "sentences" with "condenas." The data was then transformed into tables using the TPR-DB toolkit. Included in this process of data transformation is the calculation of HTra (see Carl et al. (2016b, 15–17) for details on the data compilation process,[4] and see Carl et al. (2016b, 29–33) and Carl and Schaeffer (2017, 46–48) for an in-depth definition of how HTra is calculated in the context of the CRITT TPR-DB).

For HT data, we used the studies corresponding to the following IDs: AR20[5] for Arabic (Almazroei et al. 2019), ENJA15 for Japanese (Carl et al. 2016a), and BML12 for Spanish (Mesa-Lao 2014). In each study, every participant translated two texts in three different modes. Table 1 summarizes general characteristics of the six studies we used, showing the target language, the number of participants (or MT systems in the case of the MT studies), and the modes (translation, post-editing, etc.).

[2]We used the following MT systems. Amazon Translate, Bing, DayTranslations, Google, Online English Arabic Translator, Prompt Online, Reverso, Systran, Tradukka, Translator.eu, Translator, and Yandex for Arabic; Baidu, Bing, Excite, Google, Paralink ImTranslator, Infoseek, MiraiTranslate, Pragma, So-Net, Textra, Weblio, WorldLingo, and Yandex for Japanese; and Amazon Translate, Baidu, Bing, DeepL, Google, Lilt, Pragma, Yarakuzen, and Yandex for Spanish.

[3]This was evaluated manually by the researchers, based on criteria of simple usability.

[4]The data, including a description of the multiLing texts, is publicly available on the CRITT website under the following study IDs: ARMT19 for the Arabic data; JAMT19 for the Japanese data; and ESMT19 for the Spanish data.

[5]The version of the AR20 study that we used had less sessions than the current version and can be downloaded from https://sourceforge.net/projects/tprdb/ under the version number "r561."

Table 1 Summary of TPR-DB studies

	HT			MT		
	ENJA15	**AR19**	**BML12**	**JAMT19**	**ARMT19**	**ESMT19**
TL	Japanese	Arabic	Spanish	Japanese	Arabic	Spanish
Participants	39	15	32	13	12	9
Modes	T,P,D	T,P,S	T,P,E	MT	MT	MT

T translation, *P* post-editing, *S* sight translation, *D* dictation, *E* editing, *MT* raw machine translation

For our exploratory analysis, we normalized the HTra values for each study in order to be able to directly compare HTra across the six studies in absolute terms. Otherwise, only studies with the same number of participants or MT systems would have the same range of HTra values. This normalization was accomplished by dividing each HTra value by the maximum theoretical HTra value for the corresponding study, which is the log of the number of participants. For example, the first token of the first text in ENJA15 has an HTra value of 3.03. As ENJA15 involves 39 participants, we divide this by 5.29, which is the log of 39, to arrive at 0.57. In this way, all HTra values are transformed so that they fall on a scale of 0 to 1. Since HTra values are determined logarithmically, this data transformation does carry a small risk of data-distortion because—as the number of participants increases—it is possible for transformed HTra values to be relatively lower than actual HTra values. Nonetheless, this risk was deemed to be acceptable considering the relative similarity of the six studies in terms of number of participants. Normalized HTra values will be referred to as *nHTra*.

We also created four annotation categories[6]. The first category is *WordClass*, which groups PoS tags into five classes that are more general than PoS: Noun, Verb, Adjective, Adverb, and Other. Although ST tokens are automatically assigned PoS tags when TPR-DB tables are generated,[7] this process is not completely accurate. Accordingly, one researcher manually corrected these PoS tags, and all changes were debated and vetted by the group of three researchers. Some of the most common PoS corrections were misclassified gerunds and adjectives that were automatically tagged as nouns. This is to be expected since noun forms in English can function as adjectives as well, and gerunds can function as adjectives, nouns, or verbs.

The other annotation categories are also multilevel factors. There are 3 levels in *Figurative* (Metaphoric, Fixed, and Other) as well as *Voice* (Passive, Active, and Other), and 2 levels in *Anaphora* (Anaphoric and Other). As for *Figurative*, Metaphoric refers to the tokens that are part of expressions whose intended meaning

[6] All our annotations of the multiLing source texts can be downloaded from https://devrobgilb.com/Researcher/Repository/multiLing/.

[7] The tagset used for PoS is the Penn Treebank Project tagset and can be found at https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

deviates from their literal sense, while Fixed, representing fixed expressions, refers to strings of words that are used idiomatically. For example, the sentence *British families have to cough up an extra £31,300 a year* has two metaphoric tokens 'cough up' and two fixed tokens 'have to.' As for *Voice*, in the case of *he was found guilty of four counts of murder*, the tokens 'was found' were tagged as Passive. Any verbs that were not tagged as Passive were tagged as Active, and all other tokens were tagged as Other. Finally, a token was tagged as Anaphoric if it was referring to something coming before or after it, a typical example being a pronoun.

Going forward, metrics from Arabic data will be preceded with 'AR'; Japanese, 'JA'; and Spanish, 'ES.' After this language tag, there will be another label to indicate human or machine translation: either 'HT' or 'MT.' For example, to report on human Japanese word translation entropy, we will use 'JA_HT_HTra.' If the language tag is left off (e.g., MT_HTra), this refers to the specific measure across all three languages.

4 Correlations Among HTra

In a previous study, we used the same data set to investigate whether there is a correlation between MT and HT in terms of HTra (Almazroei et al. 2019). As Fig. 1 shows, we have found that MT_HTra correlates strongly and positively with HT_HTra within all three languages. Furthermore, we found that HTra correlates across languages for both MT and HT, although the associations were weaker than the MT-HT correlations within each language.

These correlations in Fig. 1 suggest that MT systems and human translators tend to produce divergent translations over the same stretches of the ST (i.e., they both tend to have higher HTra values for the same ST tokens). But we are now left with the question, why? Since both neural and statistical MT systems are trained with human translations, it makes sense that MT output has similar traits to HT. This explains, to a degree, the rather strong correlations we observe within each language, but it is not sufficient to explain the moderate correlations across the three languages.

Another possible explanation for the similarities in HTra between HT and MT is that they stem from ST features. There is a considerable body of research behind ST features that are more likely to lead to MT errors. This area started with translatability indicators (sometimes called negative translatability indicators, or NTIs) for rule-based MT (RBMT), which included higher-order features such as sentence length, relative clauses, sentences beginning with prepositional phrases and more granular features such as compound nouns and use of gerunds (Bernth and Gdaniec 2001; Underwood and Jongejan 2001; O'Brien 2006). O'Brien (2006) conducted an experiment that demonstrated that the presence of certain NTIs led to increased temporal effort for post-editors of RBMT output. However, the reduction of NTIs, as manifested through controlled language authoring, has since been shown not to have a positive impact on quality for NMT systems, though positive effects were demonstrated for SMT and RBMT (Marzouk and Hansen-Schirra 2019).

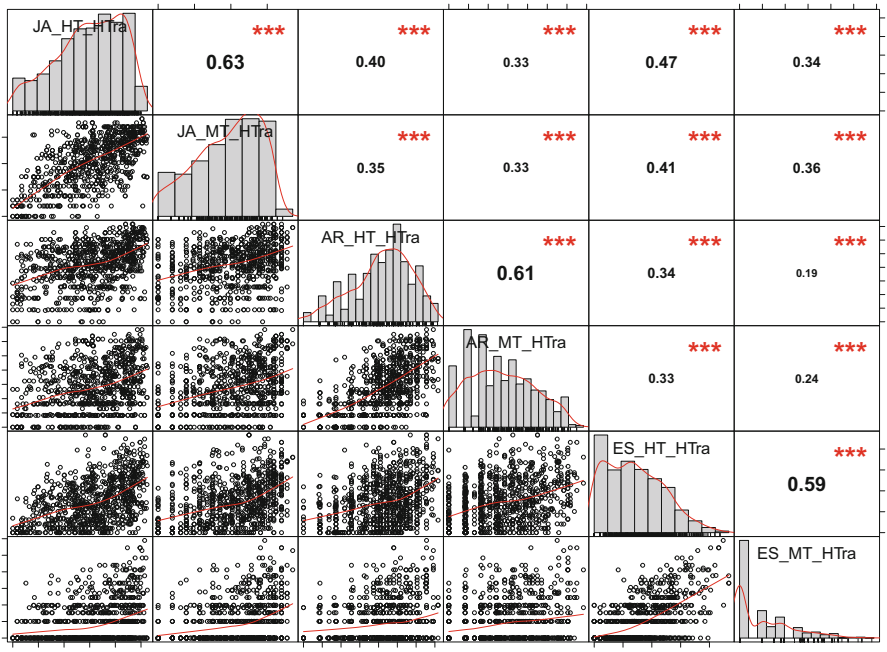


Fig. 1 Correlations among HTra

Previous research has also demonstrated that certain ST constructions are the origin not just of MT errors, but also human post-editing errors (Carl and Schaeffer 2014; Daems et al. 2014; Carl and Toledo Báez 2019). Even though HTra is not an error indicator, it is an indicator of disagreement, so it is interesting that humans and machines are similar at both the level of errors as well as translation choices. Since the common denominator for all three languages is the English ST, ST features are the likeliest source of translation entropy similarity and therefore the likeliest explanation for the moderate across-language correlations.

At the word level, which is the unit of analysis we adopt in the present study, it is also worth considering the concept of translation ambiguity. Translation ambiguity is determined experimentally by asking many participants to translate single words; words that are translated in multiple different ways are considered to be more translation-ambiguous than words that are only ever translated the same way by participants (Tokowicz 2014). Some research in the field has focused on *meaning* translation-ambiguous words (i.e., words that are translation-ambiguous because they are polysemous in the source language and would therefore have multiple semantically unrelated translations in the target language, such as the English 'nail' and its Spanish alternatives '*uña*' [fingernail] or '*clavo*' [metal nail]) and *form* translation-ambiguous words (i.e., words that have multiple synonymous translations in the target language, such as the English 'husband' and its Spanish alternatives '*esposo*' or '*marido*') (Tokowicz 2014).

Laxén and Lavaur (2010) found that *meaning* translation-ambiguous words took longer to process in translation recognition tasks than *form* translation-ambiguous words. It is true that the vast majority of translation ambiguity studies have only been conducted with single words in isolation, but the few studies that have used priming mechanisms to introduce context have shown that priming *meaning* translation-ambiguous words slightly reduced processing times while priming *form* translation-ambiguous did not (Eddington and Tokowicz 2013; Tokowicz 2014). However, priming did not cause the processing times of either type of translation-ambiguous words to drop enough that they might be considered similar to the processing times of non-translation-ambiguous words (Eddington and Tokowicz 2013; Tokowicz 2014). Bracken et al. (2017) take the dichotomous concept of *form/meaning* translation ambiguous-words and quantify it on a continuous scale. This measure could be replicated for the words in the multiLing texts in order to test its relationship with HTra.

Unfortunately, we do not have an annotation schema for ST tokens regarding translation ambiguity or polysemy in our data set and therefore cannot examine translation ambiguity here. However, it seems plausible that translation ambiguity could affect both MT and HT in a similar manner because translation-ambiguous tokens essentially lead to more possible translation alternatives, from which MT systems and human translators select the best option. It is most likely that *form* translation ambiguity is at play here, since context should be enough to disambiguate *meaning* translation-ambiguous alternatives, and translation ambiguity would most likely explain the strong *within*-language correlations reported here. Nonetheless, it is plausible that *meaning* translation-ambiguous words are exercising influence in these full-context translation sessions in subtler ways than is illustrated by homonyms such as 'nail.' Likewise, *form* translation ambiguity could conceivably explain our *across*-language correlations if certain ST words tend to be *form* translation-ambiguous for many language pairs. Further research should investigate translation ambiguity for multiple languages in these TPR-DB data sets to see if there are relationships between various types of translation-ambiguous words, production time, and HTra.

Whether we look to translatability indicators or translation ambiguity as possible explanations, when we observe the considerable strength of the correlations between MT_HTra and HT_HTra *across* languages, we are forced to ask ourselves what the common denominator is. If JA_MT_HTra moderately correlates with ES_HT_HTra, then what does a smattering of Japanese MT systems have in common with a group of thirty some-odd human translators in Spain? The glaring point of similitude is the ST that both parties were tasked with translating. This compels us to consider that, regardless of language or whether we are looking at MT systems or human translators, a significant proportion of the observed HTra can be attributed to ST features.

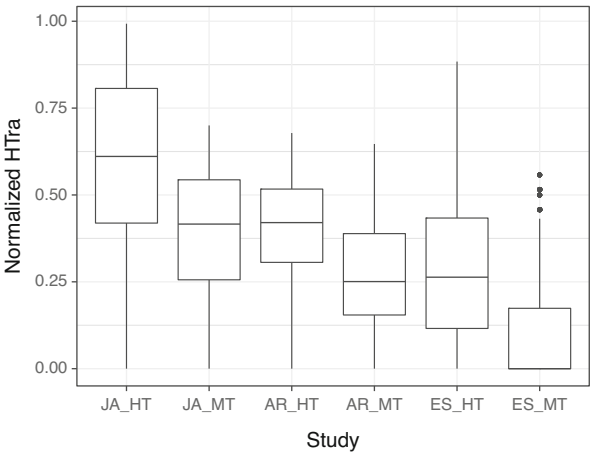


Fig. 2 Distribution of nHTra per study

5 Exploratory Analyses

In the previous section we suggested that much of the word translation entropy observed in our data could be linked to ST features. The aim of this section is to identify ST features that are associated with HTra. Using our normalized '*nHTra*' metric will allow us to compare different studies in absolute terms. Analyzing the data gleaned from all six studies in aggregate, Fig. 2 reveals a patent trend that holds for all three languages. In all three languages, nHTra values tend to be lower for MT than for HT. The figure also shows that the language that exhibits the highest nHTra values is Japanese, followed by Arabic and then Spanish.

We confirmed these visual conclusions by running a repeated measures ANOVA, which showed that there was a significant difference between HT_nHTra and MT_nHTra in all three languages. The same ANOVA confirmed that there was a significant difference in nHTra between Japanese, Arabic, and Spanish for both HT and MT. Figure 3 is an interaction plot of this repeated measures ANOVA. It echoes the information in Fig. 2, showing the same hierarchy in nHTra from Japanese down to Spanish, with MT always being lower than HT. The conclusion these results suggest is that the further a language is from English, in terms of linguistic distance the higher the average nHTra value will be.[8] In other words, translators tend to create a greater variety of translations the more distant the TL is from the SL.

For instance, Spanish is the closest to English because it is in the Indo-European language family, as is English. Unlike Spanish, both Arabic and Japanese are not at all in the same macrofamily as English. However, we could consider Japanese as being more distant from English than Arabic because Arabic uses an alphabetic writing system, whereas Japanese employs a much more complex writing system

[8]We assume that the relative distance to English is shortest for Spanish, then Arabic and Japanese, in this order, based on findings from research that quantified the linguistic distance of languages relative to English (Chiswick and Miller 2004; Isphording and Otten 2014).

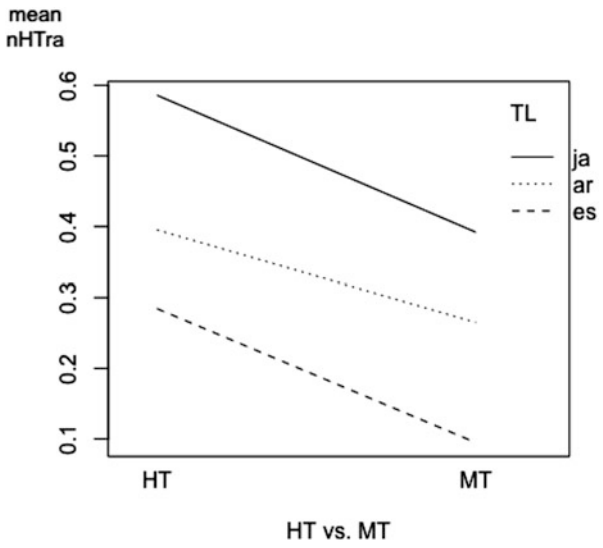


Fig. 3 Interaction plot for repeated measures ANOVA

(i.e., a combination of syllabic and logographic orthographies). This provides much more opportunity for variation in translation especially considering HTra's dependence on compositional alignments. This fits very well with research done by Carl et al. (this volume, Chap. 5) which suggests that HTra is largely a measure of *literality* in translation. They operationalize *literality* stating that a translation is considered literal to the degree that it fulfills the following three criteria: the word order is exactly the same in the ST and TT, each alignment group in the ST and TT have the same number of tokens, and each word in the ST has only one translation for a given context (ibid). A translation is less literal as it violates these criteria. HTra is mainly considered to be a measure of the third criterion, which is called *entrenchment*, though it is also dependent on the second criterion, which is known as *compositionality* (ibid.). It is important to note that literality must always be considered in the context of specific language pairs. Given our nHTra data, we could presumably order our language pairs from least to most literal as follows: EN>JA, EN>AR, and EN>ES.

Similar to our findings here, it has already been revealed that the HTra of European languages correlates better with other European languages than with Asian languages (Schaeffer et al. 2018). The fact that MT follows suit would seem to suggest that much of the source of variance in nHTra is indeed due to features of the ST, but it is also influenced considerably by the language pair. With this in mind, the goal of our exploratory analysis was to discover which aspects of the ST (i.e., *WordClass*, *Figurative*, *Voice*, and *Anaphora*) might be linked to nHTra.

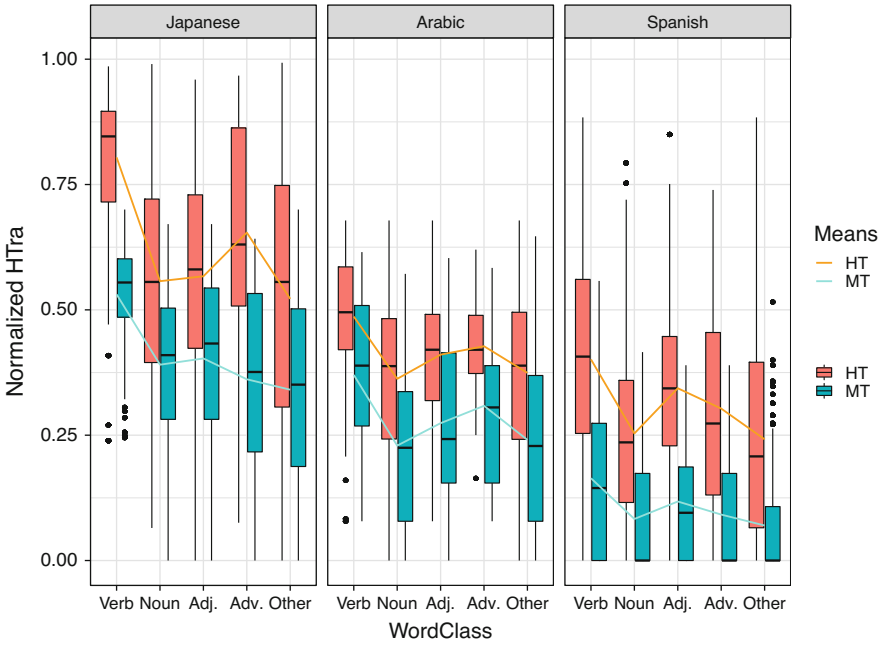


Fig. 4 nHTra per *WordClass*

Secondarily, we sought to see what influence our three respective target languages had on this as well.

5.1 Word Class

Figure 4 shows the consistent trend that nHTra is higher for verbs versus all other word classes, regardless of language or whether we are looking at MT or HT, and we verified that this difference was significant (see Appendix A). This result is consistent with psycholinguistic research on translation ambiguity. In gathering their translation ambiguity norms, where participants were asked to give translations for single words, Prior et al. found that "Word class predicted number of translations: Nouns had fewer translations than did verbs" (2007, 1029). This has held true in our data, even with the much richer context that translating an entire text provides.

When we talk about a 'richer context,' we are also referring to the fact that our data deals with verbs in context, with all of their morphological variants, and not just verb lemmas. This means that HTra is not just measuring lexical choice but is also measuring morphological variation. Due to morphological variation, we would expect in-context verbs to exhibit higher HTra values. Additionally, HTra also measures differences in how participants/aligners have chosen to align

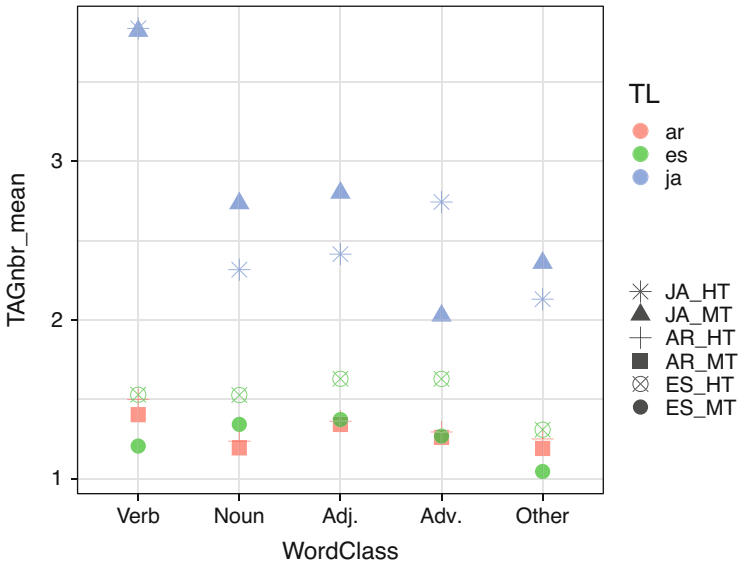


Fig. 5 Average number of tokens in a target alignment group

their translations and not just how they have chosen to translate. This could simply mean that translations of verbs tend to be less *entrenched* or *compositional* than translations of other word classes. We propose a couple of hypotheses for future research in the arena of word class and word translation entropy. First, we hypothesize that morphological variation impacts the HTra of verbs more than other word class categories. Second, we hypothesize that differences in alignment impact verbs more than other word class categories. The only evidence we offer for this first hypothesis is that, in the cases of morphologically rich languages like Spanish and Arabic, verbs are heavily inflected and can have over 40 different morphological variants (which could lead to many different TT alternatives, in other words, less *entrenched* translations). We will, however, provide some preliminary evidence for the second hypothesis.

Figure 5 shows that, in terms of the average number of words in a target alignment group (TAG), there is a considerably large disparity between verbs and other word classes for Japanese. This does not appear to be the case, however, for Spanish and Arabic. We can illustrate why we observe such a high number of words per TAG for Japanese verbs with an extreme, yet revealing alignment example from Japanese.

The highest nHTra value we observe in the Japanese data is 0.986, which belongs to the verb 'have' as in *Some of the most vulnerable countries of the world have contributed the least to climate change*. The word was mostly aligned together with 'contributed' and frequently with 'the least' as well. Close examination of the 18 cases where 'have contributed' was aligned as one alignment group revealed that none of the 18 cases shared the same translation and that there was an average

of 4.5 words[9] per TAG. One of the longest TAGs within those 18 cases was 原因_に_は_なっ_て_い_ませ_ん (literally 'have not become a factor'), which is a string of Noun + Particle + Particle + Verb + Particle + Verb + AuxiliaryVerb + AuxiliaryVerb, according to the morphological analyzer called MeCab (Kudo and Matsumoto 2002) utilized by TPR-DB. What makes Japanese unique is the existence of particles, which are suffixes that often indicate grammatical relations of words in a sentence. In addition, multiple verbs and auxiliary verbs can be combined at the morphological level in Japanese. These are the reasons why a simple ST phrase like "have contributed" can result in a very long TAG in Japanese, which in turn causes Japanese translations of verbs to be less *compositional*.

A more detailed examination of the alignments in each study is needed for a more conclusive interpretation of this data, but as we can see, there is some evidence that verbs are more subject to differing alignments than other word classes. However, this is language-dependent, as Japanese was the only language for which verbs had a significantly higher number of words per TAG (see Fig. 5). These preliminary results could be corroborated if a study were to realign translations consistently, perhaps using automatic word alignment. This is important so we can attribute increased HTra to decreased *compositionality* or *entrenchment* rather than inconsistent alignments. Additionally, future work should move us in the direction of distinguishing between two possible causes of decreased *entrenchment*: morphological variation versus lexical choice. We can do this by seeking to discover to what degree the increased HTra values for verbs (and other word classes) are due to morphological variation or lexical choice. This could also give us further perspective on how HTra relates to translation ambiguity studies in psycholinguistic research.

5.2 *Figurative*

Figure 6 shows that metaphoric tokens have higher nHTra values than fixed expression tokens and non-figurative tokens. Across studies, we found the same trend that the group Metaphoric has the highest mean of nHTra, followed by Fixed and then Other.

Although HT_nHTra being higher than MT_nHTra is a general tendency, the gap between the two is particularly noticeable for metaphoric tokens in Japanese and Spanish. This is most likely because machines translated these tokens literally, whereas human translators introduced more creative translations. Five out of 13 Japanese MT systems literally translated "cough up" from *British families have to cough up an extra £31300*, whereas none of the Japanese translators did. Translators

[9]Japanese target texts are morphologically divided into 'words' because there are no typed spaces in texts. This granular tokenization probably influences HTra, resulting in higher values compared to other languages.

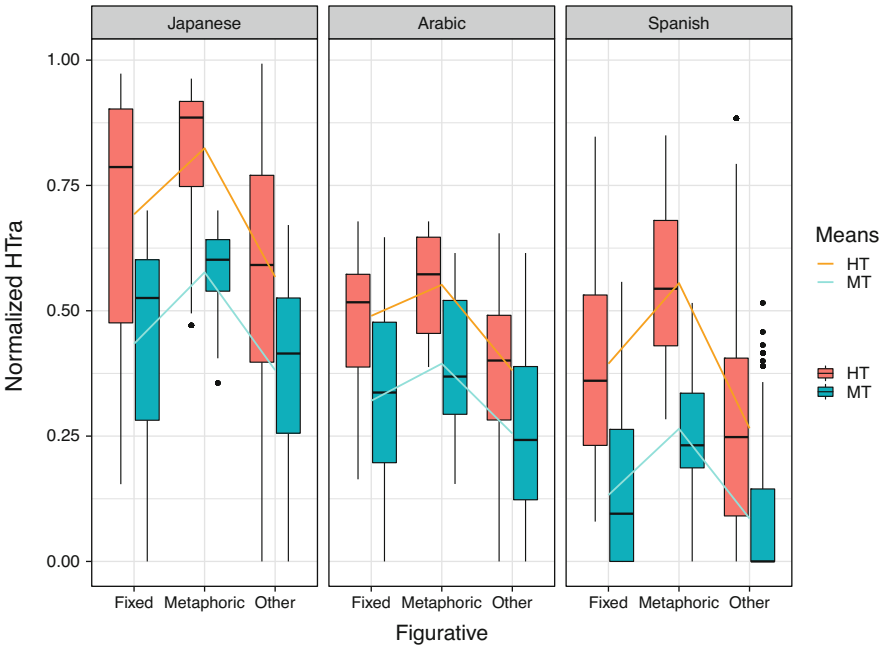


Fig. 6 nHTra per *Figurative*

must free themselves from the form and literal meaning of a metaphoric expression in order to convey the intended meaning, which leads to greater variation.

Further examination of these tokens revealed the possibility that tokens that are identified as Metaphoric and also one of the WordClass categories that tend to have high nHTra may exhibit higher HTra values than those that belong to only one category. For example, the metaphoric token of the highest nHTra value in the Japanese data was classified as Verb, which has outstandingly high nHTra (see Fig. 4). In case of Arabic and Spanish, such tokens were classified as Adjective. Although this is the second (in Spanish) or third (in Arabic) group in terms of nHTra values, the gap between its mean nHTra is not so different from that of verb compared to the Japanese data. Therefore, it is reasonable to think that we can observe a combined effect of two different categories (i.e., WordClass and Figurative) in tokens with very high nHTra values.

The fact that the Metaphoric group distinguishes itself from the other two groups so noticeably leads us back to the discussion on *meaning* translation ambiguity (see Sect. 4). Recall that Metaphoric refers to the tokens whose intended meaning deviates from their literal meaning. That is, these tokens can have different meanings depending on the context. This inherent semantic multiplicity, just like translation ambiguity, seems to be associated with HTra. In addition, the fact the Fixed group was also significantly different from the Other group in most studies (see Appendix A) supports the idea of *compositionality*, since Fixed tokens are strings

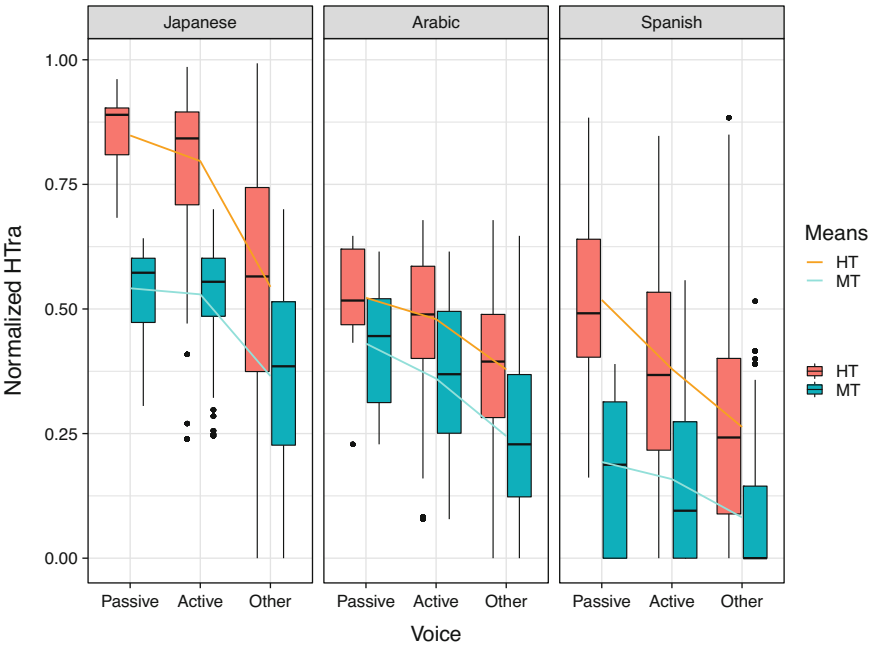


Fig. 7 nHTra per *Voice*

of words that are used idiomatically. One example of such a string, "in the wake of," is difficult to translate *compositionally* unless the target language has exactly the same saying using the same number of words.

5.3 Voice

Figure 7 demonstrates the trend that Japanese nHTra scores tend to be the highest nHTra values for both HT and MT in the *Voice* category's three levels (Passive, Active, and Other). Furthermore, the overall trend across all three languages indicates that the HT studies' passive voice expressions tend to have higher nHTra than in the MT studies. Statistical analysis showed that differences between passive and active voice tokens were only significant for ES_HT and AR_MT (see Appendix A).

Arabic passive voice does not have the copular verb 'to be' because it is agentless (i.e., it does not usually include the 'by' agentive predicate wherein the agent of the action is indicated). However, when translating an English passive sentence into Arabic, translators are faced with three options: (1) Transpose the English passive sentence into an Arabic active sentence; (2) Literally translate into an Arabic passive sentence and include the agentive predicate; (3) Translate into an Arabic passive sentence and ignore the agentive predicate. For example, Table 2 illustrates the

Table 2 Passive voice construction in Arabic

"The window was broken by the man"	
Arabic translations	**Back translations**
(1) كسر الرجل النافذة	The man broke the window.
(2) كُسرت النافذة بواسطة الرجل.	broken the window by the man
(3) كُسرت النافذة.	broken the window

three Arabic alternative translations for an English passive sentence. In the first two translations, the agent 'the man' is preserved in the meaning of the translation, whereas in the last translation the agent of the action is dropped to convey a more natural Arabic-like passive sentence.

Therefore, the existence of these features in English passive sentence structure, and the lack thereof in Arabic passive structure could be a source of inconsistency for Arabic MT systems and consequently, the generation of multiple, alternative translations of the same passive voice tokens. For Spanish, it is likely that many translators often prefer to transpose passive voice constructions into active voice. Couple this with the fact that there are multiple ways of constructing passive voice in Spanish (using either reflexive or past participle verb forms), and this could explain why the ES_HT passive tokens were significantly different from active tokens.

In the Japanese studies, the reason why the nHTra of Passive tokens was not different from the nHTra of Active tokens is probably due to the fact that main verbs and passive suffixes are oftentimes aligned together. For example, 'was given' in *he was given four life sentences* can be literally translated as 与えられた. Although morphological analyzers would tokenize this chain of a main verb, a passive auxiliary verb and a past auxiliary verb as 与え–られ–た, native speakers would conceive the chain as one verbal phrase, not three verbs. Therefore, 'was given' would be collectively aligned with 与えられた, just as a simple past 'gave' would be with 与えた.

5.4 Anaphora

Figure 8 shows the distribution of nHTra values for the *Anaphora* category across the three languages in both HT and MT. For HT, there are significant differences between the means for Anaphoric and Other in all three languages, while for MT the difference was only significant in Arabic (see Appendix A).

Consistent with the existing literature on the topic of anaphora and based on the ongoing research effort to improve MT system performance in recognizing anaphoric expressions (Safir 2004; Mitkov 2014; Voita et al. 2018), it was expected that our MT output would exhibit a tendency towards literal translation and consequently show no difference in nHTra values for Anaphoric tokens when compared to Other. This was clearly the case for Japanese. Since Japanese is a pro-

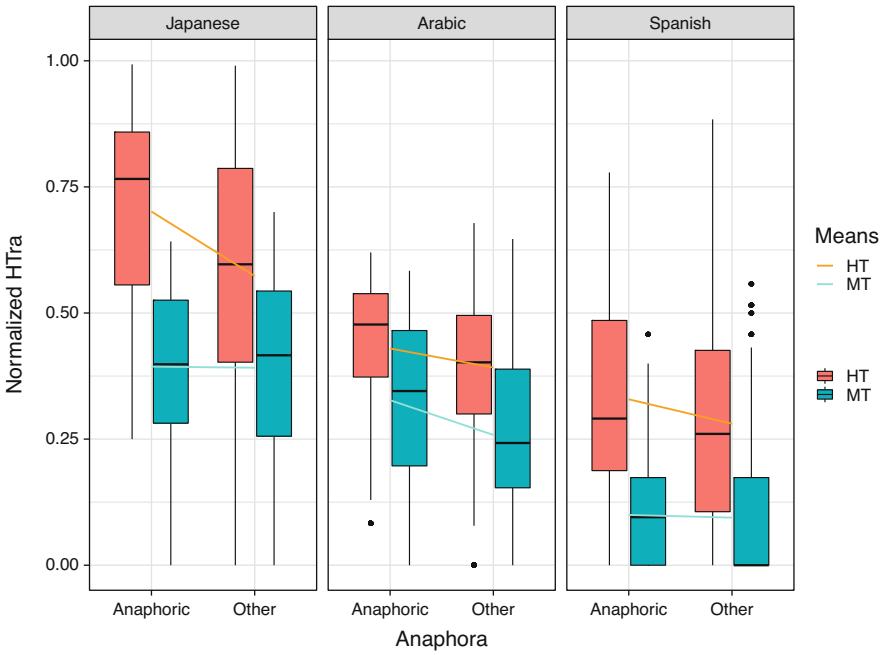


Fig. 8 nHTra per *Anaphora*

drop language, Japanese speakers not only omit subjects but also tend to avoid using pronouns altogether. In fact, excessive use of pronouns has often been considered a feature of *translationese* (Meldrum 2009). Accordingly, the human translators in the ENJA study often omitted them in their translations, whereas the MT systems translated pronouns explicitly, which lead to unnatural-sounding translations. The consequence of this for the JA_HT data is that the pronouns in the English ST were aligned together with other elements such as verbs (i.e., *non-compositional* or one-to-many alignments) while English pronouns were solely aligned to the Japanese pronouns in the JA_MT data (i.e., *compositional* or one-to-one alignment). This explains why Anaphoric has much higher nHTra values than Other for HT when there is no difference for MT.

The Spanish data revealed the same tendency as the Japanese data. The reason we observed higher nHTra values for Anaphoric tokens in ES_HT than ES_MT seems to be due to non-compositional alignments, just like Japanese. But even though Spanish is also a pro-drop language, what is actually causing the non-compositional alignments is probably the creativity of human translators. For example, if we examine ST token 76 in text 1 (the pronoun 'He' from *He will have to serve at least 30 years*, see Fig. 9), we find that the ES_HT study had seven different ways of translating the pronoun: *Norris*, *Colin Norris*, *Tendrá que* (He shall have to), *Va a tener que* (He will have to), *Deberá* (He must), *El Doctor* (The Doctor), *Él* (He). Four out of these seven possibilities are pronouns, proper nouns, or nouns while

SToken	SGroup	TGroup	target
He	He	Él	7
He	He	Norris	1
He	He_will	permanecerá	1
He	He_will	Tendrá	1
He	He_will	Va_a	1
He	He_will_at_least_30	El_doctor_por_lo_menos_30	1
He	He_will_have	deberá	1
He	He_will_have	habrá	1
He	He_will_have	lo_cual_supone_que_va_a_tener	1
He	He_will_have	Tendrá	6
He	He_will_have_to	Colin_Norris_tendrá	1
He	He_will_have_to	Deberá	1
He	He_will_have_to	Él_tendrá_que	2
He	He_will_have_to	Tendrá_que	5
He	He_will_have_to_serve	Deberá_cumplir_allí	1

Fig. 9 Spanish HT Anaphora Example

only three can be attributed to pronoun-verb alignments on account of Spanish's tendency to omit pronouns. The ES_MT systems, on the other hand, only had two possible translations, with only one of these being a verb phrase omitting the pronoun (*Tendrá que, Él*). This shows that humans are more likely to come up with creative solutions for translating anaphora, while MT is more homogeneous. In this case, human translations violate the third criterion of literality: *entrenchment*.

This increased level of creativity (or decreased *entrenchment*) shown by the translators in the ES_HT study consequently led to less *compositional* alignments. For instance, when "he" is aligned to *Colin Norris* or *El Doctor*, they both violate the second criterion of literality (i.e., *compositionality*). Despite there being only seven different translations for this token in the ES_HT study, there are 15 different SGroup-TGroup combinations (see Fig. 9). The ES_MT study does not have this issue because all items that were translated the same way were also aligned the same way. Additional investigation is needed to determine to what extent inconsistent alignments in the ES_HT data might have impacted these results.

In the Arabic data, however, AR_MT also showed statistically significant difference between Anaphora and Other, contrary to our expectation. Arabic is similar to the other two languages in the sense that subjects can be dropped,

Id	SToken	Task	SGroup	TGroup	target	count
48	he	MT	he_was_found_guilty	أدين	1	1
48	he	MT	he_was_found_guilty_of	تم_إدانته	1	1
48	he	MT	he_was_found_guilty	أدين	1	2
48	he	MT	he_was_found_guilty	أدين	1	2
48	he	MT	he	هو	1	1
48	he	MT	he_was_found	وجد	1	1
48	he	MT	he_was_found_guilty	قد_أدين	1	2
48	he	MT	he_was_found_guilty_of	أدين_بارتكاب	1	2
48	he	MT	he_was_found_guilty_of	أدين_بارتكاب	1	2
48	he	MT	he_was_found_guilty	قد_أدين	1	2
48	he	MT	he	انه	1	1

AR_MT Output

Id	SToken	Task	SGroup	TGroup	target	count
48	he	S	he	هذه_الممرض	1	1
48	he	P	he_was_found_guilty	و_تم_إدانته	1	1
48	he	T	he_was_found_guilty	تم_إدانة_أمام_المحكمة	1	1
48	he	S	he_guilty	ادانته	1	1
48	he	P	he_was_found_guilty	تم_إدانته	1	3
48	he	T	he	هو	1	1
48	he	S	he	الممرض	1	1
48	he	P	he_was_found_guilty	وقد_أدين	1	1
48	he	P	he_was_found_guilty_of	تم_إدانته	1	3
48	he	T	he	نوريس	1	1
48	he	P	he_was_found_guilty_of	تم_إدانته	1	3

AR_HT Output

Fig. 10 Arabic Anaphora example

but the Arabic MT systems provided a more divergent (less *entrenched*) array of translations for Anaphoric tokens. Figure 10 compares AR_MT and AR_HT for the ANAPHORIC token 'he' as in *Yesterday, he was found guilty of four counts of murder following a long trial*. We find that only two of the 12 MT systems have resorted to a literal translation of 'he' as هو and انه while the rest of them have omitted the pronoun. In AR_HT, there are four instances where 'he' has been translated explicitly, but translators have produced a greater variety of less *compositional* alternatives: ranging from literally translating the pronoun as هو (he), substituting the pronoun with noun equivalents such as الممرض (the nurse) or نوريس (Norris) to completely dropping the pronoun and providing a more dynamically equivalent solution such as تم إدانته (he was condemned).

It is evident that HT exhibited greater creativity than MT in the Arabic data, similar to findings in the Japanese and Spanish data, but Arabic MT systems seemed to have produced more creative translations than JA_MT and ES_MT. It is unclear at this point why this is the case since we do not know how each MT system used in this study was trained, or with what materials each was trained. Nonetheless, our study has demonstrated that higher nHTra values result from translations that violate literality criteria such as *compositionality* and *entrenchment*. Mismatches in the number of tokens in the ST and TT alignment groups violated *compositionality* (this was potentially obscured by inconsistent alignments), and higher lexical variation across translations violated *entrenchment*.

6 Concluding Remarks

This is the first study to the best of our knowledge where the output from multiple MT systems has been source-text aligned in order to calculate word translation entropy and compare it with HT. This allowed us to observe remarkable correlations

between HT_HTra and MT_HTra within and across three diverse languages. Given that we still observed significant correlations of HTra *across* languages, many of the similarities between humans and machines seem to stem from features of the source text.

Normalization of HTra values (nHTra) allowed us to compare word translation entropy values across the six studies in absolute terms, which in turn enabled us to explore ST features as they related to each language pair. In examining word class categories, we found that verbs had significantly higher nHTra values than nouns, adjectives, adverbs, and others, for both HT and MT in all languages. Regarding figurative expressions, metaphoric expressions were outstandingly different from non-figurative expressions, again for both HT and MT in all languages. For voice, however, passive voice was only significantly different from active voice in AR_MT and ES_MT. Lastly, anaphoric tokens were different from all other tokens in all studies except for JA_MT and ES_MT, which highlighted how MT systems approach the problem of anaphora more homogeneously and more literally (i.e., more *entrenched* and more *compositional*) than humans. The takeaway from these exploratory analyses is that each syntactic/semantic category investigated in this study seems to influence nHTra in the same direction regardless of target language and regardless of the mode being MT or HT.

Using the lens of HTra, our study has shed light on the relationship between MT and HT, given us insights on the nature of HTra itself, and opened up avenues for future research. First, we introduced the nHTra metric, which should be valuable for comparing separate TPR-DB studies. Second, more research is needed to investigate the impact of alignment and morphology on HTra. Attention in this area should also focus on the impact that word class has on HTra. Third, further research on the relationship between HTra and words that have been normed for translation ambiguity—such as the Spanish-English and English-Spanish words in Prior et al.'s data set—would also be very welcome. Carl (this volume, Chapter 14) attempts to delve into this topic. Finally, the present study has shown the power of analyzing and comparing how humans translate in relation to how multiple MT systems translate, instead of a single MT system; more research using this method should prove fruitful.

Acknowledgments More thanks than we can give to Kristin Yeager, the Manager of Statistical Consulting at Kent State University Libraries. Also, a special thanks to Andrew Tucker for coming up with our project's codename.

A LMEM

For our exploratory analyses, we examined the data using linear mixed-effects models (LMEM) to see whether there is a significant difference among the groups in each ST linguistic feature in terms of the nHTra values. LMEM is a better choice than a t-test because the nHTra values were uniquely distributed across studies with a high frequency of zero values.

We first ran LMEM using the "lme4" package on RStudio with all the data points from six studies. We set nHTra as the dependent variable, Text (i.e., 1–6) and TextId (i.e., a combination of Text number and word ID as in "1_5" for the fifth word of Text 1) as a crossed random effect, and one of our syntactic/semantic categories as a fixed effect. We examined a three-way interaction effect among the category, the target language, and the distinction between HT and MT. We then ran a similar analysis with the data narrowed down by the target language. For *WordClass*, Verb was significantly different from the other four groups in all six studies. For *Figurative*, the three groups (Metaphoric, Fixed and Other) were significantly different from each other in all studies except the case of Fixed vs. Other in JA_MT. For *Voice*, we only observed a significant difference between Passive and Active in AR_MT and ES_HT. And for *Anaphora*, we found significant difference between Anaphoric and Other in all studies except JA_MT and ES_MT.

References

Almazroei SA, Ogawa H, Gilbert D (2019) Investigating correlations between human translation and MT output. In: Proceedings of the second MEMENTO workshop on modelling parameters of cognitive effort in translation production, European association for machine translation, pp 11–13, MT Summit 2019, Second MEMENTO workshop on modelling parameters of cognitive effort in translation production; Dublin, 20 Aug 2019

Bernth A, Gdaniec C (2001) MTranslatability. Mach Transl 16(3):175–218

Bracken J, Degani T, Eddington C, Tokowicz N (2017) Translation semantic variability: how semantic relatedness affects learning of translation-ambiguous words. Biling Lang Congn 20(4):783–794

Carl M, Schaeffer M (2014) Word transition entropy as an indicator for expected machine translation quality. In: Proceedings of the workshop on automatic and manual metrics for operational translation evaluation, MTE 2014, European Language Resources Association, pp 45–50, LREC 2014 workshop on automatic and manual metrics for operational translation evaluation; Reykjavik, 26 May 2014

Carl M, Schaeffer MJ (2017) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. HERMES J Lang Commun Bus (56):43–57

Carl M, Toledo Báez MC (2019) Machine translation errors and the translation process: a study across different languages. J Specialised Transl (31):107–132

Carl M, Aizawa A, Yamada M (2016a) English-to-Japanese translation vs. dictation vs. post-editing: comparing translation modes in a multilingual setting. In: Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European Language Resources Association, pp 4024–4031, LREC 2016, Tenth international conference on language resources and evaluation; Portorož, 23–28 May 2016

Carl M, Schaeffer M, Bangalore S (2016b) The CRITT translation process research database. In: Carl M, Schaeffer M, Bangalore S (eds) New directions in empirical translation process research. Springer, pp 13–54

Chiswick BR, Miller PW (2004) Linguistic distance: a quantitative measure of the distance between English and other languages. IZA Discussion Papers (1246)

Daems J, Macken L, Vandepitte S (2014) On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship. In: Proceedings of the ninth international conference on language resources and evaluation (LREC 2014). European Language Resources Association, pp 62–66, Ninth International conference on language resources and evaluation; Reykjavik, 26–31 May 2014

Eddington CM, Tokowicz N (2013) Examining English-German translation ambiguity using primed translation recognition. Biling Lang Congn 16(2):442–457

Germann U (2008) Yawat: yet another word alignment tool. In: Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: demo session. Association for Computational Linguistics, pp 20–23, 46th annual meeting of the association for computational linguistics on human language technologies: Demo Session; Columbus, 16 June 2008

Guerberof Arenas A (2009) Productivity and quality in MT post-editing. In: Proceedings of XII MT summit workshop: beyond translation memories. International Association for Machine Translation, XII MT Summit Workshop: beyond translation memories, Ottawa, 26–30 Aug 2009

Isphording IE, Otten S (2013) The costs of Babylon–linguistic distance in applied economics. Review of International Econ 21(2):354–369

Isphording IE, Otten S (2014) Linguistic barriers in the destination language acquisition of immigrants. J Econ Behav Organ 105:30–50

Jia Y, Carl M, Wang X (2019) How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. J Specialised Transl (31):61–86

Koponen M, Salmi L, Nikulin M (2019) A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. Mach Transl 33(1–2):61–90

Kudo T, Matsumoto Y (2002) Japanese dependency analysis using cascaded chunking. In: Proceedings of the 6th conference on natural language learning. Association for Computational Linguistics, pp 1–7, 6th conference on natural language learning; Taipei, Aug 31 to 1 Sept 2002

Laxén J, Lavaur JM (2010) The role of semantics in translation recognition: Effects of number of translations, dominance of translations and semantic relatedness of multiple translations. Biling Lang Congn 13(2):157–183

Läubli S, Castilho S, Neubig G, Sennrich R, Shen Q, Toral A (2020) A set of recommendations for assessing human-machine parity in language translation. J Artif Intell Res 67:653–672

Marzouk S, Hansen-Schirra S (2019) Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. Mach Transl 33(1–2):179–203

Meldrum YF (2009) Translationese-specific linguistic characteristics: a corpus-based study of contemporary Japanese translationese. Honyaku kenkyū e no shōtai 3:105–132

Mesa-Lao B (2014) Gaze behaviour on source texts: an exploratory study comparing translation and post-editing. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, pp 219–245

Mitkov R (2014) Anaphora resolution. Routledge

Moorkens J, O'Brien S (2015) Post-editing evaluations: trade-offs between novice and professional participants. In: Proceedings of the 18th annual conference of the European association for machine translation. European Association for Machine Translation, pp 75–81, 18th annual conference of the European association for machine translation; Antalya, 11–13 May 2015

O'Brien S (2006) Machine-translatability and post-editing effort: an empirical study using translog and choice network analysis. Doctoral Thesis, Dublin City University

Prior A, MacWhinney B, Kroll JF (2007) Translation norms for English and Spanish: the role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. Behav Res Methods 39(4):1029–1038

Safir KJ (2004) The syntax of anaphora. Oxford University Press

Schaeffer MJ, Oster K, Nitzke J, Tardel A, Gros AK, Gutermuth S, Hansen-Schirra S, Carl M (2018) Cross-linguistic (dis)similarities in translation: process and product. In: Book of abstracts: using corpora in contrastive and translation studies conference (5th edition), centre for English corpus linguistics, pp 90–92, using corpora in contrastive and translation studies conference (5th edition); Louvain-la-Neuve, 12–14 Sept 2018

Schwartz L (2018) The history and promise of machine translation. In: Lacruz I, Jääskeläinen R (eds) Innovation and expansion in translation process research. John Benjamins Publishing Company, pp 161–190

Shannon CE (1951) Prediction and entropy of printed English. Bell Syst Tech J 30(1):50–64
Tokowicz N (2014) Translation ambiguity affects language processing, learning, and representation. In: Selected proceedings of the 2012 second language research forum. Cascadilla Press Somerville, pp 170–180, 31st Second Language Research Forum; Pittsburgh, 18–21 Oct 2012
Underwood N, Jongejan B (2001) Translatability checker: a tool to help decide whether to use MT. In: Proceedings of MT summit VIII: machine translation in the information age, European association for machine translation, pp 363–368, MT summit VIII: machine translation in the information age; Santiago de Compostela, 18–22 Sept 2001
Vanroy B, De Clercq O, Macken L (2019) Correlating process and product data to get an insight into translation difficulty. Perspectives 27(6):924–941
Voita E, Serdyukov P, Sennrich R, Titov I (2018) Context-aware neural machine translation learns anaphora resolution. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), Association for computational linguistics, pp 1264–1274, 56th Annual meeting of the association for computational linguistics; Melbourne, 15–20 July 2018

Entropy and Eye Movement: A Micro-analysis of Information Processing in Activity Units During the Translation Process

Yuxiang Wei

Abstract This chapter intends to analyze the cognitive processing of information by examining HTra within the AUs of the translation process. As a measure of the degree of uncertainty involved in translation choices, HTra reflects translation ambiguity and has been used as a predictor variable for the cognitive effort in the translator's selection of a TT item among the co-activated alternatives. In this regard, the present chapter starts from a theoretical discussion on entropy, lexical activation, cognitive effort, and the dynamic change of entropy in the mental processes of the translation choice, exploring the conceptual basis on which entropy can represent cognitive load and describe the assumed mental states, in translation. It then examines in an empirical manner the entropy values (i.e., HTra) of the words which are fixated in AUs on the basis of the CRITT TPR-DB, to shed light on the manner in which contextual information is cognitively processed. Results show that the AUs which are associated with the resolution of problems arising from high-entropy words tend to include scanpaths where low-entropy words come into play, which indicates that the cognitive processing of a highly translation-ambiguous item is facilitated by the contextual information which is provided by items of low translation-ambiguity levels. In the meantime, the chapter also analyzes in detail the process in which the average entropy of the scanpath is updated as the translator processes contextual information within the AU.

Keywords Word translation entropy · Activity unit · Translation process research · Ambiguity · Eye movement · Lexical activation · Linguistic context

Y. Wei (✉)
Dublin City University, Dublin, Ireland
e-mail: yuxiang.wei@link.cuhk.edu.hk

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_7

1 Introduction

In TPR, studies have shown that translation behavior tends to be affected by the number of alternatives into which individual ST words can be translated. For example, eye-key span (i.e., the time interval between the translator's fixation on an ST word and the keyboard input of its translation) has been shown to be longer for words which have many different translation alternatives than those corresponding to only one possible translation (Dragsted and Hansen 2008; Dragsted 2010). The influence of the number of translation alternatives on behavior is also shown in studies using decontextualized single words (e.g., Tokowicz and Kroll 2007), and research in psycholinguistics demonstrates that bilinguals often activate lexical information nonselectively from both languages during word recognition (Brysbaert 1998; Dijkstra et al. 2000; De Bruijn et al. 2001; Marian et al. 2003; Schwartz and Kroll 2006).

While it can be debatable as to how the bilingual mental lexicon is organized and which model best reflects the translating mind (e.g., Revised Hierarchical Model, see Kroll and Stewart 1994; Bilingual Interactive Activation Model, see Dijkstra and Van Heuven 1998, 2002), translation behavior examined from an empirical perspective seems to consistently reflect an evident impact from what is sometimes called "translation ambiguity" (i.e., the existence of multiple translations for the same ST word; see, e.g., Prior et al. 2011) in experimental studies of the translation process.

In this regard, an entropy-based predictor variable which mathematically describes the degree of uncertainty in translation choices, namely, HTra (see Sect. 2.1), has been considered a better measure for the variation of the translation alternatives than simply counting the number of these alternatives (Bangalore et al. 2016). Further studies on HTra show a significant positive effect on different measures of effort, including, among others, first fixation duration and total reading time (e.g., Schaeffer et al. 2016b), which indicates that words with higher HTra values tend to be more difficult to translate (Carl et al. 2019).

This chapter intends to analyze the cognitive processing of information by examining the HTra values of words which are fixated in AUs (see Sect. 2.4 for definition), and while doing so, it also examines the dynamic change of entropy during the process in which the translator resolves a problem associated with a high-HTra word. Starting from a general discussion on entropy, uncertainty, cognitive load, bilingual lexical activation, and a systems theory perspective of the translation process, it explores the conceptual basis on which entropy and entropy reduction can describe the mental processes of translation, and provides a theoretical background in Sect. 2 for the following analysis. Section 3 presents a preliminary discussion on the number of fixations in AUs and the corresponding average HTra of the scanpaths. An example of scanpath is given to show how the average entropy value is decreased to a medium level when the number of fixations is relatively large.

Section 4 takes the same scanpath into a wider scope of the translation process, presenting a detailed analysis as to how the translation ambiguity in the example

is resolved, in view of the HTra values of the individual words fixated within the relevant AUs. In the meantime, this section shows statistically that the pattern of HTra values as displayed in the example is largely representative of a general feature of AUs that are associated with the resolution of translation ambiguity in the same high-HTra word. Four additional examples are then demonstrated in detail where the same pattern is displayed.

Regarding these conceptual and empirical explorations in the chapter, a general discussion is provided in Sect. 5, summarizing the study and explaining its implications, while Sect. 6 concludes the chapter with possible avenues for future research pertaining to entropy and eye movement.

2 Theoretical Background

2.1 Entropy and Uncertainty

The entropy discussed here is largely relevant to the mathematical expression of information in information theory (Shannon 1948; Wiener 1948; Shannon and Weaver 1949; Wiener 1954), a theory which has been influential for a variety of areas (e.g., Bazzanella 2011) including translation studies (TS), where translation—especially in the "equivalence" paradigm—has often been considered a process of transcoding the ST content into the target form (e.g., Nida and Taber 1969), although the mathematical expressions, lying at the core of Shannon's theory of communication, seem much less discussed in TS.

In mathematical terms, among the properties of what Shannon calls an information channel (e.g., redundancy, information load, etc.), largely fundamental is the definition of information by probabilistic means, measuring uncertainty, surprisal, and freedom of choice. The information (i.e., surprisal) of an item is defined as the negative logarithm of its probability, which reflects the intuition that "unpredictable items should carry a large amount of information while predictable items should not" (Collins 2014, 652). If a source of message generates a discrete stochastic variable, then the amount of information for this message is represented by the weighted sum of the surprisal for each individual value that the variable can take, i.e., the entropy. This entropy would be maximum if all probabilities are equal.

This is indicative of the number of informational units needed for message encoding[1] and is considered the "selectional information-content" (MacKay 1969) in communication, for which the main point of interest is the "relative improbability of a message given an ensemble of messages" (Kockelman 2013, 116). In addition to the number of informational units in encoding, entropy also measures the *uncertainty* involved in estimating the value which a variable can take, i.e., the freedom of choice when a message is chosen (Shannon and Weaver 1949).

[1]Specifically, the value of entropy is the expectation of the bits of information for the message.

For the process of translation—a task frequently described as essentially "a chain of decision-making activities" (Angelone 2010, 17; see also Levý 1967; Tirkkonen-Condit 1993; Hervey et al. 1995)—the relative improbability of a translation option for a particular ST item, given an ensemble of translation options for the same item, can perhaps be defined in a similar manner. In this regard, Carl et al. (2016) propose an entropy-based metric for word translations, namely, HTra, as a description of the word translation choices at a given point of the ST, and syntactic choice entropy (HCross), as a representation of "local cross-lingual distortion."[2] For HTra, Schaeffer et al. (2016b) provide a succinct description:

> Word translation entropy describes the degree of uncertainty regarding which lexical TT item(s) are chosen given the sample of alternative translations for a single ST word: if the probabilities are distributed equally over a large number of items, the word translation entropy is high and there is a large degree of uncertainty regarding the outcome of the translation process. If, however, the probability distribution falls onto just one or a few items, entropy is low and the certainty of the TT item(s) to be chosen is high. (191).

On this basis, HTra has been subsequently used as a predictor variable in some studies of TPR (see also Lacruz et al., this volume, Chap. 11) and considered a perhaps "better reflection of the cognitive environment" of the translation process than simply counting the number of possible translation alternatives for a given word in the ST (Bangalore et al. 2016, 214).

Empirical studies have shown, not surprisingly, that the value of HTra has a statistically significant impact on many aspects of translation behavior, e.g., word production duration, first fixation duration, the probability of a fixation, and total reading time (Schaeffer et al. 2016b; Carl and Schaeffer 2017b).

2.2 Relative Entropy and Cognitive Load

Given its definition, and results from empirical studies, it is easy to hypothesize that some aspects of this entropy-based metric could be an indicator of cognitive load in translating. Bangalore et al. (2016), for example, propose using entropy as a measure of the translator's cognitive effort which is expended in making choices during translation. While their argument in this regard—a well-justified one indeed—is that entropy captures the *weight* of each alternative to describe probability distribution, the discussion in the present chapter is from a slightly different perspective, focusing on the dynamic *change* of probability distribution during the translator's decision-making.

In fact, similar use of Shannon's equation (i.e., information entropy) has not been infrequent in studies of monolingual processing in psycholinguistics (see, e.g.,

[2] A detailed description as to how HTra and HCross are calculated can be found in Carl et al. (2016). Some other chapters in the present book (e.g., Lacruz et al., Chap. 11; Heilmann and Llorca-Bofí, Chap. 8) also include brief and succinct illustrations of the equation for HTra.

Levy 2013; Levy and Gibson 2013), especially in relation to structural ambiguity resolution. On the basis of Attneave's (1959) application of information theory to psychology, Hale (2001) argues that the surprisal (i.e., Shannon information content, or self-information) of a word in its context can be used as a quantification of the cognitive effort which is required to process this word along the sentence. In his view, incremental sentence comprehension is a step-by-step disconfirmation of possible phrase-structural analyses for the sentence, and cognitive load can thus be interpreted as the combined difficulty of "disconfirming all disconfirmable structures at a given word" (ibid).

In expectation-based models of real-time sentence comprehension (in terms of parsing), processing difficulty (or measurable disruption) can arise from a sufficiently unexpected input which causes a shift in resource allocation "to various alternatives in the face of uncertainty" (Levy 2013, 144). The size of this shift in resource allocation, in, e.g., Levy's (2008) analysis, can be measured by the *change* (or update, to use Levy's word) of the probability distribution over possible interpretations after the current word is processed. In mathematical terms, this difference in probability distribution would be expressed by the *relative entropy* (also known as Kullback-Leibler divergence; see Kullback and Leibler 1951; Kullback 1959) of the updated distribution with respect to the old distribution.

Regarding the process of translating, perhaps a similar perspective can be adopted for the translation of each ST item. Upon encounter of a particular ST item, possible translations for this item are likely to be subliminally co-activated (Grosjean 1997; Macizo and Bajo 2006; Schwartz and Kroll 2006; Ruiz et al. 2008; Wu and Thierry 2012; Balling 2014; Bangalore et al. 2016; Schaeffer et al. 2016b; Carl et al. 2019),[3] and a higher entropy value (i.e., HTra) for this item indicates a higher level of uncertainty in the translator's decision-making in choosing from these translation alternatives. This increased level of uncertainty may be the result of a larger number of translation alternatives which are activated, or a lack of highly likely choices from these alternatives, or both (see also Sect. 2.1).

In this respect, the activation within the bilingual lexicon is often assumed, e.g., in the Bilingual Interactive Activation Plus model (BIA+; see Dijkstra and Van Heuven 2002), to be directly affected by surrounding linguistic context which provides lexical, syntactic, and semantic information. This assumption seems to be supported by experimental studies in bilingualism (e.g., Schwartz and Kroll 2006). Similarly, in the reordered access model (see, e.g., Duffy et al. 2001), the relative frequency of the alternative meanings of an ambiguous word (which often correspond to different translation alternatives) determines the order (or relative speed) in which these meanings are activated and compete for selection, while this activation can be reordered by a strong biasing context. For structural

[3]This is suggested by evidence from many studies in both TPR and bilingualism, although the *extent* of this activation is often relevant to language proficiency, particularly regarding a nondominant language (e.g., experiments also show that semantic priming is stronger for those with higher proficiency; see Faveau and Segalowitz 1983).

building framework (Gernsbacher 1990; Gernsbacher 1997), the mental processes involve a combination of enhancement and suppression, in which the activation of contextually relevant information is enhanced, whereas a suppression effect reduces the activation of information which is irrelevant to the context. Although there does not seem to be "a uniform theoretical account" as to "how sentence context exerts its influence on bilingual lexical access" (Schwartz and Kroll 2006, 209), the fact that linguistic context does aid the interpretation of ambiguous words, and also reduces the number of appropriate translations for an ST word, is perhaps without much disagreement.[4]

In terms of probabilities which are observed in the text, this results in a distribution where the different translation choices for a given ST word are not equally probable (i.e., not equally appropriate for the context). This distribution of observed probabilities can be described by its entropy value and approximated from the translation choices made by different translators regarding the ST word (i.e., HTra value). As essentially a statistical feature of the translation product, a higher entropy (i.e., HTra) in the observed translations indicates a higher level of uncertainty in the selection among the possible TT items in the sample.

In terms of the mental processes, the translator can perhaps be assumed to engage in an activation pattern where the activated items receive different degrees of priority for resource allocation (or according to the reordered access model, a pattern where the items are activated in a certain order in terms of relative time course). This pattern is largely affected by linguistic context[5] and can appear as a distribution of probabilities which are observed in the produced translation. During the subsequent selection process, the pattern would be dynamically updated to arrive at a translation choice (see below), while the distribution of probabilities which can describe this new pattern is updated accordingly. In a similar manner to the probability distributions observed in the textual material, the activation pattern in a particular mental state can also be represented by entropy (i.e., the distribution of resource allocation to various alternatives which are activated, or, perhaps equivalently, the distribution of temporary probabilities with which the activated candidates are to be selected), in turn indicating the (temporary) uncertainty level in the mental state regarding the selection among the activated candidates.

As the translator attempts to resolve the uncertainty arising from a high-entropy ST item (i.e., high HTra as observed in the translation)—typically by making use of additional contextual information—a change (i.e., update) occurs in the pattern of

[4]The disagreement in this regard is around *when* (i.e., how early), rather than *whether*, sentence context exerts its effect. See also the next footnote below, where more detail is described.

[5]It can be debatable as to how early this effect is exerted—some context-dependent accounts in monolingual processing argue that the conceptual representations which are built along the sentence have an early effect on lexical access, while in some context-independent accounts, sentence context influences not the initial activation but the subsequent selection process after the word has been accessed. Despite this disagreement, however, it seems that the effect of context concerning the pattern here, which is subsequently updated in the mental processes, can be safely assumed.

the co-activated possible (TT) translations for this ST item, as new input is received from the context (i.e., information which is provided by the surrounding items in the textual material). Some choices would become more probable while others less so, as the translator proposes and evaluates solutions on the basis of contextual information received from scrutinizing the items preceding or following the current high-entropy item. If further contextual information is inputted as a further step to arrive at a decision, the pattern continues to change.

Consequently, as the translator proceeds with the selection process for a suitable TT item, the allocation of cognitive resources becomes less evenly distributed over the various alternatives (or in other words, the probabilities which describe the updated activation pattern become less evenly distributed), concentrating on the items which are more likely than others to be chosen by the translator. Since the value of entropy represents—among other aspects—the extent to which probabilities are evenly distributed (see Sect. 2.1), this means that the updated entropy (in the assumed mental state) decreases, together with a decrease in the uncertainty involved, and the choice of a TT option for the ST item becomes more straightforward. When the resulting entropy decreases to zero, the choice would be restricted to only one option (i.e., the choice made by the translator).

If Levy's formulation (see above) can be adopted, the size of the shift of resource allocation between two certain points during this process would be represented by the *relative entropy* of the updated pattern with respect to the previous pattern.[6] In this view, the cognitive effort which is expended in the entire translation selection process can therefore be quantified via the relative entropy of the pattern when the translation choice is made at the end of the process, with respect to the beginning of the process when the items are activated. Mathematically, the value of this relative entropy would be equal to the surprisal of the TT item which is chosen by the translator.[7]

[6]From an information theory perspective, this relative entropy indicates the penalty incurred from encoding the new pattern with the old one.

[7]This is because the updated distribution is concentrated on one single item (i.e., the item chosen by the translator), whose probability equals 1. According to the definition of Kullback-Leibler divergence (i.e., relative entropy), the divergence of distribution $q(x)$ from $p(x)$ equals the expectation of the logarithmic difference between $q(x)$ and $p(x)$, with the expectation taken using $q(x)$. As $q(x)$ is concentrated on one single item (X) with a probability of 1, i.e., $q(X) = 1$, while the probabilities for all other items are 0, the divergence would be: $D(q||p) = q(X)\,[\log q(X) - \log p(X)] = -\log p(X)$. If the activation is modulated by the frequency of meanings/translations, the $p(X)$ which describes the mental state of activation would be the same as the probability observed in the text. Therefore, $-\log p(X)$ would be equivalent to the surprisal of the TT item (X) which is chosen by the translator. As mentioned above, the concept of surprisal is also termed, in different contexts, as information, self-information, or Shannon information content, all referring to essentially the same mathematical equation (i.e., the negative logarithm of probability). In some chapters of this book, the surprisal regarding a particular translation item is called word translation information and denoted by ITra (see, e.g., Heilmann and Llorca-Bofí this volume, Chap. 8; Carl this volume, Chap. 5).

If however the decrease of entropy value is used as the measurement of cognitive effort in this selection process, then after the choice is made (i.e., after the entropy decreases to zero), this decrease would equal the initial entropy when all the TT candidates are activated given the ST item (i.e., the entropy in the mental state between activation and selection).[8]

Here, whether relative entropy or the decrease of entropy is a better measurement of cognitive effort in the selection process is not the main point of focus. From another perspective, this chapter shows that within an AU (see Sect. 2.4), the above-mentioned process of integrating contextual information in response to a highly translation-ambiguous (i.e., high-HTra) word tends to include a process where the input of surrounding words at low translation-ambiguity levels comes into play. In other words, during the translator's decision-making, the update of probability distribution of the translation candidates (i.e., the shift of resource allocation to various translations) for a high-entropy (i.e., highly translation ambiguous) word seems to be facilitated by the information associated with low-entropy words in the context.

2.3 Systems Theory Perspective of the Translation Process

The above description of the dynamic change in probability distribution, borrowing Levy's (2008) formulation of resource-allocation processing difficulty, seems consistent with many aspects of a systems theory perspective on the translation process, a framework proposed in Carl et al. (2019) where the use of entropy as a description of the translation process is more inclined to the way in which entropy is defined in systems theory (or thermodynamics) rather than in information theory. From that perspective, entropy refers to the amount of disorder in a system, and the process of translating is considered "a hierarchy of interacting word and phrase translation systems which organise and integrate as dissipative structures." The expenditure of cognitive effort, or "average energy," to arrive at a translation solution is to decrease the internal entropy (i.e., disorder) of the system.

If an ST item together with all the possible translation alternatives of this item can be considered a word or phrase translation system, this chapter shows that for a highly entropic system, the decrease of its internal entropy tends to involve surrounding low-entropy systems with which the current system interacts.

[8]If the activation is assumed to be modulated by context and the frequency of the different meanings (e.g., in the reordered access model), this initial entropy value in the mental state can be, albeit arguably, considered to be equal to the entropy value which is observed in the text (i.e., HTra).

2.4 Activity Units in Translation

In TPR, translation behavior is often analyzed in terms of small segments of reading or typing activities. These segments of behavior tend to reflect a cognitive definition of a crucial and much-debated concept in translation studies—the "unit of translation" (Swadesh 1960; Rabadán 1991; Barkhudarov 1993; Bennett 1994; Malmkjær 2006; Kondo 2007; Alves and Vale 2009; Carl and Kay 2011).

Since the process of translating involves a behavioral pattern where "sudden bursts of production are followed by shorter or longer intervals with no typing activity while the source text (ST) is scrutinized" (Jakobsen 2019, 71), these intervals (i.e., the pauses between typing bursts) have been regarded as indicators of the boundaries between different production units (Dragsted 2010) and the cognitive processes concerning the change of attentional state (Schilperoord 1996).

In this regard, empirical studies have used different approaches to fragment the User Activity Data (UAD) to investigate the translator's cognitive effort and cognitive rhythm on the basis of typing pauses and gazing behavior (see Carl and Schaeffer 2017a), which include the production unit mentioned here. Another perhaps more detailed fragmentation is attention unit (Hvelplund 2016), a unit consisting of uninterrupted processing activity allocated to either the ST or the TT or to the ST with concurrent typing. Similarly, in the CRITT TPR-DB (Carl et al. 2016), this is represented by AUs which are categorized into the following types:

Type 1: ST reading.
Type 2: TT reading.
Type 4: Typing activity.
Type 5: ST reading and typing.
Type 6: TT reading and typing.
Type 8: No activity recorded.

Along the process of translating, transitions between one type of AU and another, accordingly, indicate shifts in activity.

Among all the AU types, four of them involve reading: 1, 2, 5, and 6, as can be seen above. This means that these AUs contain corresponding scanpaths where each individual word fixated is associated with an HTra value. These values can be used to calculate the mean of all fixated words to represent the HTra of the AU.[9]

As the AU types 4 and 8 do not contain fixation data, this chapter is focused on AU types 1, 2, 5, and 6.

An example of how these AU types are categorized is illustrated in the following progression graph (picture taken from Schaeffer et al. 2016a).

Here, the ST tokens are presented on the left side from the bottom to the top in sequential order, with the aligned TT tokens on the right side corresponding to the order of the ST. The horizontal axis indicates time, in milliseconds, during the

[9]It is worth mentioning that this average HTra value is not the same as the updated entropy for a particular ST item in the translator's mental state; see Sect. 2.2.

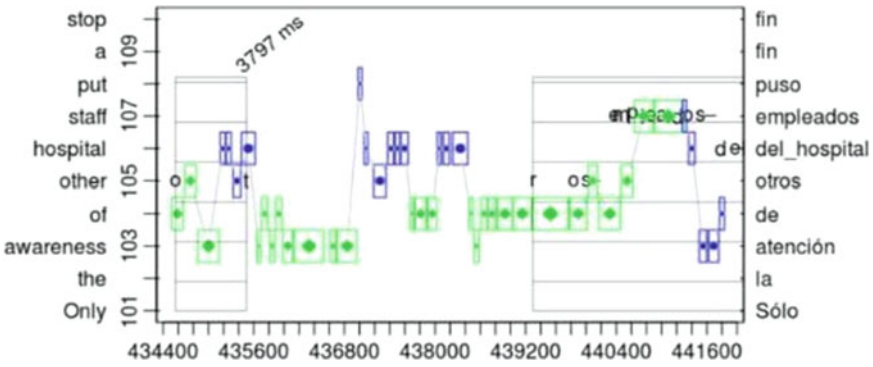


Fig. 1 Illustration of AUs within and between typing bursts (Picture taken from Schaeffer et al. 2016a)

experimental session. The blue dots indicate fixations on the ST, while the green diamonds refer to fixations on the TT.

The progression graph in Fig. 1 shows two typing bursts, "*ot*" and "*ros empleados*", as well as the movement of the eyes in this process. In the first typing burst (i.e., *"ot"*), the eyes fixate on a TT item (i.e., *de*) while typing *"o"*, which is categorized as AU Type 6 (TT reading and typing); before typing *"t"*, the eyes fixate on another two TT items (AU Type 2, TT reading) and move to the ST side (AU Type 1, ST reading). When *"t"* is typed, the eyes fixate on the ST word *hospital* (AU Type 5, ST reading and typing).

After this typing burst, the eyes move back and forth between the ST and the TT, starting from a few fixations on the target side (AU Type 2, TT reading), then switching to the source window (AU Type 1, ST reading), and then coming back to the target words for a shorter period of time (AU Type 2, TT reading), then to the source (AU Type 1), and then to the target (AU Type 2). While the eyes continue to fixate on the target item *de*, the typing activity resumes (AU Type 6, TT reading and typing), and this AU comes to an end when the eyes switch to the ST window while typing *"empleados"*, marking the start of the subsequent AU which contains concurrent typing and ST reading (AU Type 5).

These AUs within and between typing bursts, which can be analyzed via the progression graph in Fig. 1, are indicative of the cognitive processes when the translator is producing *otros empleados*.

In the following sections, the HTra values of each fixated word within such AUs, as well as the overall HTra for each AU, are discussed to gauge the cognitive processes in the translation of high-HTra (i.e., highly translation-ambiguous) items which tend to cause processing difficulty.

3 Entropy and Fixations in Activity Units

If word translation entropy and syntactic choice entropy represent the level of uncertainty involved in the translation choice at a particular point of the translating process (see Sect. 2.1), it is reasonable that a higher value of entropy would correspond to more effort involved in making that choice (see also Sect. 2.2) and therefore a larger value for fixation-based measurements of cognitive effort. This has in fact been studied and confirmed in the literature (e.g., Schaeffer et al. 2016b), as mentioned previously.

For the AUs, one might easily assume that the same correspondence can also be found between entropy and cognitive effort; if more effort is expended in an AU, the corresponding words which are processed in this AU should be more likely associated with high entropy values, so the scanpath would result in a higher level of average entropy. This would mean a visible positive relationship between HTra values and the number of fixations (nFix) within the AU.

Interestingly, in a preliminary study in Wei (2018), three scanpath measures (number of fixations, number of different words fixated, and duration) of all AUs which involve reading activity on either the ST or the TT (i.e., AU types 1, 2, 5, or 6) are analyzed in relation to the corresponding HTra values for these AUs (where the HTra value for an AU is calculated as the mean of all fixated words in the AU; see Sect. 2.4), and the relationship is found not to be a simply positive one. Based on a small dataset in an earlier version of the CRITT TPR-DB,[10] a pattern shown in Fig. 2 is discovered for all languages (Danish, German, Spanish, Hindi, Japanese, and Chinese), tasks (from-scratch translating, MT post-editing, sight translation, editing, and translation dictation), and AU types (1, 2, 5, and 6) in the data. As can be seen from the plots, this relationship does not follow a simply positive trend; instead, there seems to be a certain point of entropy as a threshold before which the number of fixations (nFix) tends to increase as entropy increases, and after this point, the number of fixations begins to show a trend of decreasing. The AUs with maximum number of fixations correspond not to the maximum of entropy but to its medium level.

When the outliers are removed by 2.5 standard deviations per subject, the pattern becomes somewhat more apparent, as can be seen from the plot on the right side in Fig. 2.

Specifically, for each AU type, this general pattern is also consistent (see Fig. 3).

As a preliminary study, Wei's (2018) discussion is fairly limited, other than illustrating the scatter plots and describing the entropy values in a single example of scanpath (i.e., the scanpath in Sect. 3.3).

[10]This includes the following studies in the multiLing dataset: BML12, ENJA15, KTHJ08, NJ12, RUC17, SG12, and STC17. Further description of the multiLing dataset in the CRITT TPR-DB can be found in Sect. 4 and on this webpage: https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies?authuser=0

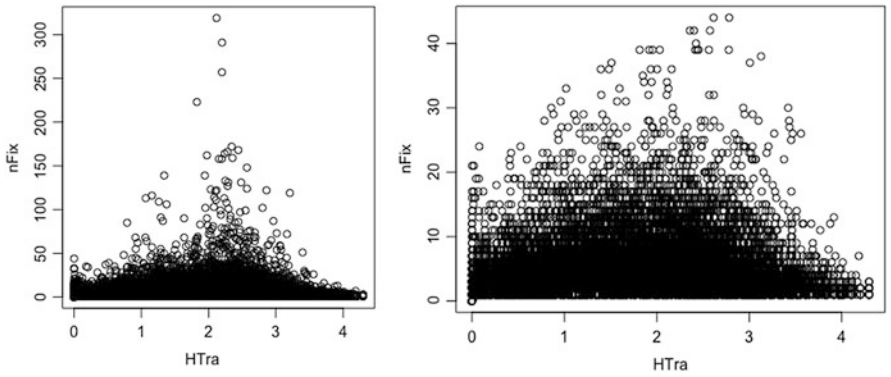


Fig. 2 Scatter plot of nFix and average HTra. Left, scatter plot with original data. Right, scatter plot when outliers are removed

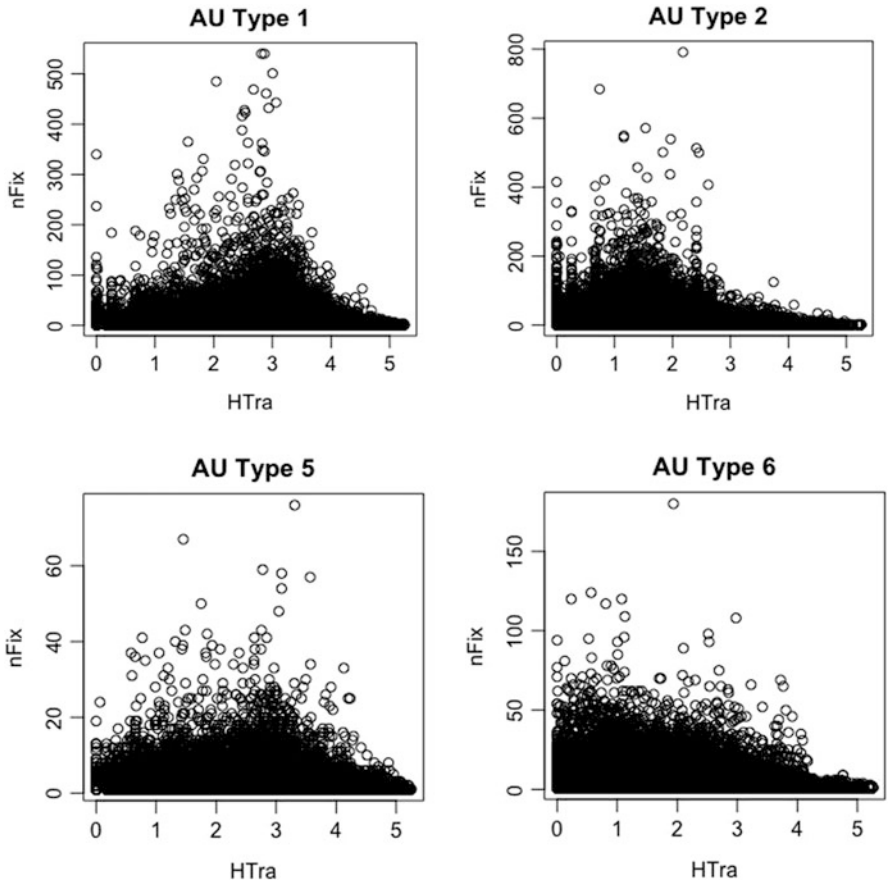


Fig. 3 Scatter plot for each AU type

In addition to the number of fixations, the same pattern is also found regarding another two features of the AU: the number of different words fixated (DFix) and the duration (Dur) of the AUs. This consistency is in fact not surprising, given that there is a "relatively high" correlation among nFix, DFix, and Dur (Schaeffer et al. 2016a, 339):

> ... the longer a coherent reading activity is (Dur), the more likely it is that more different words are fixated (DFix), resulting in a larger ScSpan. It is also more likely that progressions and regressions occur (higher Turn).

Since the two entropy-based predictors—HTra and HCross—correlate to each other for all the languages in the CRITT TPR-DB (see, e.g., Carl et al. 2019), it is not surprising either that this pattern is found to be consistent for both HTra and HCross.

3.1 Machine Translation Post-editing

As regards the pattern that the AUs with the largest number of fixations correspond to medium-level entropy rather than its maximum, one might be reminded of the findings from Krings' (2001) study on MT post-editing, where the cognitive effort of post-editing is found to reach its highest level not for poor MT output but for medium-quality translations from the MT (539–547).

Krings' explanation for this phenomenon in post-editing is that poor MT output causes a strategy shift and a "re-approximation of the post-editing process to a normal translation process" (541). As post-editing MT output of medium-level quality leads to "a disproportionate increase in coordination difficulties among ST-, MT- and TT-related processes," the cognitive effort involved is consequently higher for medium-quality MT output. In contrast, the refocused post-editing process when the output is poor, which approximates to a normal translation process, would require fewer coordination difficulties, and thus less effort, in terms of target text production.

For MT post-editing, if HTra is negatively correlated with MT quality, the above pattern would seem to be another representation of Krings' findings (at a micro level). One might perhaps hypothesize that a higher HTra value for an ST item, which indicates a higher level of translation ambiguity, poses a higher level of difficulty for the MT to produce an appropriate translation, making it more likely to result in errors that need to be post-edited. In this manner, the entropy values would be negatively correlated with MT quality. Then it follows that the instances of poor MT output to which Krings refers would be somewhat equivalent to high entropy values for the ST material, and in this respect, Krings' findings can perhaps be interpreted in terms of entropy as well, in a way which seems consistent with the findings in the abovementioned study.

This hypothesis regarding HTra and MT output has in fact been supported by empirical evidence in studies of the human translation process and of the errors

produced by MT. Translation ambiguity, which reflects the number of possible choices among all TT alternatives, is shown to be correlated with the perplexity of the MT search graph where possible translations are encoded by the engine (Carl and Schaeffer 2017b), and this perplexity of MT search graph is in turn correlated with post-editing duration (Carl and Schaeffer 2014). From another perspective, analyses of the types of MT errors in relation to translation ambiguity (measured by HTra) suggest that "a larger number of translation choices leads to increased (more evident) MT accuracy errors" (Carl and Toledo Báez 2019, 123). In Ogawa et al. (this volume, Chap. 6), the HTra values calculated from MT output are shown to correlate strongly with the HTra from human translations of the same ST material. Such findings reveal that translation ambiguity as indicated by HTra, which tends to increase the difficulty for human translation, has a largely similar effect on MT and in turn on the human post-editing process.

Here, as this pattern regarding the relationship between HTra and nFix in AUs is found to be consistent for both the post-editing of MT and translating from scratch (Wei 2018), it seems that the phenomenon may carry important information that reflects the behavior of translation in general.

3.2 Effect of Averaging

It is also important to note that since the HTra of the AU is the mean HTra of the words in the scanpath (see Sect. 2.4), the pattern as shown in the plots above may also be influenced by the statistical effect of this averaging in the calculation. If this influence is strong enough, then a larger number of fixated words in an AU would mean that this mean HTra value is more likely to approximate the mean HTra for all fixated words in the textual material during the entire task.

Figure 4 shows the mean HTra values for all fixations on the ST and on the TT, respectively, which are calculated using the same dataset in Wei (2018).

Mean HTra for all ST fixations: 2.471.
Mean HTra for all TT fixations: 1.375.

It can be seen that this statistical effect might have contributed considerably to the pattern, especially for AU type 2 (TT reading). However, since translation behaviors such as eye fixations on the words in the text and the translator's transition between AUs are not randomly occurring phenomena, questions as to what words are fixated in the scanpath, what fixated words fall into which AUs, and what HTra values are associated with these fixated words in corresponding AUs are perhaps far more important than the quantitative pattern of this correlation.

In this regard, the above pattern seems to suggest that when the translator is resolving a problem which arises from a high-HTra item in the text, the high-HTra item is unlikely to be in a long AU where all other items correspond to equally

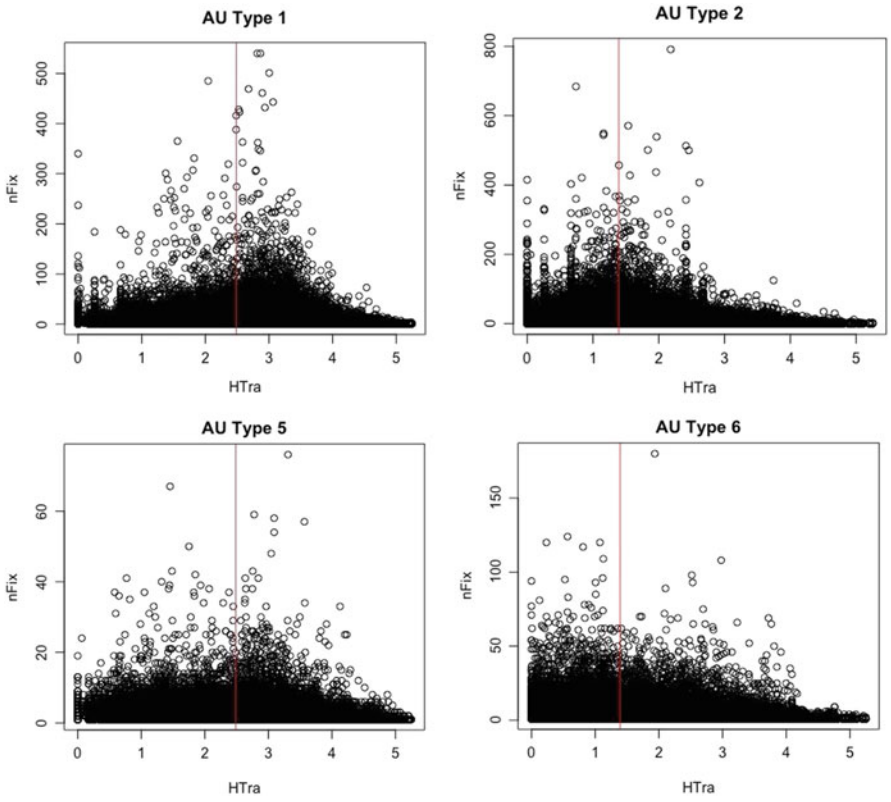


Fig. 4 Effect of averaging

high HTra values: as the AU lengthens, the number of fixations increases together with the AU's duration, and if the fixated items are all at a high level of HTra, the resulting average HTra for the AU would be equally high. This means that the values of both HTra and nFix would be large, which seems rather unlikely based on the plots in Fig. 2. Instead, the plots show that if the AU is longer, the other items which are fixated in the same AU tend to decrease the average HTra to a (medium) point which may in some way approximate the average HTra for all fixations.

In other words, during the update of the probabilities of the activated (TT) candidates regarding a high-HTra region of the ST (i.e., the decrease of entropy for this region; see Sects. 2.2 and 2.3), the AU tends to include fixations on relatively low-HTra items and therefore incorporates cognitive processing of the information associated with them. This will be illustrated in a more detailed manner in the following sections, where the scanpath of the translator and the specific entropy values are closely examined for instances where extra processing effort is expended on high-HTra items, using examples of a phrasal verb (*cough up*) which is metaphorical and idiomatic in the ST.

3.3 An Example of Scanpath

The following is an example of ST reading where the scanpath shows expenditure of extra processing effort and where the mean HTra value is at a medium level.

As can be seen from Scanpath 1.1 below, when the participant reads along the sentence "British families have to cough up an extra £31,300 a year" and encounters the metaphorical phrase *cough up*, his/her eyes fixate on *cough*, remain on this word for a period of time (indicating extra processing effort), then move back to a previous word *families* and then further back to *in* in the preceding title "Families hit with increase in cost of living," remain fixated at this point, and then move further back to *increase* (i.e., multiple regressions, which also indicate processing effort). It is at this point that the participant begins to move on to the next AU (see also the progression graph in Fig. 6).

Example 1

	HTra	HCross
AU 1.1	***2.0357***	1.3734

STid: 4 5 6 7 8 9 10 11 12 13 14 15 16
ST: ...increase in cost of living British families have to cough up an extra

17 18 19 20
£ 31,300 a year...

Scanpath 1.1

12 13 13 10 5 5 4
to cough cough families in in increase

	HTra	HCross
Cough	***3.3755***	2.6082
To	0.7654	1.0144
Families	0.2580	0.2580
In	2.0185	0.5132
Increase	2.4556	0.4262

The HTra and HCross values for this AU (denoted by AU 1.1) are shown at the top, both of which are at their medium level. The STid is an ID number given to every token in the text, and Scanpath 1.1 shows the eye movements of the participant within this unit. HTra and HCross values of each fixated word are shown in the table following Scanpath 1.1.

Perhaps the initial, lengthened fixation on *cough* in Scanpath 1.1 represents the participant's recognition of a problem regarding the metaphorical use of a word which would otherwise correspond to a different, more commonly used sense and TT equivalent (i.e., a sufficiently unexpected input), while all the following part of the scanpath contributes to the resolution of the problem where the participant seeks to arrive at a decision regarding the interpretation of the word in face of uncertainty. All the words fixated in this process are somehow related to the metaphor and would help the participant to contextualize and disambiguate the sense of *cough* in *cough up*. At the end of this process, the participant alternates into another type of AU.

Here, what is important is why the entropy of the unit is at a medium level if there is a high level of uncertainty associated with the unexpected input.

Looking at the entropy values of individual words in the scanpath, it is not hard to see that the word which causes the problem (the unexpected input), *cough* (or perhaps more accurately, *cough up*), has relatively high HTra and HCross values (3.3755/2.6082), while the entropy of all the other words on which the participant's eyes fixate are much lower. When the entropy associated with each of these fixations are calculated into a mean, the overall value (2.0357/1.3734) is considerably dragged down by the low-entropy words toward a medium level.

In other words, the scanpath of AU 1.1 is a combination of one word with very high entropy (as the problem) and many low-entropy words (which facilitate problem-solving), resulting in an average at the medium level for the entire unit.

4 "Cough Up": Analysis on Activity Units

The above analysis should provide a preliminary explanation for the pattern that AUs with larger numbers of fixation tend to correspond to medium-level entropy. However, it is important to examine the translation behavior pertaining to this example in a detailed and comprehensive manner, to genuinely understand the abovementioned observation in the scanpath. In a wider scope of the example shown above, what AU does the participant change into after disambiguating the sense in the current scanpath? Does this mark the completion of the activation, disambiguation, and translation selection processes and therefore directly lead to typing (i.e., translation production), or is this only part of the process, so that the participant subsequently switches to the TT for further processing of the disambiguation? How is this behavior represented in the translation into different languages?

From a broader perspective, it is also important to examine a larger dataset and test if the same pattern in Fig. 2 exists in the larger data regarding more languages and tasks and, more importantly, whether the phenomenon discussed in Sect. 3.3—that the overall HTra of AUs containing fixations on *cough* is dragged down by lower-HTra items—represents a general pattern in view of the dataset.

These are important questions, and it is also important to note that the examination of HTra and nFix sheds light on the relationship between uncertainty and cognitive effort, as well as on the mental processes of disambiguation, lexical selection, and uncertainty resolution.

In order to further analyze the relationship between the number of fixations within AUs and the word translation entropy corresponding to these units, a larger multilingual dataset named "multiLing"[11] in the CRITT TPR-DB is used to investigate the pattern. The multiLing dataset includes multiple studies of different tasks of translation production into various languages from the same English source texts (STs) and is therefore convenient for comprehensive analysis. Among the studies in this dataset, ten are used for the present study, incorporating all the languages in multiLing: Arabic, Chinese, Danish, German, Hindi, Japanese, and Spanish (AR19, BML12, ENJA15, KTHJ08, MS12, NJ12, RUC17, SG12, STC17, STML18).

4.1 HTra Values for AUs Containing "Cough"

A look at the HTra of all AUs in which the scanpath contains fixation(s) on this instance of *cough*, and where the total number of fixations in the AU is larger than one, shows that the phenomenon regarding the HTra values of the scanpath in Sect. 3.3 seems to represent a general trend when it comes to the same instance (i.e., *cough*) in the same ST, in view of all participants in each study and in view of different studies on various languages.

Figure 5 shows the density plots for such AUs associated with this phrasal verb, in each study and in terms of their HTra values,[12] with the red vertical line in each plot indicating the HTra of *cough* in the corresponding study. As can be seen from the plots, the AUs where *cough* is fixated tend to result in an overall HTra (i.e., average HTra for the fixated words in the scanpath) which is much lower than the HTra value of *cough* itself. This means that the other items which are fixated within the same AUs tend to be lower-HTra words.

[11] A detailed description of the multiLing dataset is available here: https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies?authuser=0

[12] Note that these density plots refer to the HTra of the abovementioned AUs, i.e., AUs which contain fixation(s) on *cough* and which contain more than one fixation, rather than the distribution of HTra values of the individual words within specific AUs.

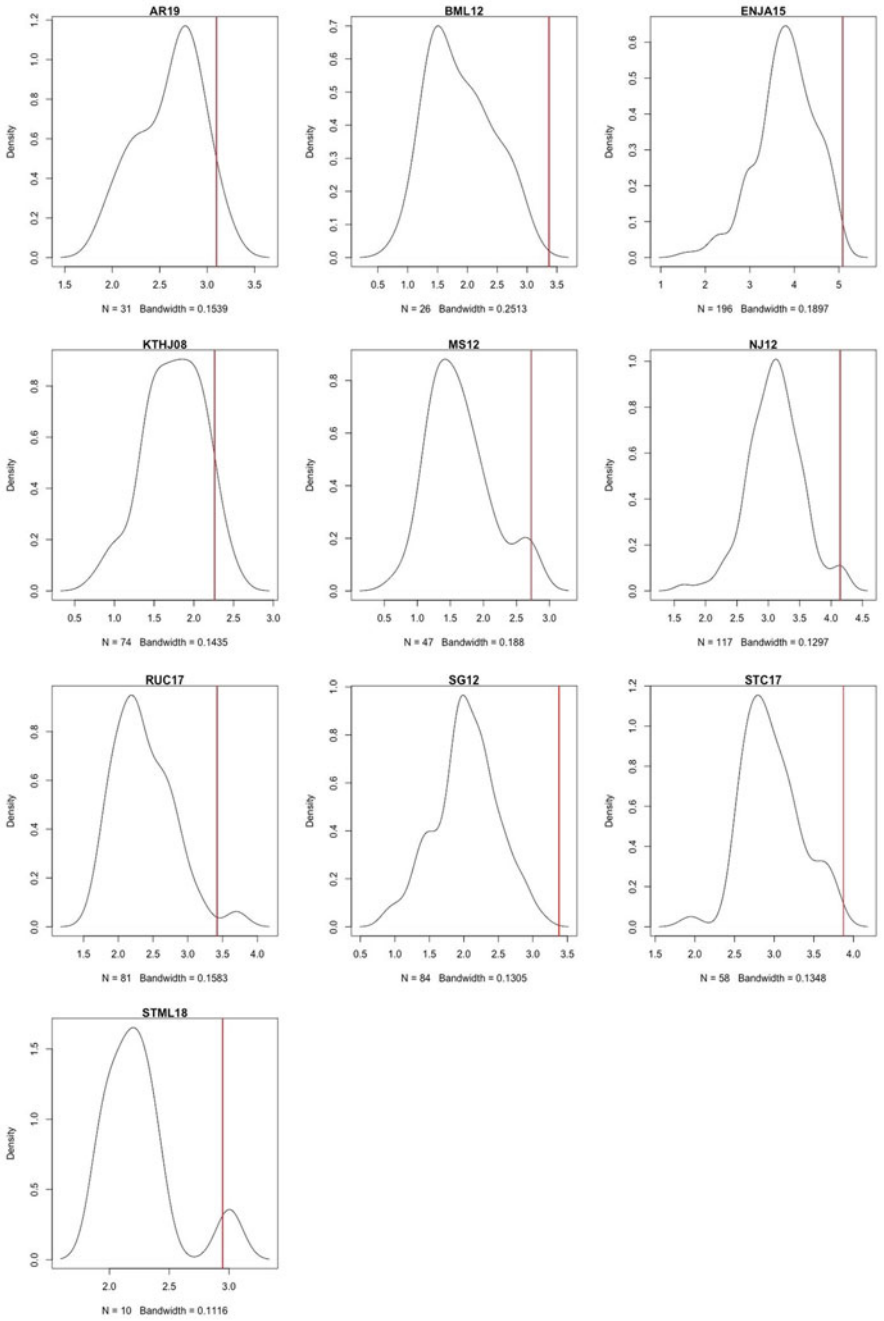


Fig. 5 HTra values of AUs containing fixation on *cough*

4.2 Progression Graph Analysis

Regarding the above example of scanpath in Sect. 3.3 (i.e., Scanpath 1.1), a wider scope of the process can be seen in the progression graph below.

The task here is MT post-editing, with German as the target language. The scanpath corresponding to the example of ST-reading AU (AU 1.1) in Sect. 3.3 occurs at 41,433 on the horizontal axis (see Fig. 6).

As can be seen from the graph (starting from 41,433), the participant's eyes fixate on *to*, proceed to *cough*, remain on *cough*, and then move to the preceding words (i.e., the scanpath in AU 1.1 as shown in Sect. 3.3). Those fixations are illustrated with blue dots, and the AU is categorized as Type 1 (i.e., ST reading).

After this encounter of *cough*, as well as the translator's gazing behavior in AU 1.1, the following AUs seem to be a continuation of the translator's problem-solving process, with fixations on a few other words in the context surrounding the word *cough*, alternating between TT reading and ST reading, before finally coming back to the original word causing the problem (i.e., *cough*) and starting an apparently linear reading process on the ST sentence (i.e., the last AU in Fig. 6 beginning at 47,275 on the horizontal axis).

In other words, these AUs altogether seem to represent the entire problem-solving process with respect to the issue arising from the metaphorical and idiomatic use of the phrasal verb *cough up* in the ST material.

Specifically, the following AUs are of particular interest:

AU 1.1: ST reading, 41,433–42,038 on horizontal axis (i.e. time).
AU 1.2: TT reading, 42,038–42,675 on horizontal axis.
AU 1.3: TT reading, 46,283–47,275 on horizontal axis.
AU 1.4: ST reading, 47,275–54,941 on horizontal axis.

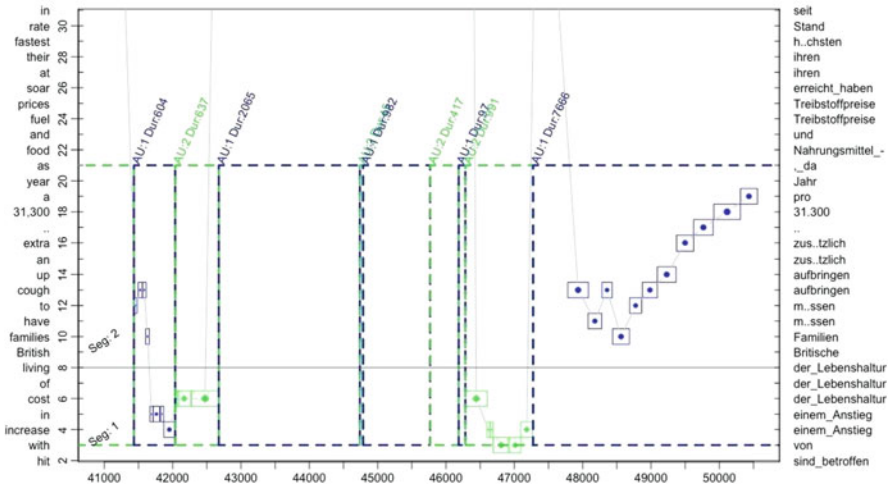


Fig. 6 Progression graph

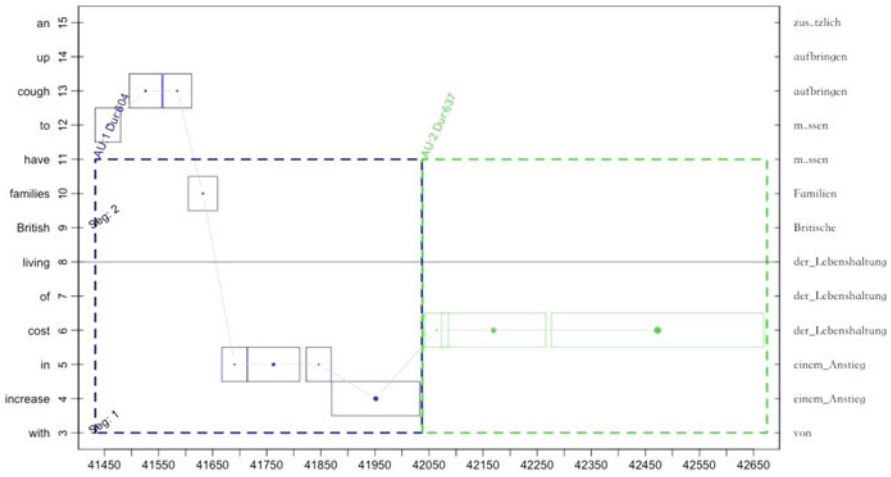


Fig. 7 AU 1.1 and AU 1.2

Returning to the question raised at the beginning of Sect. 4, Figs. 6 and 7 show that after the scanpath discussed in Sect. 3.3 (i.e., AU 1.1), the participant switches to TT reading on the machine-translated output of a word in the preceding segment, fixating for a relatively long time on *Lebenshaltungskosten*, another word which perhaps helps the participant to contextualize and disambiguate the sense of the metaphor *cough up*.

Then the fixated point is at the very end of the ST for about 4 s (see Fig. 6), which is likely an untargeted gaze while the participant processes the information.

After 46,283 on the horizontal axis (marking the start of AU 1.3, see Fig. 8), the participant's eyes come back again to the TT token *Lebenshaltungskosten* and then move to a few words in the preceding context, before finally returning to the original position where the problem arises: *cough* in the ST (i.e., the beginning of AU 1.4).

The scanpaths for the two sequences of TT reading (i.e., AU 1.2 and AU 1.3) are shown below, together with the corresponding ST and TT sentences:

(Fixated words are underlined)

ST: Families hit with increase in cost of living
British families have to cough up . . .

TT: Familien sind von einem Anstieg der Lebenshaltungskosten betroffen
Britische Familien . . .

Scanpath 1.2: Lebenshaltungskosten → Lebenshaltungskosten →
Lebenshaltungskosten.

Scanpath 1.3: Lebenshaltungskosten → Anstieg → Anstieg → von
→ von → Anstieg.

The words fixated in this process, following the ST reading illustrated in Sect. 3.3, seem to have continued to aid the participant to cognitively process the meaning of the word *cough* in the metaphorical phrase *cough up* and perhaps to disconfirm the alternative interpretations of this word. Therefore, what follows AU 1.3 is continuous reading on the ST, beginning with a gaze pattern similar to the one

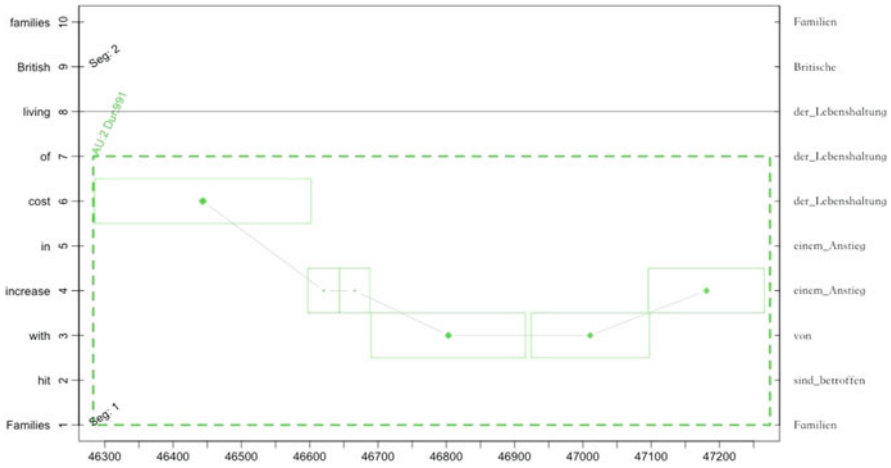


Fig. 8 AU 1.3

illustrated in Sect. 3.3 and proceeding with linear reading along the ST sentence (i.e., AU 1.4).

As mentioned, these four AUs constitute the entire process of the participant's recognition of the problem arising from the metaphorical sense of *cough up*, as well as his/her resolution of the problem and verification of the appropriateness of the machine-translated output. It can be seen that the word *cough* has cost considerable effort from the participant to integrate the metaphorical meaning in the context, and under the assumption of nonselective activation of the source and target language (see Sects. 1 and 2), perhaps the encounter of *cough* activates a semantic space which is incompatible with the context and which therefore requires extra processing effort. The long fixation on *cough* in the initial AU of ST reading (i.e., AU 1.1) signals this, while all the fixations on the other ST words, the alternation between the ST and the TT, the fixation on *cough* again, and the following regression can perhaps be considered the process in which the probability distribution of the activated items associated with *cough* keeps being updated as the participant attempts to resolve the problem, decreasing the uncertainty to the lowest level and eventually arriving at a choice among the activated items.

Here, what this chapter argues is that this problem can be represented and quantified by entropy values (i.e., HTra, for word senses and translations), with higher HTra values indicating higher levels of uncertainty pertaining to the words or expressions in question. In the meantime, the resolution of the problem which arises from this uncertainty associated with a particular high-entropy word or expression is through a process which is facilitated by a number of surrounding low-entropy words in the context.

In addition to the HTra values of Scanpath 1.1 as shown in Sect. 3.3, the corresponding entropy values for the scanpaths on TT (i.e., AU 1.2 and AU 1.3) are as follows:

	HTra	HCross
Lebenshaltungskosten	0.0435	0.0283
Anstieg	0.6139	0.3239
von	3.0836	2.4349
Scanpath 1.2	0.0435	0.0283
Scanpath 1.3	1.3421	0.9783

For Scanpath 1.2, the entropy for the entire AU is very low, clearly because the fixations are on a very low-entropy word (*Lebenshaltungskosten*). This is not surprising, given that this scanpath, as shown in the analysis above, is part of a larger process of resolving the problem associated with a high-entropy word in the preceding AU.

For Scanpath 1.3, the overall entropy is considerably dragged down by the many fixations on words with very low entropy values, which is consistent with what has been explained in Sect. 3.3.

4.3 Translation Tasks into Different Languages

As the above example is from a post-editing task, this discussion merits further analyses on examples of the same problem in translating from scratch. The following are a few examples regarding translation tasks into different languages.

Example 2

Figure 9 shows an AU associated with the same phrasal verb, for a different participant translating the same text into German. The AU involves reading on the ST (AU Type 1) during a pause of translation production, where the translator

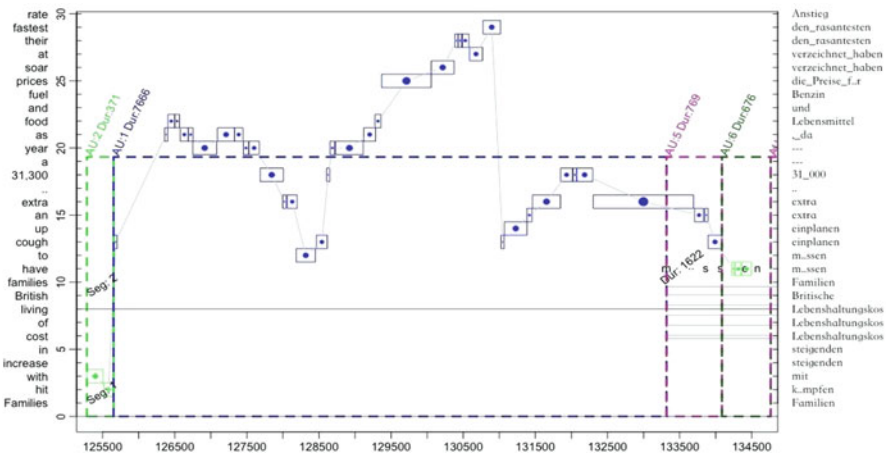


Fig. 9 Progression graph for Example 2

scrutinizes the ST sentence, fixating on *cough* and other items in its context for multiple times, with several regressions toward *cough*, before producing the TT item *müssen* (in *müssen 31,000 £ extra einplanen*).

The ST material, the TT produced by this participant, the translator's scanpath within the AU, and the relevant entropy values in this regard are as follows.

(Fixated words are underlined)

ST: Families hit with increase in cost of living
British families have to cough up an extra £31,300 a year as food and fuel prices soar at their fastest rate in 17 years.

TT: Familien kämpfen mit steigenden Lebenshaltungskosten
Britische Familien müssen 31,000 £ extra einplanen, da die Preise für Lebensmittel und Benzin den rasantesten Anstieg seit 17 Jahren verzeichnet haben.

Scanpath 2: cough → as → food → food → as → as → year → as → as → year → year → 31,300 → extra → extra → to → cough → 31,300 → year → year→ as → food → prices → soar → their → their → their → at → fastest → cough →up → an → extra → 31,300 → 31,300 → 31,300 → extra.

	HTra	HCross
AU 2	***2.5112***	2.2122
Cough	***3.3755***	2.6082
As	2.4287	2.0185
Food	2.7401	2.599
Year	1.0862	1.0862
31,300	1.567	1.3885
Extra	2.6914	2.9583
To	0.7654	1.0144
Prices	2.4464	1.4708
Soar	2.9895	3.2077
Their	4.2299	3.2618
At	4.2299	3.5555
Fastest	4.2627	3.0994

Similar to the previous example, in this AU, the translator seems to be expending extra processing effort on *cough up*, and in resolving the problem, there is a visual search for a number of items in the context which seem to facilitate disambiguation.

In terms of the entropy values, it is not hard to see that other than five fixations (on three words, namely, *their*, *at*, *fastest*), the vast majority of the fixations in the AU (31 out of 36 fixations) are on words whose HTra values are considerably smaller than *cough*. In other words, the HTra of *cough* is very high, causing the problem and uncertainty, while the rest of the AU tends to be fixations on low-HTra words. This has resulted in the fact that the overall HTra for the unit (2.5112) is lower than that of the word which causes the problem (3.3755). It also shows that the resolution of the uncertainty involved in the disambiguation process is largely facilitated by low-entropy words.

Example 3

Another example regarding this instance, for translation into Spanish, is also revealing of the same phenomenon, as can be seen in the following progression graph and the following table of the relevant entropy values.

In this example, it is apparent that the translator is expending extra effort in processing the word *cough* when producing its target translation *asumir*. At 128,265 on the horizontal axis, the translator encounters *cough*, and the typing activity comes to a pause upon completing the production unit of *Las familias británicas*, while the translator scrutinizes the part of the ST containing *cough* before proceeding with a typing burst of *tienen que asumir* (see Fig. 10).

This pause of typing activity constitutes an AU of ST reading, where the translator fixates on *cough* for a relatively long time (as represented by two consecutive fixations on the same word), searches for other words in the context, re-fixates on *cough*, continues the reading on the following words in the ST, and then engages in a backward eye-movement (i.e., a regression).

Here, the scanpath shows the same behavior: encountering the problem caused by a high-entropy word and then searching for the low-entropy words surrounding it to facilitate resolution of the problem. In terms of the entropy values, the effect of those low-entropy words results in an average entropy which is at a medium level.

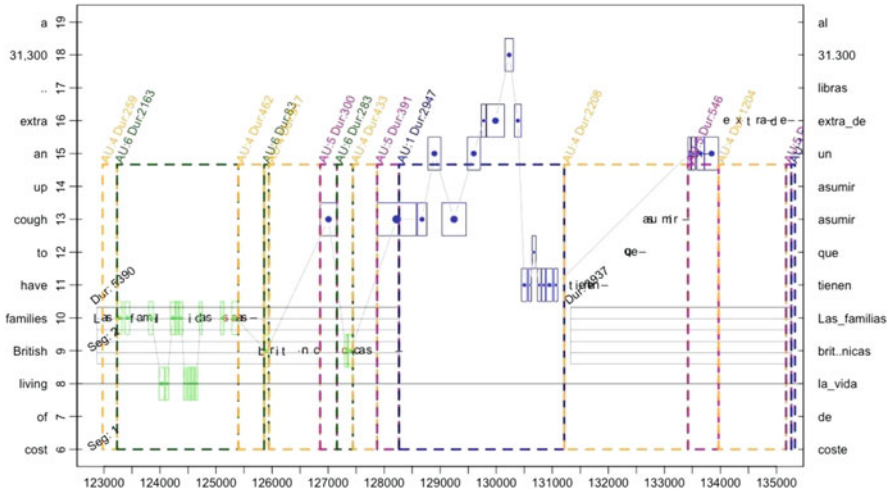


Fig. 10 Progression graph for Example 3

(Typing bursts are indicated by '///')

ST: British families have to cough up an extra £31,300 a year as . . .

TT: Las familias británicas /// tienen que asumir /// un extra de 31.300 libras al año . . .

Scanpath 3: cough → cough → an → cough → an → extra → extra → 31,300 → extra → have → have → to → have → have → have → have.

	HTra	HCross
AU 3	***1.8842***	1.1272
Cough	***3.373***	1.1055
An	2.0165	1.4071
Extra	1.8151	1.8696
31,300	0.8332	1.3039
Have	1.3786	0.65
To	1.4453	1.0912

Example 4

The following example is a translation task into Chinese, and the scanpaths of two AUs within pauses of typing show a similar pattern.

Figure 11 illustrates the AU (Type 1, ST reading) between two typing bursts by the translator: "英国的家庭"(*British families*) and "每年"(*a year*). As can be seen from the progression graph (Fig. 11), by the time the translator finishes the production of the translation for *British families* (while fixating on the TT), the eyes move back to the ST and encounter the word *cough*. In the meantime, the production of the TT comes to a pause, during which the translator scrutinizes this part of the sentence, his/her eye movements showing a clearly nonlinear reading activity with considerable regression. It is evident that the cognitive effort is elevated at this point, as indicated by the pause of typing and the apparent, multiple regression in the eye movement during this pause (which is also apparent in the previous examples).

At the end of this AU, the translator resumes his/her production of translation while fixating on the TT (see Fig. 11).

(Typing bursts are indicated by '///')

ST: British families have to cough up an extra £ 31,300 a year as food and fuel prices soar at their fastest rate in 17 years.

TT: 食品和燃料的价格以17年中最快的速度增长，导致 /// 英国的家庭 /// 每年 /// 都要额外挤出31300英镑的支出。

Scanpath 4.1: cough → 17 → £ → year → 31,300 → up → extra → extra → 31,300 → an.

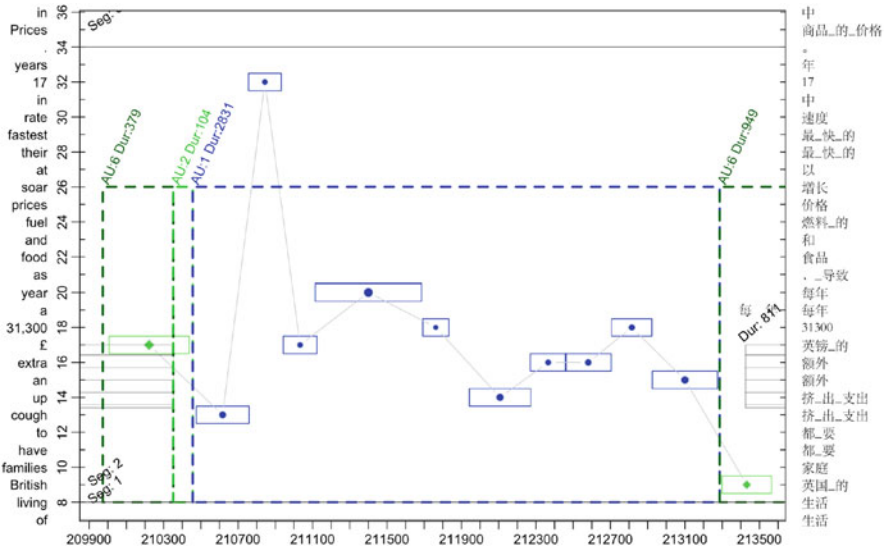


Fig. 11 Progression graph for AU 4.1

	HTra	HCross
AU 4.1	***2.2962***	2.1219
Cough	***3.4226***	2.4255
17	1.8676	2.3935
£	1.5064	1.6017
Year	0.549	2.0028
31,300	1.8676	1.5064
Up	3.4226	2.4157
Extra	2.7539	2.4876
An	2.9511	2.3923

Evidently, the HTra value of *cough* (as well as *up*, as part of the phrasal verb) is considerably high compared with all other items which are fixated in this scanpath. The high-HTra item has caused an increase of processing effort, while the scanpath involves other low-HTra items in the context which decrease the overall HTra value of the AU.

This AU is followed by a few typing bursts which are separated by pauses, as can be seen from the progression graph in Fig. 12. These bursts of typing are "每年"(*a year*), "都要"(*have to*), "额外"(*extra*), and "挤出"(*cough up*).

(Typing bursts are indicated by '///')

ST: British families have to cough up an extra £ 31,300 a year as food and fuel prices soar at their fastest rate in 17 years.

TT: 食品和燃料的价格以17年中最快的速度增长，导致英国的家庭 /// 每年 /// 都要 /// 额外 /// 挤出 /// 31300英镑的支出。

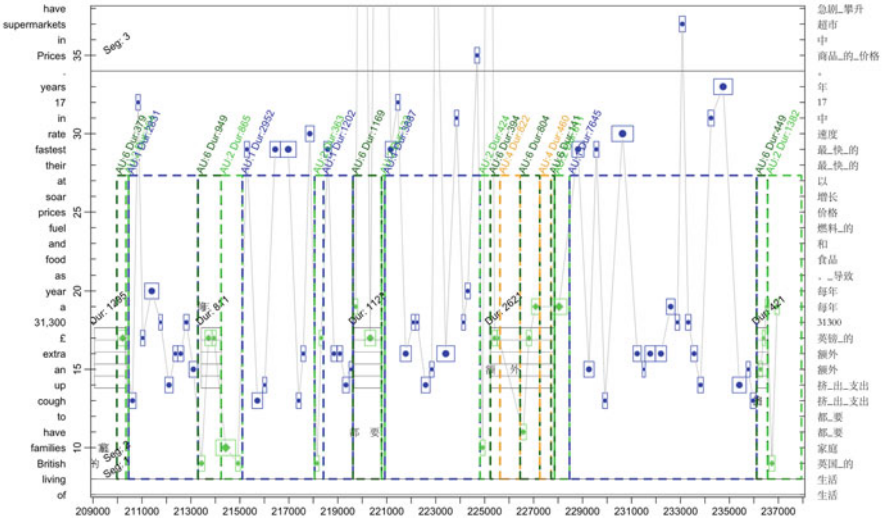


Fig. 12 Progression graph of the typing bursts after AU 4.1

In this regard, the AU (Type 1—ST reading) before the typing burst of '挤出'(i.e. the production of the translation for '*cough up*') is also particularly relevant to the translator's processing of the word '*cough*' in relation to the further steps of word sense disambiguation and—perhaps more importantly—selection of an appropriate TT item among the alternatives which are activated upon encountering this word in AU 4.1 above.

Figure 13 shows this AU (AU 4.2) in detail, and similar to what has been illustrated above, there seems to be an extra processing effort involved in the translator's production of the translation for *cough up*, as indicated by the pause of production and the nonlinear reading activity on the ST, with the eyes frequently moving back and forth in this region of the sentence.

HTra values of the words which are fixated in AU 4.2 display the same pattern as what has been illustrated above (see the scanpath and entropy values below).

ST: British families have to cough up an extra £ 31,300 a year as food and fuel prices soar at their fastest rate in 17 years. Prices in supermarkets have . . .

TT: 食品和燃料的价格以17年中最快的速度增长，导致英国的家庭每年都要额外挤出31300英镑的支出。

Scanpath 4.2: fastest → an → fastest → cough → rate → extra → an → extra → extra → a → 31,300 → supermarkets → 31,300 → extra → up → in → years → up → an → cough.

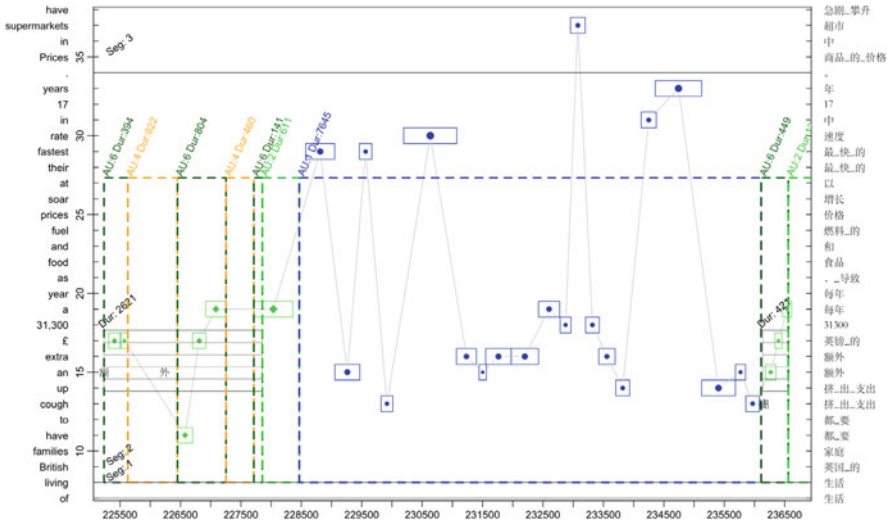


Fig. 13 Progression graph for AU 4.2

	HTra	HCross
AU 4.2	***2.5603***	2.3389
Fastest	2.2485	2.7132
An	2.9511	2.3923
Cough	***3.4226***	2.4255
Rate	2.485	2.7481
Extra	2.7539	2.4876
a	0.549	2.0028
31,300	1.8676	1.5064
Supermarkets	0.8181	1.6957
Up	3.4226	2.4157
In	3.4585	2.82
Years	2.104	2.2624

Example 5

In the following example, the number of fixations (i.e., nFix) in each AU is much smaller than in the examples discussed above, resulting in a less obvious, yet still consistent, pattern of HTra values.

ST: British families have to cough up an extra £ 31,300 a year as...

TT: 由于食物和燃油价格激增，增速为17年以来最高，/// 英国家庭 /// 每年需要 /// 额外支付 /// 31, 300英镑

Starting from the first AU (Type 5, i.e., ST reading and typing) shown in the progression graph in Fig. 14, the translator produces the TT for *British families* (“英国家庭”) while reading the subsequent part of the sentence, and as the eye movement proceeds to the instance of the metaphor (*cough up*), the production

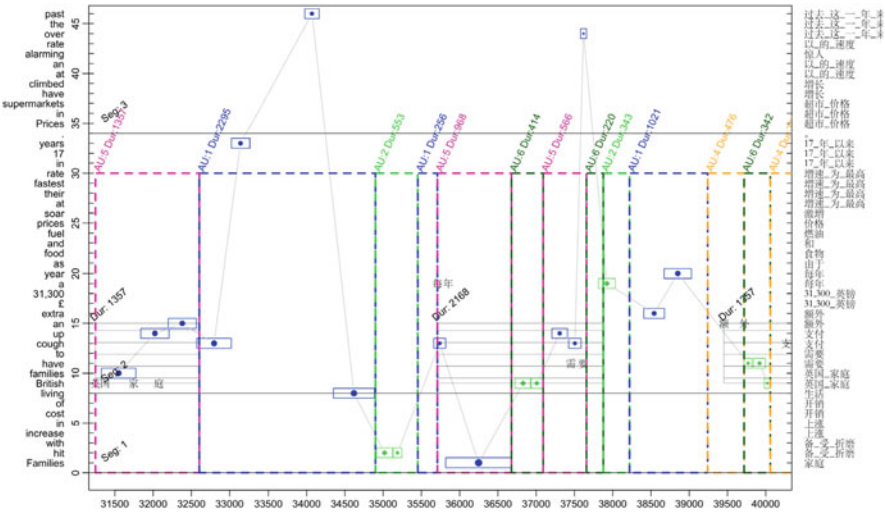


Fig. 14 Progression graph for Example 5

of translation is paused. The AU changes into Type 1 (ST reading) where a considerable amount of time is spent in reading the ST at different positions. Then the translator's eyes fixate on the TT for a short time before producing the translation again (see Fig. 14).

For the first AU at the beginning (AU 5.1, ST reading and typing), the translator's scanpath and entropy values are as follows:

Scanpath 5.1: families → up → an → cough.

	HTra	HCross
AU 5.1	***3.2817***	2.5464
Families	1.7988	2.2169
Up	3.875	2.555
An	3.5778	2.8585
Cough	***3.875***	2.555

Interestingly, a look at the entropy values here shows a phenomenon which is slightly different from the previous pattern. The fixations in this AU are mostly on high-HTra items (*up*, *an*, *cough*), contrary to the examples illustrated above, although the overall HTra for the AU is still smaller than the HTra of *cough*.

This would be understandable if one takes into account the larger context of the unit. As the progression graph and the scanpath both show, this AU is probably a stage where the problem or difficulty arises: the first encounter of the ST item in question, the nonselective activation of a semantic space associated with this item, the translator's subsequent realization that this activation may not be

compatible with the context in the ST, and consequently an update in the probability distribution of the activated items where the probabilities seem to have become less concentrated on the (incompatible) ones whose initial probabilities are relatively high. The progression graph indicates that the AU is primarily one which produces the translation of the words prior to *cough* (i.e., *British families*), while the scanpath shows that the fixation on *cough* occurs at the end of the AU. The translator is already processing the information in the subsequent words (*have to cough up an*) when translating the first phrase (*British families*), and upon encountering the problem *cough*, the typing comes to a pause and the current AU ends, so that the translator focuses on ST reading to resolve the problem associated with *cough*.

For the subsequent AU (AU 5.2, ST reading) after the typing has paused, the results are as follows:

Scanpath 5.2: cough → years → past → living.

	HTra	HCross
AU 5.2	***2.8870***	2.6442
Cough	***3.875***	2.555
Years	2.7718	2.6468
Past	3.2744	3.5
Living	1.6266	1.875

Here, the HTra for cough is still the highest among all the items fixated in the AU, but the result seems to be somewhat less obvious in terms of the pattern discovered above, i.e., that all the other words in the AU would be associated with very low entropy values.

To explain this, perhaps we can return to the plot in Fig. 2. The plot shows that larger numbers of fixation tend to correspond to medium-level entropy, and as the nFix decreases from the peak, the range of possible HTra values corresponding to the nFix would widen visibly (see Fig. 2).

In Example 5, the number of fixations in the scanpath is 4, which is rather small especially compared with Example 2 (nFix = 36), Example 3 (nFix = 16), and Example 4 (nFix = 10 for AU 4.1, nFix = 20 for AU 4.2). Therefore, the pattern here is much less obvious, although the pattern itself is still consistent to what has been displayed in the other examples.

In the meantime, a smaller nFix in this case seems to indicate that the translator is expending less effort in this AU than in many AUs of the previous examples, perhaps because the metaphor poses less of a problem for him/her. This is also represented by the fact that the translator quickly switches back to typing activity after fixating very briefly on a limited number of lower-entropy words in the ST and on the most recently typed words ("英国家庭") in the TT (see Fig. 14).

As mentioned, the translator is already processing the subsequent words when producing the translation of *British families*. In this regard, it would perhaps be meaningful to examine the eye-key span in the data and test if longer eye-key spans

tend to influence the number of fixations in AUs. Due to the scale of this study, however, the discussion will focus on entropy and fixations.

Although the nFix value is small in this case, a general tendency similar to the previous examples can still be found regarding the HTra values of the fixated words in the AU.

4.4 Dynamic Change of HTra Within the AU

On the basis of Example 5, the following section provides a close examination of the dynamic change of the average HTra values as the scanpath lengthens in each AU. It can be seen from the progression graph that upon encountering the high-HTra word, the translator pauses translation production (with concurrent reading on the ST) and devotes more attention to reading the ST. AU 5.1 shows an increase of average HTra as the eye movement proceeds, while AU 5.2 shows the opposite.

At the beginning of AU 5.1, the translator processes the information associated with *families* (HTra=1.7988) while producing its translation. The average entropy at this moment would be the entropy of the word itself, i.e., ***1.7988***.

Without pausing the production of translation, his/her eyes move on to the following part of the sentence and fixate on *up* (HTra=3.875), a high-entropy input which suddenly increases the average entropy of the scanpath: (1.7988 + 3.875)/2 = ***2.8369***. The translator maintains this AU, continuing production of the translation while fixating on more words. The next fixation is on *an* (HTra, 3.5788), which is again a high-entropy input to further increase the average entropy within the AU: (1.7988 + 3.875 + 3.5788)/3 = ***3.0842***. Then the translator proceeds with the reading and typing activities, this time encountering another high-entropy input—*cough* (HTra=3.875). The average entropy now becomes ***3.2817***, and at this point, the new entropy value seems to be high enough to trigger a breakdown of the unit, with the translator reallocating cognitive resources to resolve the problem arising from the high-entropy input.

Accordingly, the subsequent AU 5.2 displays a decrease of average entropy as the scanpath lengthens, where the translator attempts to resolve the problem by fixating more words in the context.

While pausing the production of translation, the translator continues fixating *cough* (HTra=3.875), and the average entropy for the current AU is ***3.875***. With the following fixations on *years* (HTra=2.7718), *past* (HTra=3.2744), *living* (HTra=1.6266), the average entropy value for the scanpath is incrementally decreased at each step:

(3.875 + 2.7718)/2 = ***3.3234***
(3.875 + 2.7718 + 3.2744)/3 = ***3.3071***
(3.875 + 2.7718 + 3.2744 + 1.6266)/4 = ***2.8870.***

In short, AU 5.1 is a process of incremental increase of the average entropy, while AU 5.2 is a process of decrease in the average entropy. It seems that once the average

entropy increases to a certain point, AU 5.1 breaks down, and the translator starts a new AU where effort is expended to decrease the average entropy toward a lower level.

5 General Discussion

On the basis of entropy as an indication of uncertainty and cognitive load, the above sections have discussed from a theoretical perspective how entropy and entropy reduction can describe the cognitive activities, as well as to quantify the effort, in the translation process. These sections have also demonstrated with empirical evidence that the expenditure of cognitive effort on ST items of high HTra values constitutes AUs where low-entropy words in the context come into play, thereby decreasing the average entropy of the scanpath in the AU. In the mental processes regarding a high-HTra word, linguistic context is used to gather additional information for clarification or disambiguation, and as a consequence, gaze on other words in the surrounding context is observed. It appears that the ambiguity, uncertainty, or unexpectedness involved in particular lexical items effectively directs the translator to those aspects of context which are relevant to rendering the items unambiguous, to integrating the newly activated information into what has already been activated, to enhancing or suppressing certain activation depending on its relevance to or compatibility with the context, to reducing the uncertainty (and entropy) level in the translator's mind, and to finally arriving at a selection among the co-activated alternative translations for an option which suits the ST and TT context. What seems more meaningful in this observation is that, on the one hand, the high-HTra word has led to an apparent pause of production and nonlinear reading activity (see examples above), and on the other, the words which are fixated in the surrounding context in order to facilitate the disambiguation (or resolution of the uncertainty resulting from translation ambiguity) pertaining to a high-HTra ST word tend to be at a lower level of translation ambiguity. This means that the high-HTra word has resulted in extra processing cost, while this additional effort is expended on searching for the contextual information that is provided by the lower-HTra words surrounding the current high-HTra word.

In the meantime, drawing inferences from the formulation of resource-allocation processing difficulty in psycholinguistics, the present chapter has also provided a theoretical analysis of the cognitive processes in which contextual information is integrated and in which the entropy is decreased in the mental states, while pointing out that when an update of entropy occurs as the translator resolves the problem arising from a highly translation-ambiguous item, this update tends to involve input of information from items at lower levels of translation ambiguity.

Upon encounter of a highly translation-ambiguous (i.e., high-HTra) word, it is assumed that multiple TT alternatives for this word are activated nonselectively in early priming processes. As the translator attempts to resolve the uncertainty involved in the high-HTra word and to make a selection among the activated

TT items, the pattern of activation, which eventually appears as a probability distribution in the produced text, keeps being updated to concentrate on options which are more probable than others, until a particular translation choice is made regarding this highly translation-ambiguous word. This process of change in the pattern results in a decrease of entropy in the mental state, in terms of both the mathematical expressions for the distribution of probabilities (see Sect. 2.2) and the notion of entropy as disorder in systems theory (see Sect. 2.3). This decrease of entropy in the mind through cognitive processing appears to be manifested in the behavioral data as an observable visual search for surrounding words at much lower levels of HTra, thereby also decreasing the average HTra of the scanpath in the corresponding AU. From a systems theory perspective, the decrease of entropy (i.e., disorder) of a word translation system tends to involve other systems which are much less entropic than the current system in question.

This is supported by a general analysis of the average HTra values for all AUs in the multiLing dataset and more importantly by the detailed discussion in Sect. 4 on the AUs concerning a particular ST item which causes additional processing effort. In addition, Sect. 4 also shows that the pattern regarding HTra values of the scanpaths which are described in Sects. 3.3, 4.2, and 4.3 is largely representative of the cognitive processing in general pertaining to the same ST problem.

Specifically, the phrasal verb *cough up* as a metaphor is associated with a relatively large value of word translation entropy, reflecting a high degree of uncertainty in the disambiguation (or selection) process and therefore a higher level of cognitive load in disconfirming the disconfirmable interpretations regarding this ST word. Accordingly, the AU which is primarily associated with the initial encounter of this item in Example 5 has a relatively high entropy value.

When the problem or difficulty regarding the high-entropy word in the ST causes additional processing effort (as indicated by pause of production and nonlinear reading activity), the uncertainty-resolution (disambiguation) process tends to include AUs where the scanpath consists of fixations on the high-entropy word itself and a number of surrounding, low-entropy words. The average entropy in the AU, therefore, is decreased as a result of the impact from these low-entropy words. The more fixations there are in the AU, the more likely it is for this pattern of entropy values to emerge. This seems to indicate that resolution of an instance of uncertainty regarding a translation choice, or the disambiguation of senses, depends on less ambiguous words in the context surrounding the ambiguous word in question.

It can also be seen from the discussion on scanpaths that regarding the processing of a translation-ambiguous word within an AU, there would be an input with high enough entropy (i.e. *cough* in the examples above) to trigger the expenditure of additional cognitive effort. This input tends to increase the average entropy of the AU, but as cognitive effort begins to be expended in relation to the high-entropy input, the average entropy for the AU tends to decrease (see Sect. 4.4). In the meantime, when the high-entropy input increases the average entropy of the scanpath to a certain point (which might indicate that this is beyond the cognitive capacity of the translator to maintain the relevant AU, as in AU 5.1 in Example 5), the unit breaks down into smaller ones while the translator switches to a different

type of AU, re-allocating cognitive resources to resolve the uncertainty involved in the high-entropy input and, while doing so, decreases the average entropy in the corresponding AU.

The two effects—the increase of average entropy by the translation-ambiguous input and the decrease of average entropy by the translator's expenditure of cognitive effort—seem to result in a balance in the general average HTra of AUs, so that this value tends to be at a medium level as the AU lengthens.

6 Concluding Remarks

In summary, the present chapter offers a focused piece of research and provides insights into the role of entropy in the translation process, exploring the theoretical justifications by drawing inferences from formulations of resource-allocation processing difficulty in psycholinguistics and examining empirically the entropy values of words fixated in AUs on the basis of the CRITT TPR-DB. In doing so, it also explains, in terms of entropy values, the manner in which contextual information is integrated in the cognitive processing of highly translation-ambiguous items. While the chapter has shown that the examples illustrated are representative of a general trend regarding the same phrasal verb—via statistical means on the basis of a large database (see Sect. 4.1)—it is important to note that the analysis in Sect. 4 is indeed confined to one phrasal verb, *cough up*. It remains to be tested whether the same findings would be consistent when other high-entropy items, which reflect different linguistic phenomena, are analyzed. Therefore, a meaningful avenue for further research would be to systematically analyze different categories of high-entropy items in the ST, in view of the corresponding AUs which are relevant to the resolution of the uncertainty and ambiguity involved in those items. This would contribute to a more comprehensive understanding of the cognitive processing of information regarding translation ambiguity.

Acknowledgments This research is funded by the School of Applied Language and Intercultural Studies, Faculty of Humanities and Social Sciences, Dublin City University.

References

Alves F, Vale D (2009) Probing the unit of translation in time: aspects of the design and development of a web application for storing, annotating, and querying translation process data. Across Lang Cult 10:251–273

Angelone E (2010) Uncertainty, uncertainty management and metacognitive problem solving in the translation task. In: Shreve GM, Angelone E (eds) Translation and cognition. John Benjamins, Amsterdam, pp 17–40

Attneave F (1959) Applications of information theory to psychology: a summary of basic concepts, methods, and results. Henry Holt, Oxford

Balling LW (2014) Production time across languages and tasks: a large-scale analysis using the CRITT translation process database. In: Schwieter JW, Ferreira A (eds) The development of translation competence: theories and methodologies from psycholinguistics and cognitive science. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 239–268

Bangalore S et al (2016) Syntactic variance and priming effects in translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, New York, NY, pp 211–238

Barkhudarov L (1993) The problem of the unit of translation. In: Zlateva P (ed) Translation as social action. Routledge, London, pp 39–46

Bazzanella C (2011) Redundancy, repetition, and intensity in discourse. Lang Sci 33:243–254. https://doi.org/10.1016/j.langsci.2010.10.002

Bennett P (1994) The translation unit in human and machine. Babel 40:12–20

Brysbaert M (1998) Word recognition in bilinguals: evidence against the existence of two separate lexicons. Psychol Belgica 38(3–4):163–175

Carl M (this volume) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Kay M (2011) Gazing and typing activities during translation: a comparative study of translation units of professional and student translators. Metabolism 56:952–975

Carl M, Schaeffer M (2014) Word transition entropy as an Indicator for expected machine translation quality. In: Automatic and manual metrics for operational translation evaluation workshop programme, 2014. Association for Computational Linguistics, Pittsburgh, PA, p 11

Carl M, Schaeffer M (2017a) Sketch of a noisy channel model for the translation process. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirical modelling of translation and interpreting, vol 7. Language Science Press, Berlin, p 71

Carl M, Schaeffer MJ (2017b) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. HERMES 56:43–57

Carl M, Toledo Báez C (2019) Machine translation errors and the translation process: a study across different languages. J Spec Transl 31:107–132

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, New York, NY, pp 13–54

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process. Transl Cogn Behav 2:211–232. https://doi.org/10.1075/tcb.00026.car

Collins M (2014) Information density and dependency length as complementary cognitive models. J Psycholinguist Res 43:651–681. https://doi.org/10.1007/s10936-013-9273-3

De Bruijn ER, Dijkstra T, Chwilla DJ, Schriefers HJ (2001) Language context effects on interlingual homograph recognition: evidence from event-related potentials and response times in semantic priming. Biling Lang Congn 4:155–168

Dijkstra T, Van Heuven WJB (1998) The BIA model and bilingual word recognition. In: Grainger J, Jacobs AM (eds) Localist connectionist approaches to human cognition. Lawrence Erlbaum Associates, Mahwah, NJ, pp 189–225

Dijkstra T, Van Heuven WJB (2002) The architecture of the bilingual word recognition system: from identification to decision. Biling Lang Congn 5:175–197

Dijkstra T, Timmermans M, Schriefers H (2000) On being blinded by your other language: effects of task demands on interlingual homograph recognition. J Mem Lang 42:445–464

Dragsted B (2010) Coordination of reading and writing processes in translation: an eye on uncharted terrotory. In: Shreve GM, Angelone E (eds) Translation and cognition. John Benjamins, Amsterdam, pp 41–62

Dragsted B, Hansen IG (2008) Comprehension and production in translation: a pilot study on segmentation and the coordination of reading and writing processes. Copenhagen Stud Lang 36:9–29

Duffy SA, Kambe G, Rayner K (2001) The effect of prior disambiguating context on the comprehension of ambiguous words: evidence from eye movements. In: Gorfein DS (ed) On

the consequences of meaning selection: perspectives on resolving lexical ambiguity. American Psychological Association, Washington, DC, pp 27–43
Favreau M, Segalowitz NS (1983) Automatic and controlled processes in the first- and second-language reading of fluent bilinguals. Mem Cognit 11:565–574
Gernsbacher MA (1990) Language comprehension as structure building. Lawrence Erlbaum, Hillsdale, NJ
Gernsbacher MA (1997) Attenuating interference during comprehension: the role of suppression. In: Medin DL (ed) The psychology of learning and motivation: advances in research and theory, vol 37. Academic, San Deigo, CA, pp 85–104
Grosjean F (1997) The bilingual individual. Interpreting 2:163–187
Hale J (2001) A probabilistic Earley parser as a psycholinguistic model. In: Second meeting of the north American chapter of the Association for Computational Linguistics, Pittsburgh, Pennsylvania, 2001. Association for Computational Linguistics, Pittsburgh, PA, pp 1–8
Heilmann A, Llorca-Bofí C (this volume) Analysing the effects of lexical cognates on translation properties: a multi-variate product and process based approach. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Hervey S, Higgins I, Loughridge M (1995) Thinking German translation: a course in translation method: German to English. Routledge, London
Hvelplund KT (2016) Cognitive efficiency in translation. In: Martín RM (ed) Reembedding translation process research. John Benjamins, Amsterdam
Jakobsen AL (2019) Segmentation in translation: a look at expert behavior. In: Li D, Lei VLC, He Y (eds) Researching cognitive processes of translation. Springer, Berlin, pp 71–108
Kockelman P (2013) Information is the enclosure of meaning: cybernetics, semiotics, and alternative theories of information. Lang Commun 33:115–127. https://doi.org/10.1016/j.langcom.2013.01.002
Kondo F (2007) Translation units in Japanese-English corpora: the case of frequent nouns. In: Proceedings from corpus linguistics conference series (Birmingham, 27–30 Jul 2007). University of Birmingham, Birmingham
Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. Kent State University Press, Kent, OH
Kroll JF, Stewart E (1994) Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. J Mem Lang 33:149–174
Kullback S (1959) Information theory and statistics. John Wiley & Sons, Hoboken, NJ. Republished by Dover publications in 1968; reprinted in 1978
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
Lacruz I, Ogawa H, Yoshida R, Yamada M, Ruiz Martinez D (this volume) Using a product metric to identify differential cognitive effort in translation from Japanese to English and Spanish. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Levý J (1967) Translation as a decision process. In: Levý J (ed) To honor Roman Jakobson: essays on the occasion of his seventieth birthday, vol II. Mouton, The Hague, pp 1171–1182
Levy R (2008) Expectation-based syntactic comprehension. Cognition 106:1126–1177
Levy R (2013) Memory and surprisal in human sentence comprehension. In: van Gompel RPG (ed) Sentence processing. Psychology Press, London, pp 90–126
Levy R, Gibson E (2013) Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. Front Psychol 4:229. https://doi.org/10.3389/fpsyg.2013.00229
Macizo P, Bajo MT (2006) Reading for repetition and reading for translation: do they involve the same processes? Cognition 99:1–34
MacKay DM (1969) Information, mechanism and meaning. MIT Press, Cambridge
Malmkjær K (2006) Translation units. In: Brown K (ed) The encyclopedia of languages and linguistics. Elsevier, Amsterdam, pp 92–93
Marian V, Spivey M, Hirsch J (2003) Shared and separate systems in bilingual language processing: converging evidence from eyetracking and brain imaging. Brain Lang 86:70–82
Nida EA, Taber CR (1969) The theory and practice of translation, vol 8. Brill, Leiden

Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and source-text background information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Prior A, Wintner S, Macwhinney B, Lavie A (2011) Translation ambiguity in and out of context. Appl Psycholinguist 32:93–111

Rabadán R (1991) The unit of translation revisited. In: Larson ML (ed) Translation: theory and practice: tension and interdependence. State University of New York at Binghamton, New York, NY, pp 38–48

Ruiz C, Paredes N, Macizo P, Bajo MT (2008) Activation of lexical and syntactic target language properties in translation. Acta Psychol (Amst) 128:490–500

Schaeffer M, Carl M, Lacruz I, Aizawa A (2016a) Measuring cognitive translation effort with activity units. In: Proceedings of the 19th annual conference of the European Association for Machine Translation, pp 331–341

Schaeffer M, Dragsted B, Hvelplund KT, Balling LW, Carl M (2016b) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, New York, NY, pp 183–210

Schilperoord J (1996) It's about time: temporal aspects of cognitive processes in text production. Rodopi, Amsterdam

Schwartz AI, Kroll JF (2006) Bilingual lexical activation in sentence context. J Mem Lang 55:197–212

Shannon CE (1948) A mathematical theory of communication. ACM SIGMOBILE Comput Commun Rev 5:3–55. https://doi.org/10.1145/584091.584093

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana, IL

Swadesh M (1960) On the unit of translation. Anthropol Linguist 2:39–42

Tirkkonen-Condit S (1993) Choice in translation: a challenge to theory and practice. In: Tirkkonen S, Laffling J (eds) Recent trends in empirical translation research. Faculty of Arts, University of Joensuu, Joensuu, pp 5–9

Tokowicz N, Kroll JF (2007) Number of meanings and concreteness: consequences of ambiguity within and across languages. Lang Cognit Process 22:727–779

Wei Y (2018) Cognitive effort in relation to word translation entropy and syntactic choice entropy. In: 5th international conference on cognitive research on translation and interpreting, Renmin University of China, Beijing

Wiener N (1948) Cybernetics, or, control and communication in the animal and the machine. MIT Press, New York, NY

Wiener N (1954) The human use of human beings: cybernetics and society. Avon Books, New York, NY

Wu YJ, Thierry G (2012) Unconscious translation during incidental foreign language processing. Neuroimage 59:3468–3473

Analyzing the Effects of Lexical Cognates on Translation Properties: A Multivariate Product and Process Based Approach

Arndt Heilmann and Carme Llorca-Bofí

Abstract The translation of cognates has received renewed interest in Translation Process Research (Oster (2017) Empir Model Transl Interpret 7:23; Hansen-Schirra et al. (2017) Predicting cognate translation. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 3–22) but tends to be relatively time-consuming due to the manual identification of cognates and their translations. On the basis of work by Heilmann (Profiling effects of syntactic complexity in translation: a multi-method approach. PhD thesis, 2021) and the structure of the TPR-DB, we devised a relatively simple way to determine the cognate status of ST words and detect literal, cognate translations of cognates. We assess the "cognateness" of ST items on the basis of formal (dis)similarity with aligned ST and TT words. We use these measures to show how the cognate status of ST words and the literal cognate-to-cognate translation of cognates affect properties of the translation product and process. Using multivariate statistics, we are able to show that a ST token's cognate status is a determining factor of translation ambiguity. We also find evidence for cognates affecting reading and typing behavior during translation. Additionally, we observe a moderation effect of translation experience on behavioral measures when cognates are translated literally. We interpret the results in support of the monitor model (Tirkkonen-Condit (2005) Meta 50(2):405–414) and propose a refinement of the operationalization of literality (cf. Schaeffer and Carl (2014a) Measuring the cognitive effort of literal translation processes. In: Proceedings of the EACL 2014 workshop on humans and computer–assisted translation, pp 29–37).

A. Heilmann (✉)
RWTH Aachen University, Aachen, Germany
e-mail: arndt.heilmann@ifaar.rwth-aachen.de

C. Llorca-Bofí
Universitat de València (University of Valencia), València, Spain

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_8

1 Introduction

Translations have been hypothesized to be the result of a number of translation universals (Baker 1996). In this paper, we will focus mainly on the universal known as shining through (Teich 2003) and, partially, normalization. Shining through encompasses two phenomena. In one sense, it can refer to a transfer of features of the source language (SL) to the target text (TT) (called *"genuine" shining through* (Evert and Neumann 2017)). In a second sense, it may refer to a transfer of features of a particular source text (ST) to the TT (*"individual" shining through* (Evert and Neumann 2017)). Normalization is associated with the translator's (over-)reliance on typical TL (target language) features (Hansen-Schirra 2011; Oster 2017). In this study we attempted to better understand the reasons for individual shining through phenomena using product and process data from translation experiments. This transfer of ST features to the translation is closely related to notions of priming and ST interference where ST forms impinge on the translator and affect his or her translation by making it more similar to the ST (cf. Ivir 1981). For Toury (2012), the observed similarities between STs and TTs can be described with the help of a translational law of interference. According to this law, "phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (Toury 2012, 311) and these phenomena seem to "tend to force themselves on the translator" (Toury 2012, 311).

A very visible case of ST interference is the translation of SL cognate words by a matching TL cognate. In a strict sense, cognates are words in the SL that completely overlap in form and meaning with words in the TL. For psycholinguistic purposes, however, it makes sense to use a looser definition of cognates that allows for gradual overlaps of form and meaning with words in the TL (Tokowicz et al. 2002, 437).

Toury relates the interference of ST phenomena to the cognitive effort of the translator. He proposes that in the case of interference, "the establishment of an interference-free output (or even an output where interference has been relegated to domains which are regarded as less disturbing) necessitates special conditions and/or special efforts on the translator's part" (Toury 2012, 311) and assumes further that "accomplished translators would be less affected by its [the source text's] actual make-up" (Toury 2012, 313). This means that the use of cognates in translation can be seen as kind of an automatic procedure and that in order to prevent it from happening, the translator will have to invest effort and apply of some kind of control mechanism to avoid cognate translations (e.g. see Oster 2017). While SL cognates do share formal and semantic aspects with TL items, they are not necessarily interchangeable in all contexts and linguistic items other than cognates may actually be a semantically more adequate translation option. For example, Hansen-Schirra et al. (2017) found that in translations from English to German, English/German cognates were re-translated as cognates in only 37 percent of the cases when provided in the context of a full text. In contrast, 57 percent of the cognates were re-translated as cognates when the cognates were presented in a mere list (Hansen-Schirra et al. 2017, 11). This is in line with Heilmann et al. (2019) who

also found that stimuli were translated in the most literal manner when context was scarcest. While interference is a "matter of cognition" (Toury 2012, 311), this shows that other factors are likely to play a considerable role as well and will affect the literal, cognate-to-cognate translation of a ST's cognates. For instance, the target culture's acceptance of interference in translations from a certain SL into the TL may differ for different language pairs. The extent to which interference is allowed in the translation is likely restricted be the translator's (or target culture's) estimated prestige of SL and TL (Toury 2012, 311–313) and there is some tentative evidence for the possible influence of prestige on shining through (Evert and Neumann 2017). In a case of high(er) prestige of English the translator's linguistic behavior will be more similar to the SL than to the TL, which would become apparent in the over-use of some ST features and the underuse of more typical TL features.

We will focus on explanations of cognate translation that are related more closely to questions of (bilingual) language representation and processing; while acknowledging that other, less immediate factors have the potential to affect the translation of cognates too.

To a large extent, interference effects that are due to the online processing of cognates can be accounted for by psycholinguistic insights in the organization of the bilingual language system (Kroll et al. 2010; Paradis 2004). Particularly, the notion of spreading activation as a retrieval mechanism for linguistic items from the mental lexicon plays a role here (Dell 1986).

According to Paradis' model of language comprehension, comprehension proceeds independently from the language that was used to encode a word (Paradis 1997, 203). Linguistic forms contained in all language (sub-)systems receive activation during the perception of a word if their form resembles the one that is currently perceived (see also Dijkstra and Van Heuven's bilingual interactive activation model (1998) for a Connectionist perspective on the phenomenon). The language pertaining to that word is inferred shortly after this process via the reader's metalinguistic knowledge and processed at the conceptual level (Paradis 1997, 206). Paradis calls this the direct access hypothesis (Paradis 2004, 203). Similar views are expressed for instance by Diependaele et al. (2013), who refer to this phenomenon as *interactive activation*. Sharing conceptual features in addition to formal ones, cognates will be highly activated. Experiments have shown facilitation effects during the processing of cognates in a variety of tasks. For instance, priming experiments by Voga and Grainger (2007) have suggested that quicker responses in lexical decision tasks are mediated by additional activation from the level of form, rather than cognates having "any special representational status in a bilingual's mental lexicon" (Voga and Grainger 2007, 946), corroborating the view expressed by Paradis. Cross-lingual priming effects in translation i.e. the choice to use a TL cognate for an SL word can be explained by residual activation in the extended language system from comprehension. Translators constantly switch from reception in the SL to production in the TL and they may still be affected by lingering activation in the extended system (or ongoing activation from the visual perception) (Christoffels and De Groot 2005, 462). It is possible that frequent co-activation of SL and TL cognate words will create and strengthen the neuronal links between

cognates (De Groot 2011, 320) and thus make co-activation of ST and TT cognates even stronger. In order to prevent the choice of a cognate translation, the translators will have to counteract accidental inappropriate selections by increased inhibition of unwanted options and increased self-monitoring to catch miss-selections.

De Groot (1992) found that "words similar in form to their translations are easier to translate than words dissimilar to their translations" (De Groot 1992, 1019). Part of the cognate facilitation effect can be explained by the elimination of competition. Cognates may receive such a strong activation from both form and meaning related cues that other, less active options are not viable candidates. Laxén and Lavaur (2010), Boada et al. (2013), and Prior et al. (2007) showed that subjects' responses in translation recognition and word translation tasks were slower for words with multiple translation alternatives. The more translation choices there are available, the lower the response time in oral word translation tasks (Tokowicz and Kroll 2007). Dragsted (2012) counted the number of translation alternatives stemming from multiple translators from a translation experiment and took them as an estimation of the possible translation alternatives. She found that reading time measures and pauses increased with the variety of translation solutions. This effect was replicated by Schaeffer and Carl (2014a) who used a word translation entropy measure to operationalize translation ambiguity or rather the selection pressure associated with such a set of possible choices. Schaeffer and Carl (2014a) proposed that word translation entropy can be a way of measuring the literality of translation, along with a measure of word-order distortion. We believe that a concept of literality may benefit from a revised operationalization that includes a measure of formal similarity at the lexical level (cf. Halverson 2019). We will come back to the issue of operationalizing literal translation later and propose some ideas to further develop Schaeffer and Carl (2014a)'s literality concept (see also Carl: this volume, chapter "Information and Entropy Measures of Rendered Literal Translation").

To summarize the above points: The data raised before suggests that cognates in the ST are prompting the translator to re-translate SL cognates by their respective TL cognates. This kind of interference effect will cause ST features to appear in the TT (shining through). While contextual and lingua-cultural factors play a role in the translation of cognates, the primary mechanism of cognate translation seems to reside in cognition; particularly processes of comprehension and production sharing the extended linguistic system as a resource and language non-specific access of lexical items on the basis of orthographic (or auditory) cues. Thus, activation of cognate TL items from concurrently active form and meaning cues in the language system as well as possible direct neuronal links between cognate words have the potential to supersede the activation of alternatives with mere semantic overlap.

We want to substantiate the findings from psycholinguistic experiments using more ecologically valid translation data. While translation process studies on cognates tend to work with full texts, they tend to rely mainly on mono-variate statistical analyses of the effects of cognates. This makes it difficult to ascertain if indeed effects of cognates are responsible for the observed effects or other (lexical) confounders. With the help of multivariate statistics, we address the following four interrelated hypotheses:

1. Product Based Hypotheses:
   (a) ST cognates reduce the number of translation alternatives
   (b) Cognate-to-cognate translations contribute to the reduction of translation alternatives
2. Process Based Hypotheses:
   (a) ST cognates facilitate translation
   (b) Cognate-to-cognate translations contribute to this facilitation

Facilitation effects will be measured with the help of eye tracking and keystroke logging measures. In case of facilitation, we should find shorter reading and typing durations for cognates and/or cognate-to-cognate translations. When encountering a cognate of the SL in a ST, the translator is basically confronted with two translation strategies: (1) the translation of the cognate by its respective TL cognate (e.g. *evidence* (engl.) by *Evidenz* (Ger.)) or (2) a translation that avoids the use of a TL cognate (e.g. choosing a non-cognate word (e.g. *evidence* (engl.) by *Beweis* (Ger.) or even the non-translation of an ST's cognate)). The perspective on the cognate status of the ST words allows us to predict the effects of cognateness in general (i.e. the effects of an SL word sharing formal and semantic similarity). The latter offers a refined perspective as to what actually happened in the translation in reaction to this similarity.

In the following section, we present our method of measuring formal correspondences and present the data and statistical means used to test the tool on translation process and product data.

2 Measuring Formal Correspondence

Literal translation involves a wide range of phenomena and this makes it very difficult to find an operationalization of this vague concept. A promising attempt to characterize and operationalize central aspects of the concept of literality is the approach by Schaeffer and Carl (2017). They interpret literality as a combination of three factors: One factor is the number of available translation choices for a specific source text item. The fewer possible translation solutions a source word has, the higher the literality. Another factor is word-order correspondence of ST and TT. If word order overlaps, this can be counted as a sign of literality as well. The last aspect they cover in their literality assessment are alignment pairs. A single ST word that translates as a single word in the TL can be seen as more literal than a single ST word that requires multiple TL words for its translation. We follow Halverson's (2019) suggestion that it may be beneficial to think of literality also in terms of formal correspondence and thus ways that "can capture degrees of formal similarity also at other levels (e.g., the lexical level). The latter is of interest given evidence that translators may avoid cognates, which would otherwise be obvious choices due to priming" (Halverson 2019).

We propose a simple way to measure formal correspondence between ST and TT at the lexical level using data readily made available by the TPR-DB. The method is a simplification of the cognate assessment used in Heilmann (2021) and relies merely on orthographic correspondence. Following Heilmann (2021), we operationalize not only formal correspondences of ST words and their translations but also a graded estimate (the *cognateness* of a word if you will) by how much formal overlap a SL word and a TL word can *potentially* share. This is an attempt to measure the *cognateness* or cognate status of a given word and follow the reasoning of De Groot (1992) that the amount of shared features is important. De Groot, for instance, used graded estimates of orthographic, semantic, and phonological similarities of words that she acquired with the help of human bilingual raters. We restrict ourselves to orthographic similarity here which can be determined in an automatized fashion with relatively little effort thanks to the structure and information of the TPR-DB. The basis for our measure of (dis)similarity is the Levenshtein distance. This measure is an editing distance that calculates the smallest number of character insertions, deletions, or substitutions required to change one word into another. The method thus presupposes that SL and TL share a fair amount of orthographic convention to work well (though see Heilmann (2021) for a proposal to (also) use auditory transcription of ST and TT words with the Levenshtein distance).

The highest Levenshtein distance that two sequences can achieve is limited by the length of the longer of two compared strings. For example, *chair* and *top* have a Levenshtein distance of 5 because all 3 characters of *top* have to be substituted and 2 have to be added to convert *top* to the five letter word *chair*. The Levenshtein distance of *bat* and *cat* on the other hand is but 1 because only one character has to be substituted. Following Heilmann's (2021) procedure we normalized the editing distance by dividing the Levenshtein distance of a token pair by the length of the longer sequence so that the resulting normalized distance measured dissimilarity from 0 to 1. Thus, 1 is maximal dissimilarity as in *chair* and *top*.

However, since we are looking for similarity rather than dissimilarity, formal similarity is then calculated as:

$1 - Norm.Levensthein(STToken, TGroup)$. In this case, 1.0 means maximal similarity and 0.0 no similarity at all. For example, a German translation of the English word *beer* with its German cognate *Bier*[1] would be: $1 - Norm.Levensthein(beer, bier) = 0.75$. Thus, the formal overlap is 0.75 which indicates a relatively high similarity rating as opposed to English *add* and its translation *Werbung* $1 - Norm.Levensthein(add, werbung) = 0.0$.

With the help of a Python script all translations of a token were collected from the TPR-DBs source token table (.st-tables). We restricted ourselves here to a fraction of the multiLing study subset of the TPR-DB (BML12, SG12, KTHJ08). These tables contain a large number of measures pertaining to each ST token gathered across

[1]Due to German spelling conventions noun initial letters are always capitalized but for all our comparison we ignored upper case.

Table 1 German translations pertaining to the lemma *academic* by the participants of study SG12 and the associated similarity ratings. The cognate rating of *academic* for the language pair English–German is 0.636

Lemma	SToken	TGroup	Translator	1-NormLev. = FormalSimilarity
Academic	Academic	Wissenschaft	P09_SG12	0.167
		Wissenschaft	P22_SG12	0.167
		wissenschaftliche	P01_SG12	0.235
		wissenschaftliche	P02_SG12	0.235
		wissenschaftliche (x 13)	ssL . . .	. . .
		wissenschaftliche	P23_SG12	0.235
		wissenschaftliche	P24_SG12	0.235
		akademische	P03_SG12	0.636
		akademische	P06_SG12	0.636
		akademische	P15_SG12	0.636
		akademische	P20_SG12	0.636
Max. FormalSimilarity = CognateRating:				0.636

different translation tasks. We used the columns *Lemma*, *SToken*, and *TGroup* for our evaluation of similarity. *Lemma* contains the uninflected form of each ST token, as identified by TreeTagger (Schmid 1995). We grouped all ST tokens belonging to a lemma under the respective lemma, along with each aligned translation (TGroup) of the ST token (see Table 1). For each language pair, we evaluated the similarity of each TGroup with each SToken. In order to determine the cognate status of a ST word, we gathered all translations of ST lemma in a specific language pair and searched for the highest similarity among all translations of a SToken in a language pair (see Table 1). This similarity rating served as an approximation of the cognateness of a ST token's lemma.

The more translations there are of a given ST token, the better the accuracy of the measure due to reduction of a sampling error. If a ST token has indeed a cognate translation but none of the translators uses it, it may be a sign that the semantic similarity is very low and/or the cognate translation may be contextually completely inappropriate.

3 Data and Participants

We analyze the effects of cognates (or rather cognate status) on the translation product and process with the help of multivariate statistical means. For these analyses, we used a subsample of the multiLing data set of the TPR-DB. The multiLing data set contains multilingual translations of the same set of six English ST into Hindi, Japanese, Chinese, Danish, German, Spanish. These translations consist of translations from scratch, post-edited machine translation (MT) and edited MT without access to the ST. Due to the dependence of our method to determine

formal similarity on the basis of orthographic similarity we were restricted to use the studies BML12 (Spanish), KTHJ08 (Danish), and SG12 (German) from the TPR-DB, which share most of their alphabet with English. The statistical analysis of the data is restricted to translation from scratch. It may be interesting to see how the results obtained for this mode of translation compare to other modes of translation.

For our analyses, we used R (R Core Team 2017) and the package lme4 (Baayen et al. 2008) for (generalized) linear mixed regression modelling. To test the statistical significance of the effects of our measures of interest, we used the R package lmerTest (Kuznetsova et al. 2015) and used type-II ANOVAs to calculate the statistical significance of our findings. The lmerTest-package uses Satterthwaite approximation to estimate the degrees of freedom. Kurtosis and skewness were calculated with the help of the package *moments* (Komsta and Novomestky 2015) and for the calculation of R^2 we used the *MuMIn* package (Barton 2009). To check for multicollinearity issues, we included variance of inflation factors (VIFs). These were calculated with the help of the vif.mer() function (Frank 2014).

We ran each linear mixed regression model with and without outlying data points to assess if the results were affected by overly influential data points. Outliers were identified as data points whose model residuals exceeded 3 residual standard deviations. If the removal of the outliers changed the statistical significance of our dependent variables in the model or resulted in a sign change, we report both results for reasons of transparency. All linear models were checked for multicollinearity, skew, and kurtosis of residuals. Skewness of $>|2|$ and kurtosis >7 were selected as indication of a severe deviation from the normality assumption regarding model residuals (Kim 2013). VIFs and model fits are reported in the results tables to increase readability. Note that our dependent variables were log-transformed when the distribution of residuals was notably right skewed.

The models used the categorical variables text, item, and participant as random variables. The model predicting translation ambiguity was run without participant as a random effect because translation ambiguity in the form of HTra is calculated across participants and thus there is no idiosyncratic behavior with respect to translation ambiguity (though there is for translation (self) information, e.g. a translator's habit of always picking an untypical translation). Note that we modelled Study as a fixed, rather than a random effect—though conceptually the latter would be more correct. We modelled it this way because having only three studies in our data set, our random effect would only have three levels. However, in order to work as intended, a random effect needs far more levels (>6) to reliably estimate random variance accurately (Harrison et al. 2018).

We excluded all tokens that were identified as symbols and cardinals, e.g. full stops, commas, and all kinds of numbers from the analysis as they were likely to skew the results regarding the cognateness: It is very likely that a full stop is translated as a full stop and a number as the same number which would likely affect the interpretation of more literal cognate-to-cognate translations.

Our variables of interest are the cognate ratings of the ST tokens[2] and the formal similarity of translations with their aligned source (i.e. literal, cognate translations vs. non-cognate translation). For both variables we expect that experience will moderate the translators' responses. Therefore, we added the years of professional translation experience as an interaction effect with the two variables of interest.

- **CognateRating**: The maximal formal similarity of ST item and its translations (see max. values in Table 1. This variable measures the cognate status of a lexical item and thus operationalizes a property of the linguistic system rather than a property of the ST.
- **FormalSimilarity**: This variable operationalizes the formal similarity of a particular translation with its aligned ST token (see individual ratings in Table 1). It thus measures a property of the TT (in relation to the ST); i.e. the result of choosing a more (or less) literal translation strategy.
- **ExperienceYears**: We controlled for the experience of the participants by including the number of years as a professional translator. Note that the meta data of KTHJ08 referred to translators with less than two years of experience simply as "<2" years. We substituted this with the value 0 in order to be able to use an interval-scaled variable.

A number of control variables were included to avoid confounding. Following De Groot (1992, 1011), we control for a number of factors that can affect word translation performance. We controlled for the effect of word length, imageability (in form of concreteness ratings), and word familiarity (via word frequency).[3] The same factors have not only been shown to influence translation performance but also translation probabilities (Prior et al. 2007). The latter authors also factor in the part of speech of the source token, which we will be doing as well. Having controlled for a number of lexical characteristics, our study is one of the few that studies the effect of cognates on translation in an ecological valid setting but still exerts controls over likely confounds of cognate processing and literal translation. Lastly, we added a variable tracking the translators' progress through the ST (using the sequential numbering if ST segments) (see Schaeffer and Carl 2017) and control for the fact that some of the translated ST words are named entities.

- **LenS**: The length of a source text token (in characters).
- **Concreteness**: We used Brysbaert and Diependaele's (2013) list of concreteness ratings to ward off potential concreteness effects that may affect mono-lingual and bilingual processing. Concreteness (or imageability) affect translation ambiguity (Prior et al. 2007) and word translation performance (De Groot 1992).

[2]The cognate ratings are of course calculated for each language separately i.e. for the Spanish portion the ST token *victims* received a high cognate rating of 0.75 (*víctimas*) whereas the cognate rating for Danish and German was 0.14 and 0 respectively because there is no respective cognate translation.

[3]De Groot (1992, 1011) also controlled for context availability, but since our translations were produced in rather ecologically valid settings with full-text translations, this was unnecessary.

The ratings are provided on a Likert-like scale, ranging from 1–5. The higher the rating, the more concrete the token was rated. For a few lemmas not covered by the list, we used their method to collect the missing values. We equate concreteness with imageability here, assuming that if there is a difference between them, the overlap of imageability and concreteness is sufficiently high.

- **PoS**: The parts-of-speech of the ST tokens were modelled as a categorical predictor which summarized the very detailed PoS categories from the .st-tables to more general labels such as *noun*, *(full)verb*, *modal*, *adjective*, *adverb*, *preposition*, *conjunction*, and *determiner*. PoS not fitting either category were grouped under a category called *other*.
- **Names**: The variable *Name* was modelled as a binary variable (TRUE/FALSE) to inform our linear models that a given token is a name. We can expect very low translation ambiguity and very high formal similarity of ST tokens and TT tokens for many of these items. We tagged all tokens that were proper nouns such as *Colin* or *Darfur* as NameTRUE to filter out this effect.
- **Prob1**: Word frequency is a known influence on reading behavior. The interval-scaled variable frequency uses the frequency information (Prob1) from the .st-tables of the TPR-DB. The frequency information is taken from the BNC (Carl et al. 2016) and a proxy for possible entrenchment effects that may affect translation ambiguity and reading speed.
- **Study**: We use *Study* with the three levels BML12, KTHJ08, and SG12 as a fixed effect to control for the language and study specific variation.

3.1 Prediction of Translation Choice

If indeed formal correspondence leads to stronger activation of cognates than other items, we should see a more homogeneous set of translations for SToken with a high cognate rating. In order to assess the homogeneity of translation we used TPR-DB's word translation entropy measure (HTra) and predicted it by the variable *CognateRating*. The model residuals were fairly normally distributed (skewness: 0.03, kurtosis: 4.1) and there was no sign of heteroskedasticity. The removal of outliers did not affect the interpretation of the results, so we interpret and present the model with all data points below.

The results of the model (see Table 2) suggest that the cognate rating has a small statistically significant effect on the number of translation options, indicating a slight trend towards a lower amount of translation choice when cognates are involved. Thus, the cognate status of the ST token seems to be indeed a co-determining factor of the amount of translation choice. However, the model cannot show how the reduction of translation choice occurs. While we can assume that this is partially due to cognate-to-cognate translation, it may be equally plausible to assume that the reduction of translation choices has been due to the translators' shared attempts to avoid cognate translations and settle for a shared non-cognate translation (see the translation 'wissenschaftlich' in Table 1).

Table 2 Model results from predicting HTra by Cognate Rating

HTra by CognateRating					
(R^2m: 0.1/R^2c: 0.72)	Estimate	SE	*F*	p	VIF
(Intercept)	1.332	0.207	–	–	–
Concreteness	0.048	0.046	1.086	0.298	1.913
***Study**KTHJ08*	−0.065	0.01	184.236	<**0.001**	1.239
***Study**SG12*	0.141	0.01			1.234
***PoS**Adv*	0.036	0.187	8.246	<**0.001**	1.489
***PoS**Conj*	−0.735	0.217			1.544
***PoS**Det*	0.041	0.169			2.516
***PoS**Modal*	−0.269	0.313			1.142
***PoS**Noun*	−0.091	0.131			3.693
***PoS**Other*	0.521	0.206			1.808
***PoS**Prep*	0.086	0.156			2.717
***PoS**Pro*	−0.007	0.222			1.381
***PoS**Verb*	0.504	0.133			2.609
Prob1	−0.021	0.013	2.591	0.108	1.278
LenS	0.022	0.016	1.95	0.163	1.861
NameTRUE	−0.935	0.217	18.617	<**0.001**	1.066
CognateRating	−0.081	0.027	9.045	<**0.01**	1.028

3.1.1 Predictability of (Cognate) Translation Choices

In order to settle this issue, we wanted to find out whether cognate translation solutions constituted a considerable number of the translation solutions made by the translators. Since HTra generalizes over different translation solutions, we decided to use a closely related measure i.e. word translation (self) information (ITra for short).

In this context, ITra can be used to measure the predictability of a specific translation solution given a particular ST token. HTra on the other hand measures the heterogeneity of the translation solutions of a ST token. HTra operationalizes selection pressure associated with a specific ST word very well, whereas ITra operationalizes the likelihood or predictability of a translation option. Self-information is actually a component that is used as part of the calculation of the entropy value (see the underlined portion of the entropy formula (H) below). Information is a very prolific concept in mono-lingual reading studies and is used to model mono-lingual reading behavior (Frank and Thompson 2012; Demberg and Keller 2008). In this research it is commonly referred to as surprisal i.e. the higher the self-information the more surprising/less predictable the next word.

$$ITra(x) = -\log_2 p(x) \tag{1}$$

$$H = \sum_{i=1}^{n} p(x_i) * \underline{-\log_2 p(x_i)} \tag{2}$$

The higher the word translation information (ITra), the less predictable and infrequent a translation solution is. A typical translation option will have a high predictability/low information i.e. one can expect the translator to translate a ST word in a specific way. In order to calculate the self-information, we used data from post-editing, editing, and translation from scratch provided by a subset of the TPR-DB's *multiling* study (see Sect. 3) and used the translation probability variable (ProbT) from the TPR-DB. To make the concept more relatable, we refer back to Table 1. The ITra of "wissenschaftliche" is very low (0.30) and thus can be termed the standard solution, whereas "akademische" is much less predictable or standard (1.75)—despite being a cognate of "academic." The change of word class from "academic" (adjective) to "Wissenschaft" (noun) is even less predictable (2.44).

The average value of ITra in the data set was 1.4 (sd: 1.2) over all languages. By using HTra alone, it would be impossible to determine whether one translator behaved less typical than another translator when translating the same ST token. HTra in this sense is similar to our CognateRating variable. It operationalizes a property of the (currently active) language system, while the variable FormalSimilarity rating actually captures the result of the activation of the language system i.e. the similarity of a particular translation with a ST word. It is possible to argue that unpredictable (or rare) translation options are less literal than those that have a high predictability. They may be seen as more creative. In this context, predictability shares aspects of the notion of default or standard. If we wanted to restrict the notion of default to initial processing, process measures would have to be included in this case too (Halverson 2019). We will come back to the idea of literality at a later stage again.

We predicted the self-information of translation solutions by FormalSimilarity of translations with the ST items to estimate the predictability of cognate-to-cognate translations. FormalSimilarity and ExperienceYears were modelled in form of an interaction effect to probe if experienced translators were differently likely to opt for a cognate solution compared to less experienced translators. The model's residuals were fairly normally distributed (skewness: 0.43, kurtosis: 2.6) and they did not show signs of heteroskedasticity. The removal of outliers did not have a noteworthy effect on the results so we left the outlying points in the model (see Table 3).

Indeed, we found a statistically significant negative correlation between FormalSimilarity and ITra meaning that cognate-to-cognate translation solutions are more predictable. Note that this does not mean that translators will always opt for the cognate, but rather that translators are not capable of escaping the reflex of cognate translation completely and will opt for it in a considerable number of cases on the basis of orthographic similarity. It would be interesting to see if the effects were stronger if we took phonetic (Heilmann 2021) and semantic similarity/overlap into account. Semantic similarity may be assessed using Continuous Vector Space Models in the future for example (see Carl: this volume, "Translation Norms,

Table 3 Model results for the prediction of ITra by FormalSimilarity

ITra by FormalSimilarity					
(R^2m: 0.06/R^2c: 0.51)	β	SE	F	p	VIF
(Intercept)	1.364	0.151	–	–	–
Concreteness	0.029	0.033	0.777	0.378	1.919
StudyKTHJ08	−0.519	0.018	424.791	<**0.001**	1.318
StudySG12	−0.13	0.018			1.243
PoS*Adv*	−0.094	0.135	6.324	<**0.001**	1.488
PoS*Conj*	−0.819	0.155			1.547
PoS*Det*	−0.277	0.121			2.537
PoS*Modal*	−0.486	0.226			1.142
PoS*Noun*	−0.087	0.094			3.714
PoS*Other*	−0.027	0.148			1.812
PoS*Prep*	−0.148	0.112			2.732
PoS*Pro*	−0.177	0.159			1.385
PoS*Verb*	0.127	0.096			2.625
Prob1	−0.033	0.01	11.821	<**0.01**	1.286
LenS	0.041	0.011	13.148	<**0.001**	1.865
NameTRUE	0.102	0.155	0.432	0.511	1.086
zFormalSimilarity	−0.476	0.011	1897.228	<**0.001**	1.065
zExperienceYears	0.021	0.007	9.03	<**0.01**	1.068
zFormalSimilarity:zExperienceYears	0.002	0.007	0.1	0.752	1.002

Translation Behavior, and Continuous Vector Space Models"). Stronger reductions in self-information may still be achieved rather by settling for a common (non-cognate) option. Following the gravitational pull hypothesis (Halverson 2003), this TL option may be a linguistic item that occurs relatively frequently in the TL and due to the resulting entrenchment becomes the most likely candidate for selection after the inhibition of cognates.[4]

3.2 Prediction of Translation Process Properties

Having determined how CognateStatus and FormalSimilarity affect translation choice, we now investigate the effects of cognates and strategies of cognate translations on the translation process. The following model investigates how

[4]The example of the German translations of "academic" in Table 1 as the most frequent solution "wissenschaftliche" instead of "akademisch" may be a result of the gravitational pull of "wissenschaftlich." A quick Google search revealed only 1.400.000 results for *akademisch* and 16.600.000 results for *wissenschaftlich*. This frequency difference would support this idea.

Table 4 Model results from predicting Total Reading Time of the ST (TRTS) by CognateRating

TrtS by CognateRating					
(R^2m: 0.2/R^2c: 0.41)	Estimate	SE	*F*	p	VIF
(Intercept)	4.864	0.125	–	–	–
Concreteness	0.011	0.016	0.484	0.487	1.933
***Study**KTHJ08*	0.157	0.02	1521.635	<**0.001**	1.615
***Study**SG12*	0.885	0.017			1.427
TaskT	0.638	0.016	1556.051	<**0.001**	1.172
***PoS**Adv*	−0.108	0.066	2.925	<**0.01**	1.488
***PoS**Conj*	−0.066	0.077			1.556
***PoS**Det*	−0.084	0.059			2.507
***PoS**Modal*	−0.165	0.112			1.142
***PoS**Noun*	−0.047	0.046			3.755
***PoS**Other*	−0.255	0.075			1.744
***PoS**Prep*	−0.191	0.055			2.682
***PoS**Pro*	−0.194	0.078			1.373
***PoS**Verb*	−0.026	0.047			2.662
Prob1	−0.015	0.005	10.118	<**0.01**	1.275
LenS	0.099	0.005	328.584	<**0.001**	1.827
NameTRUE	0.288	0.074	15.22	<**0.001**	1.136
HTra	0.051	0.008	36.752	<**0.001**	1.08
zCognateRating	−0.034	0.01	12.429	<**0.001**	1.152
zExperienceYears	−0.102	0.008	155.297	<**0.001**	1.065
STseg	−0.032	0.006	30.433	<**0.001**	1.022
zCognateRating:zExperienceYears	0.007	0.006	1.212	0.271	1.001

cognates and cognate translations strategies influence the translation process with respect to reading times and translation duration as well as editing behavior.

Our next model predicts the Total Reading Time of the ST token (TrtS) by the cognate rating of the ST token while controlling for various confounds (see Table 4).

The variable TrtS was log scaled and all values of 0 were removed before the analysis. The variables CognateRating and ExperienceYears that were used to probe an interaction were z-scored to reduce multicollinearity and to make it more interpretable. The model was homoskedastic but the residuals displayed a slight skew (skewness: −0.35) and displayed a small tendency towards leptokurticity (kurtosis: 4.14), but both were not high enough to cause noteworthy problems regarding the model's validity. Outliers did not have an effect on the interpretation of the results and were left in the model.

The results of the model suggest a facilitation effect for the reading of cognates during translation that is however unaffected by the experience of the translator. Due to the activating cues of form and meaning, cognates likely lead to the activation and rapid selection of a cognate word in the TL that can be used for translation. Thus, the translator's attention can be turned towards the next ST word more quickly or turned towards the target text.

3.2.1 Reading Time and Formal Similarity of ST and TT

In the next model, we analyzed the effects of the literality of the actual translation choices rather than the general effects of the ST on whose basis these choices were made. We attempted to cover different aspects of the literal translation of the ST tokens and introduced these as variables in the linear regression model. The first is the typicality of the translation option that was chosen. As mentioned before, we can interpret a low translation self-information (ITra) as an indicator for a non-creative, typical translation strategy. The second variable was intended to cover word-order overlap. A low Cross Value would be indicative of a higher syntactic similarity of ST and TT. Lastly, we looked at the orthographic formal similarity of ST tokens with their translation (*FormalSimilarity*). A high formal similarity indicates a literal cognate-to-cognate based translation. The residual outliers affected our interpretation of the model because the interaction effect of FormalSimilarity and ExperienceYears shifted from non-significant (Est = 0.012 SE = 0.007 p = 0.107) to marginally significant (Est = 0.012 SE = 0.007 p = 0.08). We interpret the model without outliers (1.1% of the data were removed). The full model results are depicted below (see Table 5). The cleaned model showed no signs of heteroscedasticity and its residuals were distributed fairly normally (skewness: −0.3, kurtosis: 3.07)

The marginally statistically significant interaction between formal similarity of translation choices and experience (see Table 5 and Fig. 1) on reading time suggest that with more translation experience, literal cognate-to-cognate translations may slow reading speed. More experienced translators might try and gauge more carefully than less experienced peers if a cognate translation is adequate and engage in more frequent re-readings of a ST token. This supports the notion of increased self-monitoring during cognate translation (Oster 2017), at least for experienced translators. However, ST reading is but one manifestation of monitoring behavior and other measures are necessary to triangulate this finding.

3.2.2 Transl. Duration and ST Cognate Status

We repeated the two linear mixed models above but exchanged total reading time for translation duration (i.e. the time spent typing the translation of a ST token). We removed data points of zero duration and log scaled the variable Dur. The model showed some deviation from normality (kurtosis: 4.87) due to a number of outliers left and right to the mean (skewness: 0.55). No signs of heteroscedasticity of the residuals could be observed. A removal of the outliers did not have a notable effect on our variables of interest and only *Prob1* changed from non-significant (Est = -0.019 SE = 0.012 p = 0.112) to marginally significant (Est = -0.023 SE = 0.013 p = 0.067). Because our variables of interest were not affected, we present the results for the model with all data points (see Table 6).

Table 5 Model results from predicting Total Reading Time of the ST (TRTS) by FormalSimilarity

TrtS by FormalSimilarity					
(R^2m: 0.21/R^2c: 0.49)	Estimate	SE	*F*	p	VIF
(Intercept)	5.451	0.148	–	–	–
Concreteness	0.015	0.016	0.871	0.351	1.937
StudyKTHJ08	0.077	0.021	1316.516	<**0.001**	1.557
StudySG12	0.937	0.021			1.472
PoS*Adv*	−0.117	0.069	5.023	<**0.001**	1.471
PoS*Conj*	−0.176	0.079			1.56
PoS*Det*	−0.144	0.062			2.555
PoS*Modal*	−0.191	0.119			1.135
PoS*Noun*	−0.087	0.047			3.769
PoS*Other*	−0.259	0.078			1.741
PoS*Prep*	−0.276	0.057			2.686
PoS*Pro*	−0.308	0.081			1.379
PoS*Verb*	−0.021	0.048			2.663
Prob1	−0.01	0.005	4.661	**0.031**	1.287
LenS	0.11	0.006	368.989	<**0.001**	1.839
NameTRUE	0.184	0.075	6.006	**0.015**	1.158
ITra	0.018	0.007	7.066	<**0.01**	1.224
absCross	0.008	0.003	7.465	<**0.01**	1.115
zFormalSimilarity	0	0.011	0.002	0.963	1.328
zExperienceYears	−0.123	0.009	168.353	<**0.001**	1.074
STseg	−0.023	0.006	15.262	<**0.001**	1.026
zFormalSimilarity:zExperienceYears	0.012	0.007	3.067	**0.08**	1.002

When exposed to cognates, it seems that the effect on translation duration differs greatly from that of cognates on total reading time. Rather than facilitating, the presence of cognates slows the translators down—irrespective of experience.

3.2.3 Follow-Up Model to Transl. Duration: Number of Revisions

Unfortunately, the model on translation duration is without any inkling of why the duration increased. It may be a larger number of revision[5] attempts of cognate translations that is responsible or more hesitation, for example. To ascertain this, we used a follow-up model that predicted the number of revision attempts. In order to count such events, we made use of the TPR-DB's *Edit* variable and counted the number of opening square brackets that indicated revisions e.g. there

[5]Note that we tried to predict the TPR-DB variable InEff (translation inefficiency) at first. Unfortunately, the data was far from normally distributed even after a log-transformation of the variable.

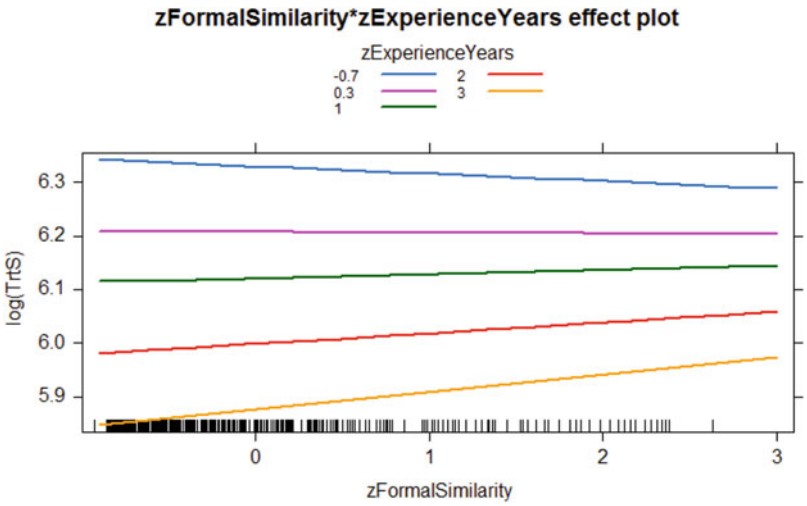


Fig. 1 Interaction effect between FormalSimilarity and Experience on the reading time of a ST token

are 3 revisions in the *Edit Gefährte,_Begleiter[Begleiter]Kamerad[Kamerad][,_]* (SG12_P14_T5). Because the number of revisions[6] represents a count variable, we used a non-parametric Poisson-regression model. With the help of the dispersion_glmer() function of the *blmeco*-package (Korner-Nievergelt et al. 2015) we checked if the model suffered from overdispersion that may affect the model's validity. The square root of the scale parameter was $0.75 < \mathbf{1.01} < 1.4$, so there was no indication of overdispersion. The model results are presented in Table 7.

The model indicates that words with a higher cognate rating tend to elicit to more revisions—at least for relatively inexperienced translators. More experienced translators seem to actually exhibit a tendency towards fewer revisions in their translation of cognates (Fig. 2). This may suggest that experienced translators choose and write their translations more carefully. While they take more time to come to a decision (partially due to the re-reading of ST cognate words if they decide to go for a cognate translation (Fig. 1)) they seem more likely to be content with their translation. It is possible to assume that less experienced translators revise their translations more often due to less effective initial monitoring, which leads to a premature cognate translation that is revised after the fact. Unfortunately, we did not have the tools to apply the FormalSimilarity measure to initial translations of cognates which would have helped to ensure that the revisions were indeed caused by cognate translations and were not systematically caused by a different mechanism (e.g. systematically fewer or more typos in response to translation of cognates respectively).

[6]Note that the concept of revision here is markedly different from that of Munit (Carl: this volume, "Micro Units and the First Translational Response Universal") which captures later returns to an earlier draft.

Table 6 Model results from predicting translation duration (Dur) by CognateRating

Duration by CognateRating					
(R^2m: 0.23/R^2c: 0.33)	Estimate	SE	*F*	p	VIF
(Intercept)	6.448	0.105	–	–	–
Concreteness	0.068	0.02	11.156	<**0.01**	1.949
StudyKTHJ08	−0.05	0.019	112.094	<**0.001**	1.415
StudySG12	0.217	0.019			1.322
PoS*Adv*	−0.096	0.084	12.813	<**0.001**	1.495
PoS*Conj*	−0.747	0.097			1.576
PoS*Det*	−0.326	0.075			2.564
PoS*Modal*	−0.222	0.141			1.146
PoS*Noun*	0.071	0.059			3.766
PoS*Other*	−0.273	0.092			1.821
PoS*Prep*	−0.472	0.07			2.77
PoS*Pro*	−0.367	0.098			1.395
PoS*Verb*	−0.173	0.06			2.66
Prob1	−0.01	0.006	2.526	0.112	1.3
LenS	0.082	0.007	136.709	<**0.001**	1.864
NameTRUE	−0.043	0.095	0.207	0.65	1.118
HTra	0.278	0.009	876.971	<**0.001**	1.064
zCognateRating	0.04	0.011	13.257	<**0.001**	1.117
zExperienceYears	−0.042	0.009	22.299	<**0.001**	1.085
STseg	−0.008	0.007	1.238	0.267	1.021
zCognateRating:zExperienceYears	−0.004	0.006	0.393	0.531	1.002

3.2.4 Translation Duration and Formal Similarity of ST and TT

Lastly, we looked again at the effect of literal translation strategies (ITra, Cross, and FormalSimilarity) on the log-scaled translation duration. We removed data points of zero duration and log scaled the variable Dur here again. The model displayed higher kurtosis (kurtosis: 5.01) due to outliers to the left and right to the mean (skewness: 0.38). However, outlier removal did not affect the interpretation of our variables of interest despite improvements of normality. The distribution of residuals was homoscedastic (Table 8).

We find a statistically significant interaction effect of experience and formal similarity again. It seems that the choice for a literal i.e. cognate-to-cognate translation decreases the overall translation duration, though experienced translators seem to benefit to a larger degree from choosing a more literal cognate-to-cognate translation than less experienced translators. It may be the case that experienced translators have thought through their decision more carefully (Fig. 3).

Table 7 Model results from predicting the Number of Revisions by CognateRating

Revisions predicted by CognateRating					
(R^2m: 0.063/R^2c: 0.14)	Estimate	SE	z	p	VIF
(Intercept)	−1.619	0.131	−12.315	<**0.001**	–
Concreteness	0.056	0.025	2.196	0.028	1.937
StudyKTHJ08	−0.374	0.031	−11.969	<**0.001**	1.387
StudySG12	0.031	0.029	1.061	0.289	1.295
PoSAdv	−0.059	0.106	−0.554	0.58	1.48
PoSConj	−0.379	0.136	−2.797	<**0.01**	1.415
PoSDet	−0.052	0.096	−0.542	0.588	2.421
PoSModal	−0.27	0.19	−1.419	0.156	1.123
PoSNoun	0.095	0.073	1.312	0.19	3.739
PoSOther	0.058	0.116	0.502	0.616	1.817
PoSPrep	−0.222	0.089	−2.488	**0.013**	2.565
PoSPro	−0.231	0.128	−1.806	**0.071**	1.348
PoSVerb	−0.149	0.074	−2.006	**0.045**	2.624
Prob1	0.002	0.008	0.207	0.836	1.284
NameTRUE	−0.104	0.122	−0.852	0.394	1.132
HTra	0.233	0.014	17.08	<**0.001**	1.094
LenS	0.038	0.009	4.329	<**0.001**	1.824
zCognateRating	0.029	0.016	1.737	**0.082**	1.162
zExperienceYears	−0.046	0.016	−2.878	<**0.01**	1.096
STseg	0	0.009	0.004	0.997	1.019
zCognateRating:zExperienceYears	−0.029	0.012	−2.39	**0.017**	1.008

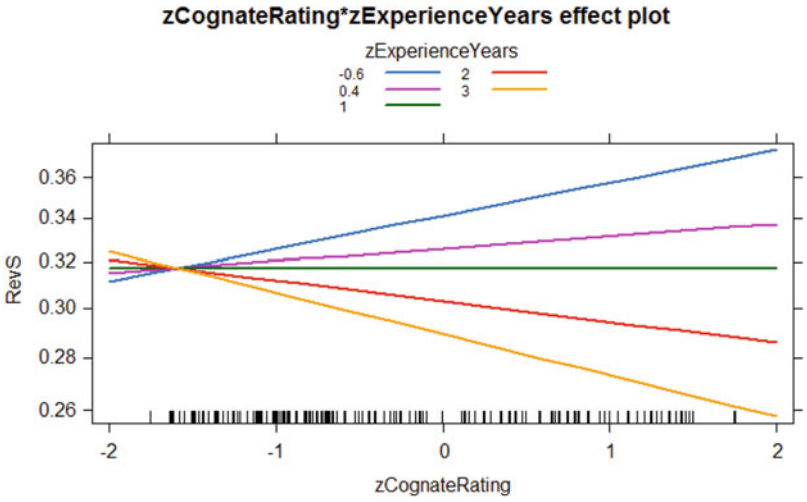


Fig. 2 Interaction effect between CognateRating and Experience on the number of revisions

Table 8 Model results from predicting translation duration (Dur) by FormalSimilarity

Dur predicted by FormalSimilarity					
(R^2m: 0.22/R^2c: 0.35)	β	SE	F	p	VIF
(Intercept)	6.311	0.1	–	–	–
Concreteness	0.057	0.02	7.794	<**0.01**	1.928
StudyKTHJ08	0.121	0.018	84.884	0	1.465
StudySG12	0.229	0.018			1.34
PoS*Adv*	−0.142	0.084	13.429	<**0.001**	1.491
PoS*Conj*	−0.757	0.097			1.566
PoS*Det*	−0.331	0.075			2.574
PoS*Modal*	−0.253	0.14			1.145
PoS*Noun*	0.099	0.059			3.753
PoS*Other*	−0.29	0.092			1.819
PoS*Prep*	−0.461	0.069			2.754
PoS*Pro*	−0.381	0.098			1.396
PoS*Verb*	−0.165	0.059			2.653
Prob1	−0.01	0.006	2.789	**0.095**	1.299
LenS	0.085	0.007	145.755	<**0.001**	1.87
NameTRUE	0.037	0.095	0.155	0.694	1.115
ITra	0.272	0.006	1926.318	<**0.001**	1.202
absCross	0.059	0.003	423.644	<**0.001**	1.121
zFormalSimilarity	−0.029	0.01	8.246	<**0.01**	1.229
zExperienceYears	−0.061	0.008	53.145	<**0.001**	1.088
STseg	−0.013	0.007	3.427	**0.066**	1.018
zFormalSimilarity:zExperienceYears	−0.025	0.006	15.592	<**0.001**	1.002

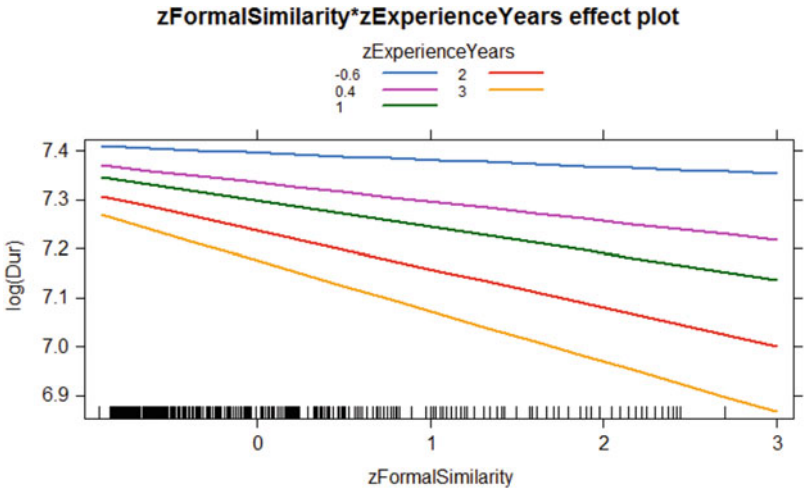


Fig. 3 Interaction effect between FormalSimilarity and Experience on the translation duration (Dur)

4 Discussion

This paper has proposed a relatively simple way of measuring *cognateness* and formal similarity of ST and TT tokens and has used a combination of product and process based analyses to test their effects on translation properties. By bridging the gap between product and process aspects of translation, we gained a more holistic impression of the effects of cognates and the literal translation thereof. While we found evidence for our product based hypotheses, the evidence was mixed for the process based hypothesis:

Hypothesis 1 a: Cognates reduce the number of translation alternatives Holding alternative explanatory factors constant, we could show that presence of a cognate is likely co-responsible for the reduction in the number of translation choices available to the translators. This suggests that the available translation choices that are associated with an ST item are reduced, the higher the cognate status of the ST item is. Another contribution of cognates to the reduction of HTra may rather occur indirectly. If a cognate translation is actively inhibited, it makes sense that the most strongly activated TL option that is not a cognate is selected over the others. This will most likely be a strongly entrenched target language option i.e. one that is frequently used in the TL (such as "wissenschaftlich" vs. "academic" in the qualitative example before (see Table 1)). Such an item likely exhibits a strong gravitational pull (Halverson 2003). This effectively restricts the number of translation choices to the cognate solution and another non-cognate solution that is frequently used in the TL. Together, these two should be the most dominant choices in a pool of other possible translation choices because they would receive the highest activation due to formal overlap of ST and TT words on the one hand, and entrenchment of TL words on the other. With our operationalization of cognate status by formal overlap, it is possible to control statistically (or experimentally) for effects of interference in translation experiments or corpus studies. This may allow us to quantify the influence of standardization (or normalization in terms of universals) on translation properties more accurately. But this was unfortunately not possible in the scope of this study.

A lower translation ambiguity would, according to Schaeffer and Carl's (2014a) definition of literality, indicate that cognates are associated with a higher literality of translation. We propose, however, that we reserve the notion of literality to the characteristics of a particular translation solution (in response to the ST), rather than an abstract average over a number of translation choices (such as HTra) (cf. Heilmann et al. 2019). We will pick this aspect up again in the conclusion section further below.

Hypothesis 1 b: Cognate-to-cognate translations contribute to the reduction of translation alternatives We could show that a literal cognate-to-cognate translation decreases the self-information of the translation solution and thereby translation ambiguity (as operationalized by word translation entropy).

Thus, cognate translations tend to frequently and systematically emerge as translations of their respective ST cognates. The reduction in HTra is thus not driven solely by the translator's attempt to avoid cognates and thereby settle for an entrenched, conventional, non-cognate target language option (gravitational pull) (Halverson 2010).

The higher the formal overlap of ST and TT cognate is, the more likely a cognate-to-cognate translation becomes. The effect may have been stronger if only initial translations of cognates had been considered as other studies have shown that first translation drafts contain more cognates than the final product (Oster 2017). The fact that we still found a negative correlation with word translation information may indicate that cognate translations are surprisingly persistent.

The choice of a cognate translation solution seemed to be largely independent from the translator's experience i.e. we could not find evidence that with more experience translators use more or less cognate translations. This stands in contrast to findings from Hansen-Schirra et al. (2017), who did find that translators with more experience were less inclined to use cognates. However, this may be explained by sample differences. Their sample consisted translation students and expertise was measured in terms of the number of semesters enrolled.

Hypothesis 2 a: Cognates facilitate translation Regarding the translation process, we could show that cognates exert a facilitation effect during the reading of ST words. This effect may be explained by the facilitated access due to the formal overlap of ST and TT cognates that strongly activates and selects a cognate for translation. Due to this quick access, the translator may turn to production relatively quickly.

The translation duration, on the other hand, seemed increased by the presence of cognates in the ST. This counteracted the facilitation effect in reading (or pre-translation) to some degree.

Interestingly, the increase of translation duration for the translation of cognates by more experienced translators did not seem to be caused by more revision. This means that the longer translation duration for cognate translation is mainly driven by longer pauses.

Hesitation may indicate that a cognate translation has been activated and selected for a first draft, but due to verbal self-monitoring the translation process is cut short before it is written down (Oster 2017, 35). This finding can thus be taken as an indication for a better developed self-monitoring skill in more experienced translators when it comes to the translation of cognates.

Hypothesis 2 b: Cognate-to-cognate translations of ST cognates contribute to this facilitation We found weak (i.e. marginally significant) evidence for a moderating effect of translation experience on cognate translation. Experienced translators showed increases in reading time when they chose translations with a high formal similarity. This could be taken as another indication of more self-monitoring during cognate translation (see Oster 2017, 35) in experienced translators.

By carefully choosing cognate translation, it seems that experienced translators engage in relatively unchallenged translation for items with high formal similarity, while less experienced translators do not benefit as much from cognate translation.

Thus, the small advantage regarding the reduction of reading time for inexperienced translators for cognate-to-cognate translations does not necessarily extend to the translation duration. Particularly, experienced translators seem to benefit from deciding for a cognate translation, making up for the additional increase in reading time (see Fig. 1). Future studies might use more sophisticated process measures such as the eye-key span (Schaeffer and Carl 2017) to help get a more detailed picture of the effects of cognates and literal translations thereof.

5 Conclusion and Outlook

The study of literal translation provides a great opportunity for process-oriented translation studies to bridge the gap to product based analyses of translation.

Because we studied the translation of complete texts, we can assume that the results obtained are more ecologically valid than more or less de-contextualized[7] psycholinguistic studies of cognates. With the help of our multivariate statistical analyses and the data from the TPR-DB we could still ensure control over many likely confounds. In general, we see the results as generally compatible with the Monitor Model (Tirkkonen-Condit 2005) in that self-monitoring during translation interrupts effortless production (effortless reading, effort due to monitoring during writing). However, having had no direct access to the initial first draft of a cognate translation leaves some results open for alternative interpretations. Future work may benefit not only from including the formal similarity of final translations with the ST token but also that of an intermediate version of translations before they are revised as well as from more elaborate process measures such as the eye-key span (Dragsted 2010).

Interestingly, the choice of a literal cognate-to-cognate translation revealed processing differences that were moderated by the translator's experience in reading, translation duration, and revision. Despite these processing differences, we could not find evidence for Toury's (2012) assumption that accomplished translators are more capable of freeing themselves from ST interference. It is possible that experience in years is an insufficient operationalization of accomplishment or expertise and that other factors may play a role here. Experience in years may be measuring habituation and not much more. Future studies should try to learn

[7]Due to the rigorous experimental control of confounding factors, words in psycholinguistic experiments tend to be presented in isolation or within single sentence without any notion of textual coherence or progression. This may result in observations that are experimental artifacts. The observed processing behavior may not surface (to the same extent) in more naturalistic text reception and production processes.

more about their participants and intricate measures such as Schaeffer et al.'s (2020) competence questionnaire, which may help with this in the future.

Following up on our description of surprisal and formal similarity, we deem it useful to adjust Schaeffer and Carl's (2014ba) argument and subsequently operationalization of literality. Thus, in order to measure the literality of a translation, we suggest to (a) substitute the translation ambiguity (HTra) by translation information, (b) keep the measure of word-order correspondence (*Cross*), and (c) add a measure of formal (orthographic) similarity (such as our FormalSimilarity) to the list of measures. The substitution of the translation ambiguity measure (HTra) by ITra keeps the general reasoning of Schaeffer and Carl intact but helps to separate source language phenomena from target text phenomena more clearly. With these modifications, literality would be clearly restricted to the properties of a particular translation solution (a translation strategy) rather than to the effect of the potential of translation strategies (HTra). From our perspective, HTra has proven to be one of the most valuable variables at our disposal to explain differences in translation performance. We only argue that it may be beneficial to separate effects at the system level from those associated with a specific instance of the said system (e.g. a translation strategy).

Translation self-information allows the researcher to characterize and compare the typicality of different translations of the same source text item. This makes it possible to grade different translation options by their predictability/literality, which would allow statements like "translator X translated ST Token Y more literal/predictable than Z." This is currently not possible with HTra because HTra is a model of overall choice and thus attributes the same value to translations of the same token.

As becomes evident from the process models above, the operationalizations of literality in form of self-information of the translation, word-order distortion (a form of syntactic formal dissimilarity), and orthographic formal similarity in concert explained independent fractions of the overall variation in the process data. The evidence from this study suggests that all literal translation strategies are associated with lower processing effort (but also increased monitoring for more experienced translators in case of cognate-to-cognate translations), which can be brought in line with the prediction of the monitor model (Tirkkonen-Condit 2005). The inclusion of these (and other[8] operationalizations of literality in process studies) allows us to see how specific translation choices or strategies affect the translation process. Due to them being rooted in the translation product, research on literality with the help of such measures offers a unique opportunity of closing the gap between product and process based research.

Acknowledgments This study was supported by the German Research Foundation (DFG) project TRICKLET, research grant no. NE1822/2-1

[8]see Carl: this volume, chapter "Information and Entropy Measures of Rendered Literal Translation", for other operationalizations of aspects of literality.

References

Baayen H, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Memory Lang 59(4):390–412

Baker M (1996) Corpus-based translation studies: the challenges that lie ahead. In: Somers H (ed) Benjamins translation library, vol. 18. John Benjamins, Amsterdam, pp 175–186

Barton K (2009) MuMIn: Multi-Model Inference. R Package Version 1. 0. 0. http://xn--rforge-wg0c.xn--rproject-sn3d.org/projects/mumin/

Boada R, Sánchez-Casas R, Gavilán JM, García-Albea JE, Tokowicz N (2013) Effect of multiple translations and cognate status on translation recognition performance of balanced bilinguals. Biling Lang Cogn 16(1):183–197

Brysbaert M, Diependaele K (2013) Dealing with zero word frequencies: a review of the existing rules of thumb and a suggestion for an evidence-based choice. Behav Res Methods 45(2):422–430

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Schaeffer M, Bangalore S (eds) New directions in empirical translation process research. Springer, Cham, pp 13–54

Christoffels IK, De Groot AMB (2005) Simultaneous interpreting: a cognitive perspective. In: Kroll JF, De Groot AMB (eds) Handbook of bilingualism: psycholinguistic approaches. Oxford University Press, pp 454–480

De Groot AMB (1992) Determinants of word translation. J Exp Psychol Learn Memory Cogn 18(5):1001–1018

De Groot AMB (2011) Language and cognition in bilinguals and multilinguals: an introduction. Psychology Press, New York

Dell GS (1986) A spreading-activation theory of retrieval in sentence production. Psychol Rev 93(3):283–321

Demberg V, Keller F (2008) Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 109(2):193–210

Diependaele K, Lemhöfer K, Brysbaert M (2013) The word frequency effect in first-and second-language word recognition: a lexical entrenchment account. Q J Exp Psychol 66(5):843–863

Dijkstra T, Van Heuven WJB (1998) The BIA model and bilingual word recognition. In: Localist connectionist approaches to human cognition, pp 189–225

Dragsted B (2010) Coordination of reading and writing processes in translation: an eye on uncharted territory. In: Shreve GM, Angelone E (eds) American translators association scholarly monograph series. John Benjamins, Amsterdam, pp 41–62

Dragsted B (2012) Indicators of difficulty in translation – correlating product and process data. Across Lang Cult 13(1):81–98

Evert S, Neumann S (2017) The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In: De Sutter G, Lefer M-A, Delaere I (eds) Empirical translation studies. De Gruyter, Berlin/Boston, pp 47–80

Frank A (2014) Diagnosing Collinearity in Mixed Models from Lme4, Vif.mer Function. https://raw.githubusercontent.com/aufrank/R-hacks/master/mer-utils.R

Frank S, Thompson R (2012) Early effects of word surprisal on pupil size during reading. In: Proceedings of the annual meeting of the cognitive science society. vol 34

Halverson S (2003) The cognitive basis of translation universals. Target Int J Transl Stud 15(2):197–241

Halverson S (2010) Cognitive translation studies: developments in theory and method. In: Shreve GM, Angelone E (eds) American translators association scholarly monograph series, vol XV. John Benjamins, Amsterdam, pp 349–369

Halverson S (2019) 'Default' translation: a construct for cognitive translation and interpreting studies. Transl Cogn Behav 2(2):187–210

Hansen-Schirra S (2011) Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language. In: Kranich S, Becher V, Höder S, House J (eds) Multilingual discourse production: diachronic and synchronic perspectives. John Benjamins, Amsterdam, pp 133–162

Hansen-Schirra S, Nitzke J, Oster K (2017) Predicting cognate translation. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 3–22

Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, Robinson BS, Hodgson DJ, Inger R (2018) A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ 6:1–32

Heilmann A (2021) Profiling effects of syntactic complexity in translation: a multi-method approach. PhD thesis

Heilmann A, Serbina T, Couto-Vale D, Neumann S (2019) Shorter than a text, longer than a sentence: source text length for ecologically valid translation experiments. Target 1:98–125

Ivir V (1981) Formal correspondence vs. translation equivalence revisited. Poet Today 2(4):51–59

Kim H-Y (2013) Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative Dent Endod 38(1):52–54

Komsta L, Novomestky F (2015) Moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. https://CRAN.xn--rproject-sn3d.org/package=moments

Korner-Nievergelt F, Roth T, von Felten S, Guelat J, Almasi B, Korner-Nievergelt P (2015) Bayesian data analysis in ecology using linear models with R, BUGS and stan. Elsevier

Kroll JF, Van Hell JG, Tokowicz N, Green DW (2010) The revised hierarchical model: a critical review and assessment. Biling Lang Cogn 13(3):373–381

Kuznetsova A, Brockhoff PB, Christensen RHB (2015) Tests in Linear Mixed Effects Models. https://cran.r-project.org/web/packages/lmerTest/index.htmln0qTH7jYKtICMhw7qBx5m. Visited on 23 Nov 2019

Laxén J, Lavaur J-M (2010) The role of Semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. Biling Lang Cogn 13(2):157–183

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 10:707–710

Oster K (2017) The influence of self-monitoring on the translation of cognates. Empir Model Transl Interpret 7:23

Paradis M (1997) The cognitive neuropsychology of bilingualism. In: De Groot AMB, Kroll JF (eds) Tutor biling psycholinguistic perspect. Erlbaum, New Jersey, pp 331–355

Paradis M (2004) A neurolinguistic theory of bilingualism, vol 18. John Benjamins, Amsterdam

Prior A, MacWhinney B, Kroll JF (2007) Translation norms for English and Spanish: the role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. Behav Res Methods 39(4):1029–1038

R Core Team (2017) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.xn--rproject-sn3d.org/

Schaeffer M, Carl M (2014a) Measuring the cognitive effort of literal translation processes. In: Proceedings of the EACL 2014 workshop on humans and computer–assisted translation, pp 29–37

Schaeffer M, Carl M (2014b) Measuring the cognitive effort of literal translation processes. In: Proceedings of the EACL 2014 workshop on humans and computer-assisted translation, pp 29–37

Schaeffer M, Carl M (2017) Language processing and translation. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 117–154

Schaeffer M, Huepe D, Hansen-Schirra S, Hofmann S, Muñoz E, Kogan B, Herrera E, Ibáñez A, Garcia AM (2020) The translation and interpreting competence questionnaire: an online tool for research on translators and interpreters. Perspectives 28(1):90–108

Schmid H (1995) Treetagger\textbar a language independent part–of–speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart 43:28

Teich E (2003) Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. De Gruyter, Berlin

Tirkkonen-Condit S (2005) The monitor model revisited: evidence from process research. Meta 50(2):405–414

Tokowicz N, Kroll JF (2007) Number of meanings and concreteness: consequences of ambiguity within and across languages. Lang Cogn Process 22(5):727–779

Tokowicz N, Kroll JF, De Groot AMB, Van Hell JG (2002) Number-of-translation norms for Dutch—English translation pairs: a new tool for examining language production. Behav Res Methods Instrum Comput 34(3):435–451

Toury G (2012) Descriptive translation studies – and beyond. John Benjamins, Amsterdam

Voga M, Grainger J (2007) Cognate status and cross-script translation priming. Memory Cogn 35(5):938–952

Part III
Translation Segmentation and Translation Difficulty

Micro Units and the First Translational Response Universal

Michael Carl

Abstract In this chapter, we investigate the *first translational response universal* which posits that translators mentally segment texts into units that can be kept in memory and processed by their available cognitive resources. We investigate the observable traces of the first translational response, i.e., keystrokes, gaze patterns, and pauses of the translator's first translation draft, which reveal characteristics that cohere, among other things, with the language pair, the properties of the text to be translated, and the "longest stretch of translation that a translator can deal with at once" (Malmkjær, Translation universals, 91, 2011). Our findings suggest that the typing pause that precedes the initial translational response depends—among other things—on the information content of the source and the non-literality of the translation, i.e., the syntactic and semantic cross-linguistic similarity. We relate our findings to the *literal translation hypothesis* showing that the first translational response is likely to be more literal than successive revised versions. If is also more literal when the preceding pause is short. We conclude that translators tend to proceed in the most compositional manner and produce the shortest possible stretch of translation that the context allows for.

1 Introduction

"No translator is able to work at once with an entire text" (Malmkjær 2011: 88). The segmentation of a text into smaller translation units is thus essential, and the investigation of the segment and segmentation properties is likely to provide valuable insights into the translation process. Numerous attempts have been made to determine basic translation units with different goals. Many researchers (e.g., O'Brien 2006, Lacruz and Shreve 2014, Kumpulainen 2015) have argued that pauses in the flow of keystrokes are indicators of cognitive effort, where more

M. Carl (✉)
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_9

and longer pauses indicate extended cognitive effort. This chapter investigates the *first translational response* and with this the segmentation during the initial translation of a stretch of source text (ST). Malmkjær suggests the first translational response to be a translation universal, as "something has to be the first thing that comes to your mind when you are faced with a linguistic item to translate," and while this phenomenon occurs in translation, "it is not present in unilingual language events," thus justifying that it is a translation universal. According to Malmkjær (2011:88–89), "the pairings of ST and TT [target text] that emerge as first translational responses might tell us something about the interlingual relationships, and the linguistic-conceptual relationships that exist in the translating bilingual's mind." Data-Bukowska (2019:185) suggests "that it may be rooted in the cognitive mechanism of priming" but acknowledges that only very little research into potential universals of translation has been carried out in the past years, including the first translational response universal.

This chapter investigates variation in the first translational response across different translators and different languages to pinpoint commonalities that underlie initial segmentation and translational responses. We investigate translation process data of approximately 240 translation sessions from English into six different languages and relate the text production pause that preceded the first translational response to properties of the final translation. We discuss several notions of translation units and describe how we automatically determine micro units in the translation data that relate to ST segments.

The notion "first translational response" is closely related to that of *default translation* and the *literal translation hypothesis* (see Carl this volume, Chap. 5). Default translations have been of interest in translation studies for some time (Carl and Dragsted 2012, Halverson 2015, 2019). Carl and Dragsted (2012) make a distinction between "default translations" and "challenged translation," where default translations are first translational responses that are produced quickly with no apparent problems, while "challenged translations" require more time to be formulated and are presumably more problematic. We hypothesize that default translations are more literal (i.e., monotone, compositional, and entrenched; see Carl this volume, Chap. 5) than challenged translations. As Halverson (2015, 313) puts it, "translators first opt for formal correspondents and only deviate from them when forced to do so." For Chesterman (2011, 26), the literal translation hypothesis "makes a claim about the translation process ... in which translators tend to proceed from more literal versions to less literal ones," while Tirkkonen-Condit (2004, 183) observes a "tendency of the translating process to proceed literally to a certain extent," where the literal translation is typically a translator's first choice (Tirkkonen-Condit 2005). For Malmkjær (2011, 89), a first translational response "involves simultaneous suppression and activation of the right features of the linguistic systems at the right time in the right proportions to each other." More entangled activation and suppression of those features leads presumably to longer pauses and more challenged and less literal translations. It is, according to Malmkjær, determined by the amount of paired text a translator can hold in short-term memory. While there is, thus, a body of considerations about the

importance and implications of first translational choices, the *first translational response universal* has—to the best of our knowledge—not been investigated in a multilingual rigorous empirical setting.

In order to investigate properties of the first translational choices, Halverson (2019, 191) suggests investigating the "relationships between the default phase of production and aggregate properties of translated text," as this relation might provide insights into cognitive processes that reveal specific knowledge and experience of translators. Following the notion of Carl and Dragsted (2012), Halverson suggests to "identify stretches of non-problematic translation performance" (2019: 198) by classifying first translational responses into unchallenged default translations and (presumably) challenged, non-default translations, based on the typing pause and gazing behavior (i.e., the *translation act*) that precede the typing burst (i.e., the *translation event*). However, it is unclear what exactly such a threshold for the typing pause should be. Halverson discusses several approaches related to finding a suitable pause threshold as a "means of identifying default translation in the [recorded translation process] data" (Halverson 2019: 198). However, numerous suggestions have been made regarding how to define a pause threshold—static or dynamic, by taking into account averaged typing behavior, etc.—to segment the flow of recorded keystrokes into production units (see, e.g., Kumpulainen, 2015, for an overview) with the aim of discriminating between different processes of translation production. Too manifold are the reasons for and manifestations of translation difficulties, so that there is little hope of finding one such pause threshold—whether dynamic or static—that fits all purposes of clearly separating instances of problematic from non-problematic translation production.

Instead of engaging in a troublesome search for more sophisticated pause thresholds and a binary (re-) definition of default and challenged translation, we suggest looking at the issue the other way around, from the product to the process. In line with Chesterman's (2015) notion of "reverse-engeneered processes,"[1] we suggest assessing "aggregate properties of translated text" (e.g., translational variation and cross-linguistic reordering as detected in the translation product) and following the revision traces (*Munits*, see Sect. 2) back from the final translation product to the first transational response. This trace of *Munits* illustrates how the translation of ST segments has emerged in time. Instead of defining behavioral patterns in the process data (e.g., keystroke pauses) and searching for correlations in the product, we quantify properties of the translation product and investigate the traces of recorded behavioral data that have produced these translations. Given completeness of activity data, we circumvent Chesterman's (2015, 15) doubt whether a "reconstructed process necessarily represent the *actual process* which

[1]Chesterman distinguishes between "models of actual processes" and "models of reversed engineered processes," one looking forward, the other backward into the translation process. However, given complete information, the two models are in fact equivalent and can be mapped into each other without loss of information via Bayes theorem or, equivalently, the noisy channel model. With respect to keylogging and eye-tracking, we have (almost) complete information how a translation was produced and can thus look backward into the production process.

terminated in the a given translation" (original italics): given the complete trace of keystrokes and gaze data, those processes are reversible.

Jakobsen (2011) conceptualizes the translation production process as a cycle of cognitive and motor processes, reflected in eye movements and keystrokes, respectively. He distinguishes between several instances of *translation acts* (cf. Chesterman 2015) which are characterized by gaze movements on the source or target texts and *translation events* in which the translation is typed. Alves and Vale (2009, 2011) operationalize this cycle by means of translation units (TUs), as well as micro and macro translation units. A TU, according to them, "begins with a reading phase [i.e. *a translation act*] ... and evolves in a continuous production [i.e. *a translation event*] until it is interrupted by a pause [i.e. another translation act]" (Alves and Vale 2009: 257). The outcome of a TU is thus product data in the form of a piece of translated text. Alves and Vale further introduce the notion of *micro TU* which represents the translation events of TUs: it is "the flow of continuous TT production" (Alves and Vale 2009) that refers to a particular ST segment. In addition, they establish the notion of the *macro TU* "as a collection of micro TUs that comprises all the interim text productions that follow the translator's focus on the same ST segment" (Alves and Vale 2009). A macro TU thus comprises all edit operations related to the translation of an ST segment, including the first translational response and all successive revisions of it.

However, there seems to be an incompatibility in the definition of TUs—which is supposed to be preceded by a typing pause of a predefined minimum length—and the *macro TUs*, which subsume all micro TUs that relate to the same ST segment. As we will show, it is likely that different TUs refer only partially to intersecting ST segments (one TU may refer to several ST segments), and it may well be the case that different micro TUs which refer to the same ST segment are not preceded by the minimum TU-defining preceding pause. To overcome this incompatibility, we define a micro unit (*Munit*) as consisting of parts of TUs that relate to a well-defined ST segment. A collection of *Munits* that relates to the same ST segment reliably traces the translation production, from the first contributing keystrokes, and includes all possible successive revisions of it, but this implies that *Munits* can be preceded by pauses well below the TU minimum pause threshold.

Reversing the order of *Munits* for a given ST segment leads back to *Munit1*. *Munit1* is the first micro unit that relates to the translation production of an ST segment and thus represents a first translational response. It starts with the first keystroke that contributes to the translation of the ST segment and lasts until the first draft of the translation for that ST segment is finished. We refer to the duration of *Munit1* as *Dur1* and the typing pause that preceded *Munit1* as *Pause1*. Depending on the duration of *Pause1*, the first translational response (i.e., *Pause1* + *Munit1*) could be clustered into default or challenged translations. However, instead of searching for a (arbitrary) pause threshold, we take *Pause1* and *Dur1* as continuous variables. The aim of this chapter is, then, to investigate (1) parameters that determine the duration of *Pause1*. We also investigate (2) the relation between properties of the final translation and the revision process, i.e., the number of *Munits* that has contributed to the translation of the ST segment. We find that numerous variables

play an important role in the first translational response, i.e., the duration of *Pause1* and *Munit1* as well as for the revision process.

In order to (partially) automate the annotation and evaluation process, we introduce the notion of alignment groups (AGs) which represent possible locations of the "translator's focus on the same ST segment" and which also represent segments of translation equivalence of the ST and TT. We define Production Units (PUs) as continuous sequences of translation production (keystrokes), and the intersection between AGs and PUs will then allow us to automatically compute *Munits*.

Section 2 explains in detail how *Munits* are generated as an intersection of PUs and AGs. From the discussion of several examples, we develop hypotheses for factors that determine the duration of *Pause1* and the relation between *Munit1* and properties of the final translation product. Section 3 investigates the impact of *Dur1* and several product properties on the duration of *Pause1*. First, we investigate the relation between the distribution of visual attention on the source and the target text during *Pause1* and the duration of *Pause1*. Then, we assess the effect of *Dur1*, as well as properties of the final translation product on *Pause1*. Section 4 looks at revision patterns, i.e., the number of *Munits* per AG. Our data confirms the literal translation hypothesis, which posits that more revision of the first translational response leads to less literal translations. Thus, translations that are more often revised are less monotone, less compositional, and less entrenched than those that are less often revised. This effect is most pronounced for verbs and less so for articles and adjectives. Section 5 discusses the results in the light of bilingualism and priming studies.[2]

2 Micro Units and Revision Behavior

Alves and Vale introduce the micro translation units as a "flow of continuous TT production" (2011, 107) that relates to a given ST segment. For them, a micro TU ends if a keystroke pause occurs that is longer than "the standard threshold of five/six seconds." Micro TUs aggregate into macro TUs. A macro TU is a collection of

[2]The empirical analysis is based on translation process data from six language pairs, English-to-Danish, German, Spanish, Hindi, Japanese, and Chinese, extracted from the *multiLing* dataset. It consists of 34,106 words. The *multiLing* corpus was introduced in Carl (this volume, Chap. 5, Appendix). We only take from-scratch translations into six languages (Danish, German, Spanish, Hindi, Japanese, and Chinese) to investigate properties of default translations.

TL	da	de	es	hi	ja	zh	Total
#TL words	8249	4732	6808	3008	5882	5427	34,106

micro TUs "that comprises all the interim text productions that correspond to the translator's focus on the same ST segment." Alves and Vale suggest a taxonomy of macro TUs, depending on the revision behavior of translators and according to whether micro TUs are processed solely during the drafting or revision phase. For instance, consider the following three edit strings in the micro TUs in Example 1, two of which were produced in the drafting phase, while the last one (following the symbol "~") occurred in the revision phase. As a translation of *meter of blood-sugar-level*, initially *medidor de índice* was typed. Then, later in the drafting phase, *índice* is changed into *glicemis* and later in the revision phase to *glicemia*.

Example 1

English source segment:

Meter of blood-sugar-level.

Portuguese translation: Medidor de índice|Medidor de glicemis ~ Medidor de glycemia.

Alves and Vale investigate the translation behavior of 12 professional translators and provide a classification of translators according to their revision patterns in the drafting and the revision phase. They notice that revision patterns are substantially different in the different phases and consider online revision "to be where translation takes place par excellence" (2011: 120).

Alves and Vale develop and describe a browser-based tool, LITTERAE,[3] with which micro and macro TUs can be manually annotated and searched. While LITTERAE constitutes important pioneering work that allows for empirical investigation of translation processes, some additional steps need to be introduced if we want to automate the labor-intensive annotation, linking and evaluation of micro Tus, and their aggregation into macro TUs. We will describe this automatization process in the next subsection and then discuss some examples.

2.1 Alignment Groups, Production Unit, and Micro Units

In order to automatize the generation of *Munits*, we make use of production units (PUs) and illustrate this on a made-up example. Assume a first draft translation "AFG" has been produced in one initial text production unit (PU-*a*). Assume further that a subsequent revision takes place (PU-*b*) which substitutes "F" with "D." This substitution results in an intermediate solution "ADG." Assume another revision PU-*c* that replaces "G" with "E," which provides an interim solution "ADE." A successive PU-*d* inserts "BC," which leads to the final translation "ABCDE." According to the definition of macro TUs, to aggregate several micro TUs with the same ST focus into one macro TU, we need to know for each of

[3] http://letra.letras.ufmg.br/litterae/

the translation actions whether or not the translator's focus was on the same ST segment. But, obviously, by only looking at TT edit operations, we cannot know which modification relates to which ST segment and whether different modifications in close proximity should be clustered into the same macro TU.

For instance, with the four productions (a, b, c, d) described above, we have four successive versions of the TT, which are reproduced in Table 1. But the mere fact that "ABC" was written in one initial production burst does not necessarily mean that substituting "B" by "D" and substituting "C" by "E" are operations that refer to the same ST segment. It might well be the case that one production contains up to ten or more words (Jakobsen 2005), which is likely to include several ST segments that we do not want to subsume into one macro TU, but rather consider them individually. While the LITTERAE tool may provide possibilities for human annotators to choose labels and ST-TT relations on a case-to-case basis, we need to find a principled way to assign edit operations to ST segments automatically and thus to generate *Munits* for large amounts of test.

Within the CRITT TPR-DB,[4] this is achieved through an intersection of AGs and production units (PUs). PUs—represented in letters (*a, b, c, d*) in Table 1—are very similar to *Munits*, with the only (but important) difference that we do not know (yet) to which ST segment(s) the edit operations relate. In addition, we fragment the translation product into a set of AGs—indexed as numbers (1–5; see below) on the right side in Table 1—which represent translation equivalents, and the ST equivalent represents a possible focus of the translator's mind. In order to intersect PUs with AGs, we start from the final translation product and produce the *Munits* through successive application of edit operations as specified by PUs in the reverse order. Thus, starting from the final translation "ABCDE" and reversing the operation of the last PU-*d*, i.e., deleting "BC," provide "ADE." Successive substitution of "E" → "G" results in "ADE," and so on, until the beginning of the list of PUs is reached. If we know during this reverse process which ST token(s) align with which TT token(s), we can map each edit operation onto an AG. A PU (or a part of it) becomes a *Munit* as it intersects with and is assigned to a specific AG. AGs, thus,

Table 1 Intersecting production units (PUs) and alignment groups (AGs) to compute micro units (*Munits*)

	Production unit		Different segmentations (I … V) with their AGs (1 … 5)				
TT string	PU	Operation	I	II	III	IV	V
AFG	a	Insertion: AFG	1	1, 4, 5	1,2	1,2	1,2
ADG	b	Substitution: F → D	1	4	1	1	–
ADE	c	Substitution: G → E	1	5	1	2	2
ABCDE	d	Insertion: BC	1	2,3	2	3	1

[4]The CRITT Translation Process Research Database is a publicly available repository of translation production sessions that is described in more detail in Carl (this volume, Chap. 5). The keystroke-mapping algorithm presented here is also slightly differently described in Carl (2012).

subsume one or more *Munits* and replace the notion of macro TU. We illustrate this process with several examples.

Assume "ABCDE" has been produced as a translation of the ST segment "abcde." Let us further assume we have some knowledge that tells us which groups of ST tokens align with which groups of tokens in the final TT. There are many different ways in which a translation can be segmented into AGs, each of which results in a different distribution of the edit operations and different *Munits*. The following list enumerates five possible bilingual segmentations of the hypothetical ST segment and the TT segment into a different number of AGs. Each AG consists of one or more ST and TT tokens within parentheses which are co-indexed with identical numbers:

1. 1:(abcde) ←→ 1:(ABCDE)
   The translation consists of one AG which is not decomposed into smaller translation equivalents (e.g., representing a metaphor that cannot be compositionally translated).
2. 1: (a) 2: (b) 3: (c) 4: (d) 5: (e) ←→ 1:(A) 2:(B) 3:(C) 4:(D) 5:(E)
   Each of the five ST tokens has a correspondence in the TT and the order of tokens in ST, and their TT translations is identical, i.e., a most monotone and compositional translation.
3. 1:(ab) 2:(cde) ←→ 2:(ABC) 1:(DE)
   The ST and the TT are made up of two AGs, one consisting of two words, the other of three words, and the order is reversed on the target side.
4. 1:(ab) 2:(cd) 3: (e) ←→ 1:(A) 3:(BC) 1:(D) 2:(E)
   The translation consists of three AGs of which one (ab) is mapped to a discontinuous target segment "A..D."
5. 1:(ab) c 2:(de) ←→ 1:(ABC) D 2:(E)
   The translation consists of two AGs, but there is an omission in a source item "c" which has not been translated, and there is an insertion of item "D" in the translation which does not have a correspondence in the ST.

Table 1 shows how the four PUs are assigned to different AGs, depending on how the translation is segmented. For instance, in segmentation I, all five ST and TT tokens are grouped into AG-1. Accordingly, all four PUs (*a, b, c, d*) are linked to I:AG-1. They all become *Munits* of segmentation I:AG-1, as all modifications relate to the same ST segment I:AG-1. In contrast, segmentation (II) is the most compositional alignment which consists of five AGs. Also here, each of the five AGs is linked to those parts of the PUs that modify the associated translation. For instance, PU-*c* substitutes "G" with "E," and because "E" is the translation of a source segment II:AG-5, PU-*c* becomes *Munit-c* in II:AG-5. Notice that PU-*a* also contributes to the production of II:AG-5, II:AG-1, and II:AG-4. While one PU may contribute to several AGs, a *Munit*, by definition, can only contribute to one AG—for an AG aggregates *Munits* with the *same ST focus* and AGs do not overlap. We may thus need to distribute the operations of PU-*a* into several *Munits*, so that each *Munit* only consists of operations for the AG of which it is part. For instance, in the context of segmentation II, PU-*a*:{AFG} will be split into three, *Munit-a-1*:{A},

Table 2 AGs and *Munits*

Segmentation	I	II					III		IV			V	
AG	1	1	2	3	4	5	1	2	1	2	3	1	2
ST segment	abcde	a	b	c	d	e	ab	cde	ab	cd	e	ab	de
TT segment	ABCDE	A	B	C	D	E	DE	ABC	A..D	E	BC	ABC	E
Munit1	*a*	*a-1*	*d-2*	*d-3*	*a-4*	*a-5*	*a-1*	*a-2*	*a-1*	*a-1*	*d*	*a-1*	*a-2*
Munit2	*b*				*b*	*c*	*b*	*d*	*b*	*c*		*d*	*c*
Munit3	*c*						*c*						
Munit4	*d*												

Munit-a-4:{F}, and *Munit-a-5*:{G} so as to intersect with each of the translations in segmentation II AG-1, AG-4, and AG-5, respectively.

Table 2 shows the sequences of *Munits* aggregated under different segmentations and AGs as discussed above. It shows that a different segmentation of the translation may lead to different *Munits* and different sequences of *Munits*. Every AG has at least one *Munit* (*Munit1*) which constitutes its first translational response. Some AGs include several *Munits*. For instance, segmentation II:AG-1 contains all four *Munits*, while many AGs for most of the other segmentations only contain *Munit1*. As a tendency it thus appears that more compositional segmentations (i.e., shorter AGs) will subsume fewer revisions (fewer *Munits*). As we will demonstrate below, this is also the case when one takes into account the number of source tokens involved. Note that in segmentation (V), the ST token "c" does not occur in any of its AGs, as it is an omission. Also, PU-*b* is not listed as a *Munit*, as the TT token "D" was not aligned to any source token; it is an insertion, and thus the edit operation that relates to PU-*b* does not seem to imply an ST focus. Note also that the process described here has limitations when deletions (or substitutions) cover several words and more than the translation in one AG. The entire deleted string will then be attributed to the leftmost TT word.

2.2 Examples of Verbal and Nominal Translation

The previous section established a relation between three concepts:

- **Alignment group** (AG): interlingual correspondence of source words and target words which constitute a translation equivalence within the final translation product.
- **Production unit** (PU): sequence of coherent typing. Within the TPR-DB, a PU is a flow of continuous TT production, with delays of inter-keystroke pauses of less than 1 second.
- **Micro unit** (*Munit*): sequence of coherent typing activities that relates to the words in an AG. Unlike PUs, *Munits* are not constrained by minimum preceding or maximum internal pauses.

We discuss two examples to elaborate on the relation between these notions. Example 2 illustrates a rather complex verbal AG-6 that stretches over four ST words and four TT words and involves several PUs.

Example 2

English source, segmented into AGs:

1:(the government's) 2:(insistence) 3:(that those in the) 4:(public) 5:(sectors).
6:(have to receive below-inflation) 7:(salary) 8:(increase).

Spanish translation, segmented into AGs:

2:(la insistencia) 1:(del gobierno) 3:(en dar) 6:(una inflación menor en) 8:(la subida).
7:(de sueldos) 5:(al sector) 4:(publico).

Example 2 is an English-Spanish translation,[5] with eight AGs, indicated in parentheses which are identically co-indexed. We look at the production of AG-6, which involved four PUs, three in the drafting phase and one in the revision phase.

$$\text{PU} : a : \{\text{dar una inflación mayor}\} \mid b : \{\text{en}\} \mid c : \left\{\left[\text{mayor}\right]\text{menos}\right\} \sim d : \{[\text{s}]\,\text{r}\}\,.$$

The translator starts out with PU-*a* by typing "dar una inflación mayor" which lasts 4985 ms and which is preceded by a typing pause of 7328 ms. This is followed by another typing pause of 5062 ms after which PU-*b*: "en" (with a blank space) was typed within 114 ms. After finishing the translation of the entire sentence, the translator returns to this segment, produces PU-*c* by deleting "mayor" (deletions are annotated in square brackets in PUs), and replaces it with "menos." This took 1422 ms. Later, during the revision phase, the translator replaces—within PU-*d*—the "s" in "menos" with "r," which only took 93 ms and results in the final translation "dar una inflación menor en," as show in Example 2.

However, the produced translations observed within these PUs do not necessarily coincide with the segmentation of AG-6. In particular "dar" in PU-*a* is part of AG-3, and since no other content was produced between PU-*a* and PU-*b*, both contribute to the same *Munit1-a-6,b-6*. The duration of this micro unit is 9938 ms, as it also includes the typing pause of 4985 ms between PU-*a* and PU-*b*, while *Munit2-c* and *Munit3-d* are identical to PU-*c* and PU-*d*, respectively. There are thus three Munits within AG-6:

$$\text{Munit} : a - 6, b : (\text{una inflación mayor en}) \mid c : \left(\left[\text{mayor}\right]\text{menos}\right) \sim d : ([\text{s}]\,\text{r})\,.$$

These three Munits have durations of 9938, 1422, and 93 ms, respectively, leading to a total production duration of 11,453 ms for AG-6. As this discussion

[5]The translation is from the study BML12; it is a part of a segment in session P26_T2.

shows, the number of *Munits* and PUs that are associated with an AG can be different. As each *Munit* represents a revision, this suggests that translations of more complex AGs are more often revised. It also suggests that they are produced in a less fluent manner, involving more PUs, and it shows that all words within one multi-word AG, such as AG-6: (*have to receive below-inflation*) are linked to the same *Munits*.

Example 3

English source:

> Increasing mobility and technological advances resulted **in the increasing exposure of** people to cultures and societies.

Spanish translation:

> Una_mayor movilidad y los_nuevos avances tecnológicos supusieron **un augmento en la exposición de** los pueblos a culturas y sociedades.

In order to assess gaze activities, we discuss another segment in an English-to-Spanish translation, highlighted in bold in Example 3. Example 3 shows how a nominal phrase is translated within one long PU which is distributed over several compositional AGs that represent word-for-word translations.

Figure 1 shows the production pattern of this segment, with two PUs. The striped blocs in Fig. 1 represent the stretch of time in which PU-*a*: {un augmento en la} and PU-*b*: {exposición de} were produced. The figure shows gaze activities on the source (in blue asterisks) and on the target text (in green diamonds), before, during, and after the two production bursts. PU-*a* starts at timestamp 13,660 and is preceded by a pause of 1875 ms in which the English chunk "increasing exposure of people" is scanned. The typing of the translation "un ayum[yum]ento en la" lasts 3250 ms during which the translator mostly monitors the production of the translation (green diamonds). Toward the end of PU-*a*, the gaze returns to hover over the source, and gaze activity is recorded on "exposure of people." Then the eyes move back to the target, and the PU-*b* "exposición de" is produced while monitoring the typing activity. PU-*b* starts at time stamp 141,031 and is separated from PU-*a* by typing

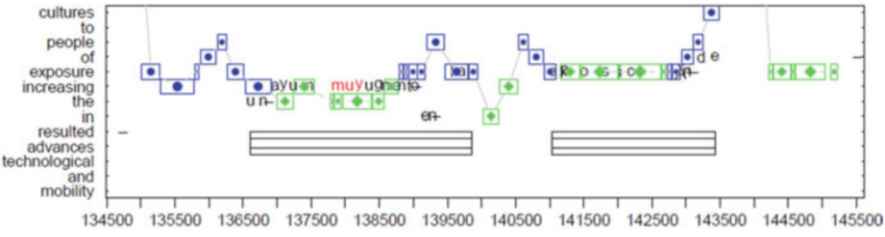


Fig. 1 Translation progression graph showing two production units PU-*a* and PU-*b* in relation to other process data

AG	2	3	1	4	5
Pause1 **(ms)**	1875	93	125	234	109
Dur1 **(ms)**	266	2079	156	3547	2297
TT segment	un	augmento	en	la exposición	de
ST-TT Alignment					
ST segment	in	the	increasing	exposure	of

Fig. 2 AGs and aspects of associated *Munits*

pause of 1172 ms. PU-*b* lasts 2391 ms and shows a similar gazing pattern as PU-*a*, with TT monitoring activities, while toward the end of PU-*b*, the gaze moves back to the ST to scan the next piece of text.

The progression graph in Fig. 1 shows the coordination of reading and typing activities and places each event (keystroke and fixation) in relation to the source and the emerging target text. Each line in the figure represents the activities that are related to the source word on the left *y*-axis. Figure 1 shows that almost every English ST word in this section is aligned to one Spanish target word (except for *exposure—la exposition*).

Figure 2 shows how the information from Fig. 1 is distributed over the five AGs. It shows the corresponding AGs with their alignment links between the source and the target segments, as well as the duration of the *Munit*, and the pause that precedes each *Munit*. The inversion of English-Spanish *in–en* also becomes apparent. The activities of PU-*a* which lasts 3250 ms are distributed over four *Munits* that are part of AG-1–4. Only 266 ms of it are allocated to *Munit*-a-2 which is the duration needed to produce the translation of "the" for AG-2. Similarly, for AG-3, only the duration of keystroke sequences that relate to the translation of "increasing" is extracted from PU-*a*. The production of "augment" lasted 2079 ms, as there was an immediately corrected typo (a[yum]ento). Notice that AG-4 and the production of "la exposición" stretches over PU-*a* and PU-*b*, where "la" is produced in PU-*a* and "exposición" is in PU-*b*. The pause between the two PUs is, therefore, considered part of the translation *Munit* production duration.

A *Munit* is preceded by a processing *Pause1* (i.e., a translation act). Alves and Vale (2011) assume that "This [first] pause may be a pause for planning or searching for a translation alternative, an assessment of the previous production or the beginning of a new reading phase" (2011: 107). Within the TPR-DB, we associate the duration of the *Munit*, as well as the duration of the preceding pause with AGs. In addition, we also record gaze data (ST fixations, TT fixations) among others with each AG.[6] The above considerations lead to four hypotheses, which we will test in the next sections:

[6]For a complete description of the features in the CRITT TPRDB, see https://sites.google.com/site/centretranslationinnovation/tpr-db/features

Length of *Pause1*:

1. More extensive scanning of the ST leads to a longer *Pause1*.
2. More complex and less compositional AGs are preceded by a longer *Pause1*.

Number of *Munits* per AG (Self revision):

3. More complex AGs will be revised more frequently than less complex ones.
4. Verbal AGs have a longer *Dur1* in the default phase.

3 Determinants of *Pause1* Duration

The *Pause1* preceding the first translation response is part of the translation act. Its duration is an indicator of effort for which eye movements provide "data that speaks to cognitive aspects of the translation process" (Halverson 2019: 198). As mentioned above, numerous suggestions have been made about how to define a pause threshold that could discriminate between problematic and non-problematic translation. According to Franchak and Adolph (2014, 3), a boundary at which an action shifts from possible to impossible (or for that matter from non-problematic to problematic) is termed a *critical point.* However, according to the "extent that performance is variable across repeated trials," this "boundary" may not be categorical. In line with these considerations, we consider *Pause1* a continuous variable.

The distribution of the $Log_{10}Pause1 = log_{10}(Pause1)$[7] for the data that we use in this study is shown in Fig. 3, for *Pause1 > 10 ms*. The graph suggests that there might indeed be two distributions overlapping, one with a peak shortly after 10^2 (i.e., around 250 ms) and another flatter distribution with a peak close to 10^3 (i.e., around 1000 ms). It is nevertheless far from obvious where in that continuum a boundary should be located that would separate the pause into two categories indicating challenged and non-challenged processing. Instead of considering a pause a binary categorical variable, separating problematic from non-problematic translation, we consider *Pause1* a continuous variable and investigate parameters that might affect its duration.

We investigate the duration of *Pause1* from two angles: first we look at the distribution of observed gaze behavior on the source or the target text during *Pause1*. Then we investigate properties of the translation burst of *Munit1* that follows *Pause1*.

[7]We use $\log_{10}$ ($Log_{10}Pause1$) in this figure for better interpretation. We use $\log_e$ ($LogPause1$) in the analyses below.

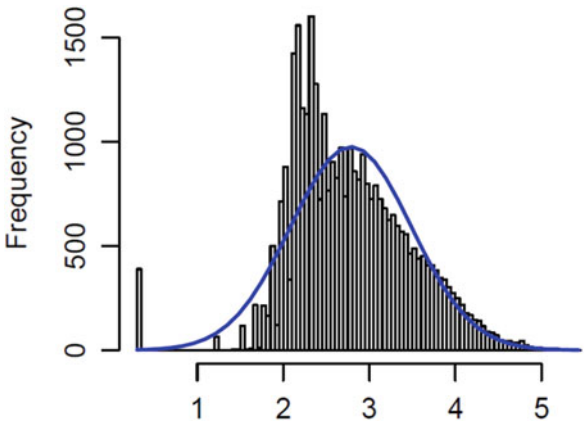


Fig. 3 Distribution of Log_{10} Pause1 length across six languages

3.1 Source and Target Text Reading Patterns During Pause1

To assess whether translators spent more time on the source or the target text time during *Pause1*, we computed the relative source and target text reading *(Pause*1*TrtSR and Pause*1*TrtTR*, respectively) based on the observed gaze duration on source and target text during *Pause1* as shown in Eq. (1).

$$\begin{aligned} &\text{(a) Pause1TrtTR} = \frac{\text{total reading time on target during Pause1}}{\text{pause duration (Pause1)}} \\ &\text{(b) Pause1TrtSR} = \frac{\text{total reading time on source text during Pause1}}{\text{pause duration (Pause1)}}. \end{aligned} \tag{1}$$

We ran two linear regression models and tested the effect of *Pause*1*TrtSR*, *Pause*1*TrtTR,* target language (*TL*) and word class (*PoS*) as independent variables on *Pause1*. All four independent variables have a significant effect on *Pause1*.

Figure 4 (top) plots the relative distribution of gaze on the target text (left) and source text (right) for the six languages as a function of *Pause1* duration (F: 418.8, df = 34,088, $p < 2.2e{-}16$). Translators spend proportionally more time reading the target text when *Pause1* is short. This effect is significant for all languages. With increasing length of *Pause1*, translators are more likely to spend proportionally more time consulting the source text. This tendency is also significant for all languages except Danish. The effect is slightly stronger for the source text and different across different languages. An explanation for this observation may be that longer pauses are needed to accumulate a new chunk of information from the source text, while reading or refreshing the target text prior to the production of a first translational response can be achieved in relatively shorter pauses.

Figure 4 (bottom) shows the effect of source and target text reading behavior during *Pause1* depending on the word class of the source word for which a translation is being produced (F: 46.52, df = 14,071, $p < 2.2e{-}16$). There is a significant negative main effect of proportional target text reading on *Pause1* duration and a significant positive effect for proportional source text reading. The

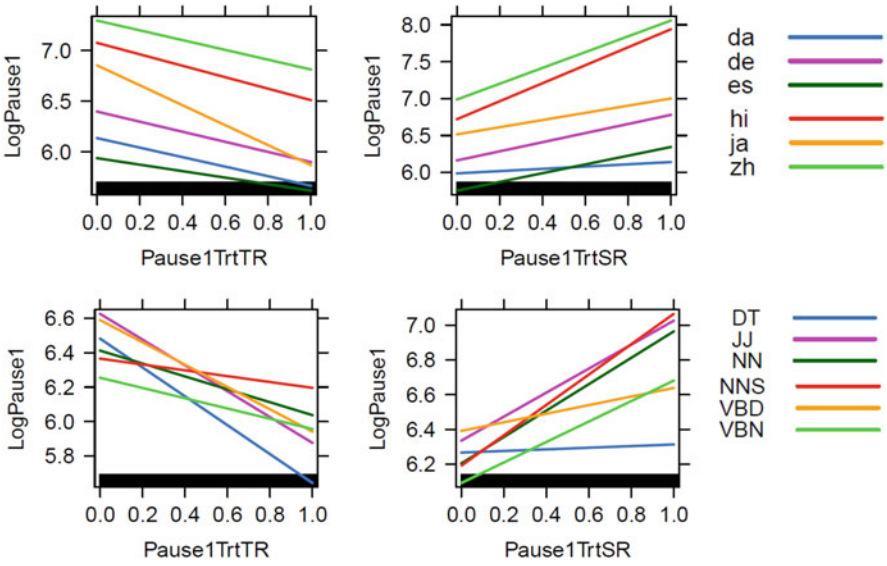


Fig. 4 Relative reading time of target (left) and source (right) text during *Pause1*: top effect for target six languages, bottom effect for six major word classes

figure shows six of the more frequent and important word classes, determiner (DT), adjective (JJ) noun-singular (NN), noun-plural (NNS), verb-past tense (VBD), and verb-past participle (VBN), and the interaction effect of word class with gaze distribution on *Pause1*. The determiner (DT) has a particularly strong negative effect on *Pause1* (Fig. 4, bottom left) for proportional target text reading, but no significant effect for source text reading (Fig. 4, bottom right), while the effect is the opposite for plural noun (NNS). An explanation might be that the translation of nouns—in our data in particular plural nouns—requires more contextual information in the source than the translation of a determiner.

3.2 Pause1 and the Translation Product

In addition to the source and target text reading patterns, the duration of *Pause1* also depends on a number of source and target text properties for which the translation is being produced. We examine various parameters of the translation product and the production process as predictor variables, and we fit a regression model to assess their effect on *LogPause1*.

Using a hierarchical regression analysis, we have identified seven variables that have a significant effect on *LogPause1*. Table 3 plots the main effects of these variables without the six target languages and 31 *PoS* tags which are not shown in the table. The overall fit for this model was $r^2 = 0.218$ (df = 34,905, p: $<2.2e{-}16$).

Table 3 Effects of HSTC, SAGnbr, PUnbr, Id, LogDur1, and Prob1 on total LogPause1, all with a significant effect. Thirty-one PoS tags (partially not significant) are not listed in the table

	Estimate	Std. error	*t*-Value	Pr(>|*t*|)
(Intercept)	6.2317	0.0497	125.5120	<2*e*−16
HSTC	0.2125	0.0081	26.1360	<2*e*−16
SAGnbr	0.1650	0.0096	17.1320	<2*e*−16
PUnbr	0.0617	0.0043	14.356	<2*e*−16
Prob1	−0.1171	0.0108	−10.8080	<2*e*−16
Id	−0.0024	0.0002	−12.8240	<2*e*−16
LogDur1	0.1594	0.0070	22.548	<2*e*−16

We also tested the interaction with the six languages (TL). The final analysis that was run in R is shown in Eq. (2):

$$\text{LogPause1} \sim (\text{HSTC} + \text{SAGnbr} + \text{PUnbr} + \text{Id} + \text{Prob1} + \text{PoS}) \times \text{TL} \tag{2}$$

The following independent variables have a significant effect on *LogPause1*:

- *HSTC*: a *joint ST-TT- alignment crossing entropy*, as introduced in Carl (this volume, Chap. 5). It takes into account the joint probability of the source group (s), target group (t), and a distortion (c) for k alternative translations. *HSTC* indicates the degree to which a source token allows for rendered literal translations and shown in Eq. 3:

$$HSTC = \sum\nolimits_k p\,(s_k, t_k, c_k) \times \log_2\,(1/p\,(s_k, t_k, c_k)) \tag{3}$$

HSTC has a positive effect on *LogPause1* suggesting that more literal translations imply shorter durations of *Pause1*, and less literal translations are preceded by a longer *Pause1*. As can be seen in Fig. 5, the effect is negative for Hindi (hi) and not significant for Japanese (ja). This might be due to the fact that Hindi and Japanese have on average larger AGs, which indicates less compositionality.

- *SAGnbr*: indicates the number of source tokens in an AG. It is an indicator of complexity and non-compositionality of the AG. It is added as a confounding variable, as *Dur1*, *HSTC*, and *PUnbr* may depend on the length of the AG. *SAGnbr* has a positive effect on *LogPause1* indicating, as with *HSTC*, that less compositional translations are preceded by longer *LogPause1*. Also as for *HSTC*, there is no significant effect of *SAGnbr* for Hindi.
- *LogDur1*: the log production duration of *Munit1*. It is positively correlated with *LogPause1*: longer *Dur1* is preceded by longer pauses. The effect for Danish and Spanish is not as strong as for the other languages.
- *PUnbr*: represents the number of production units by which the translation was produced. As mentioned above, a PU is defined by a duration of 1 s or more in which no keystroke occurs. Keystroke pauses are indicators of disfluent writing and indicators of elevated cognitive effort. They are more likely to occur

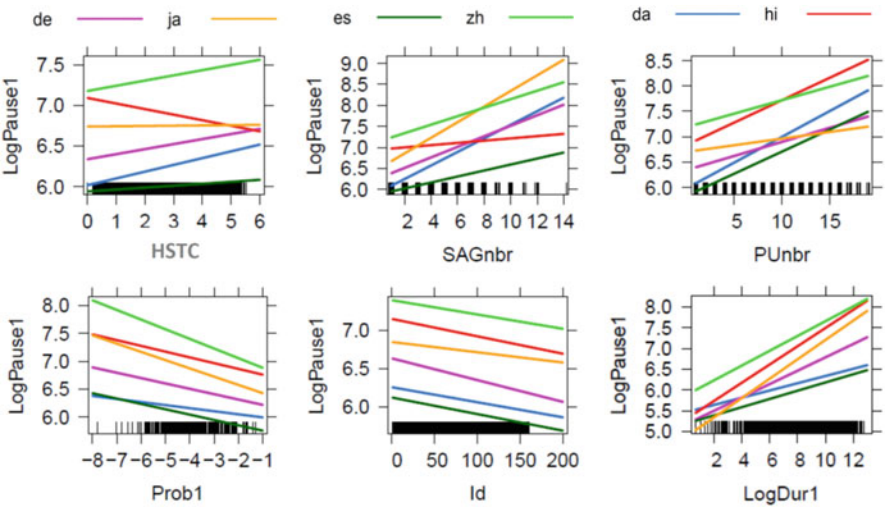


Fig. 5 Interaction effects of various parameters on pause duration (*LogPause1*) preceding first translation responses

where translation material is complex, new or unfamiliar (e.g., Carl and Kay 2011). There is a significant positive effect of *PUnbr* on *LogPause1* and a weak correlation between *PUnbr* and the other indicators of translation literality, i.e., *HSTC* ($r = 0.24$) and length of the AG ($r = 0.29$).

- *Id*: is the word number in a text. *Id* has a significant negative effect on *LogPause1*, indicating a facilitation effect. The facilitation effect predicts that text production (and translation) becomes more fluent further into the text (cf., e.g., Schaeffer et al. 2016). The facilitation effect is similar across all six target languages.
- *Prob1*: is the log frequency of the English source text word (according to the BNC) which has a negative effect on *Pause1*: more frequent source words lead to shorter *LogPause1*. *Prob1* is related to the self-information (*I*) of a word ($I = -Prob1$). It is thus possible to say that the self-information of a word has a positive effect on *Pause1*.
- *PoS*: 26 of the 31 part of speech tags have a significant effect on *LogPause1*. While most *PoS* tags have a similar effect on *LogPause1*, some categories, including articles, adverbs, particles, pronouns, some adjectives, and WH-words have quite different effects for different languages. Ogawa et al. (this volume, Chap. 6) address some of these differences in more detail.

4 Munits and Revision Patterns

Dragsted and Hansen (2008) as well as Schaeffer and Carl (2017) measure the *immediacy* of first translational response as the lag of time between the first fixation on a source word and the moment when the translator starts typing the translation

of the word. This so-called *eye-key span* is part of a translation act as defined above and has been interpreted as an indicator of translation effort. The more time elapses between the first fixation on a source word and the typing of its translation, the more effortful the translation is assumed to be. Schaeffer and Carl (2017) show that the eye-key span correlates with the variation of word choices (*HTra*) and syntactic reordering (*Cross*) in a corpus of alternative translations.

We assess the *durability* of the first translation response, i.e., how likely it is that the first translation will survive until the final product. Most first translation responses will show few or no revisions during the translation process (they are more durable), whereas other initial translations will be revised once or more often; the first translation is hence less durable. The revision of a translation is a translation event; it can be quantified, for instance, by the number of keystrokes, the amount of modification, or the lag of time between successive modifications. Within the CRITT TPR-DB, a revision is recorded as a *Munit*. As explained above, a *Munit* consists of one or more successive keystroke(s)—which may be an insertion or a deletion—that relate(s) to an AG. The number of *Munits* that the translation of the words in an AG is involved in is thus an indicator of its translation effort, since each revolving modification is an indicator of restructuring or reconsidering the translation, which implies additional considerations on the part of the translator.

Table 4 shows the distribution of revisions (*Munits)* per word across the six languages in the translation data. The first (draft) translation version is counted as *Munit1*, and thus only values >1 are actually proper revisions. The table shows that for Danish, only slightly more than 5% of the words are revised at least once (Munit >1), while for Japanese, this is the case for almost 20% of the words. That is, the vast majority of textual material in the final translation product corresponds to their first translation renderings.

Carl and Schaeffer (2017) use *Munits* to assess revision processes of procedural and conceptual encodings in the translation process. Relevance theory (Gutt 2000) predicts that conceptually encoded information is easier to translate than procedurally encoded information, as conceptual encoding exhibits a "relatively stronger interpretive resemblance between source and target texts" (Alves and Gonçalves (2003) : 20) and also "conceptual representations can be brought to consciousness; procedures can not" (Wilson and Sperber 1993: 16). While translators learn how to consciously manipulate information that is conceptually and procedurally encoded, translations of conceptual representations may be easier to learn and reason about. Sekino (2015) shows that the processing effort is greater when dealing with procedural encodings, as compared to conceptual encodings in from-scratch

Table 4 Percentage of revised translations across six languages

Munit	da	de	es	hi	ja	zh
1	94.35	83.52	88.32	81.78	80.80	85.80
2	5.20	12.62	9.65	14.50	14.29	12.34
3	0.42	3.01	1.62	2.61	3.69	1.44
4	0.02	0.63	0.28	0.73	0.97	0.35

translations and in post-editing tasks in terms of keystrokes, fixation counts, and fixation duration. Similarly, Carl and Schaeffer (2017) report that:

> While the perception [i.e. translation act] of procedurally encoded information seems to be less effortful than that of conceptually encoded information, our findings indicate the reverse relation for translation production. Taking the number of revisions [i.e. translation event (Munit)] as an indicator for the effort in translation production, our dataset shows that the generation of translations for procedurally encoded information is more difficult than that of conceptually heavy words (Carl and Schaeffer 2017:108)

Based on the *HTra* and *Cross* measures, Carl and Schaeffer (2017) also show that conceptual encodings lead to more literal translations than procedural encodings.

We assess the effect of *HTra*, *CrossS*, and *HSTC* on the number of revisions (*Munit)* in two regression analyses shown in Eq. (4). We only considered from-scratch translations (T) that had a production duration *Dur* > *10 ms*[8] and took out instances with 2.5 SD (standard deviation). All three independent variables had a significant effect on the number of revisions (*Munit*) (df = 33,917, $p < 2.2e{-}16$). Figure 6 plots a regression analysis showing significant interaction effects between the six target languages. The strongest effect of *HSTC* on revision can be observed for German and Hindi, the weakest for Danish. Japanese word-order has a relatively weak effect on revision.

$$\begin{aligned} &\text{(a)}\ \ \text{Munit} \sim (\text{HTra} + \text{abs}\,(\text{Cross})) \times \text{TL}. \\ &\text{(b)}\ \ \text{Munit} \sim \text{HSTC} \times \text{TL}. \end{aligned} \tag{4}$$

HTra and *HSTC* have a relatively stronger effect on *Munit* for German (de) and Hindi (hi) than for Danish (da). This may be due to the fact that Danish has fewer revisions in general than the other languages. It might play a role here that part of the Danish data was collected under time constraints (Hvelplund 2011) and that 50% of

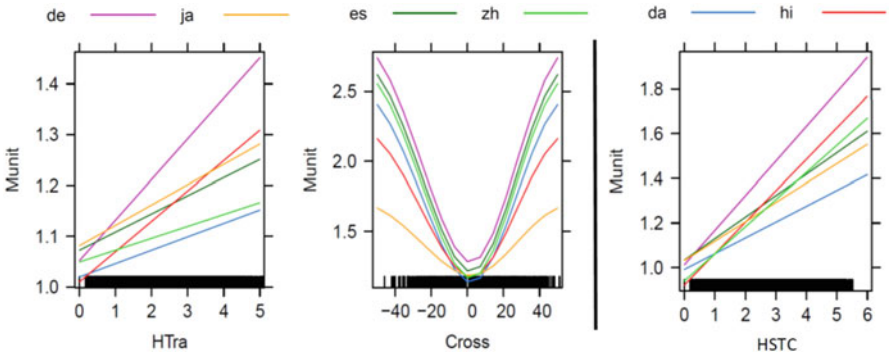


Fig. 6 Effect of semantic and syntactic (left) and literal similarity (right) on revision

[8]This excludes words with no alignment links and words that were copied and pasted.

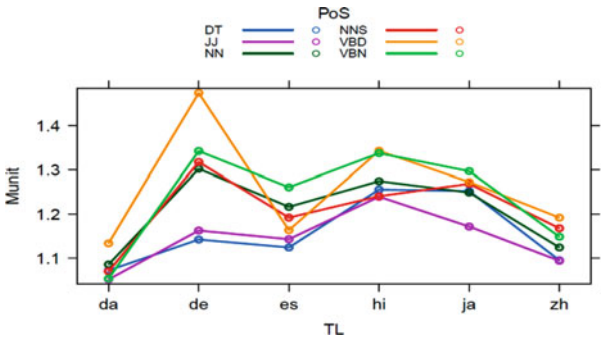


Fig. 7 revision behavior for several word classes and languages

the translators were professional, while for German, all translators were students with no time constraints.

Figure 7 shows the effect of the six *PoS* tags that were discussed in Sect. 3.1 on *Munit*. The analysis suggests that different word classes require different revision behavior. Verbs (VBD and VBN) are most often revised, while determiner (DT) and adjectives are least often revised, and thus most durable. Nouns (NN and NNS) seem to occupy a middle position. It is also interesting to observe that past tense verbs (VBD) seem to be most frequently revised for most languages other than Spanish. Ogawa et al. (this volume, Chap. 6) give a much more detailed description of those differences and similarities across three languages (Arabic, Japanese, Spanish).

5 Discussion and Conclusion

In this chapter, we investigate Malmkjær's (2011) *first translational response universal* which posits that translators mentally segment texts into smaller units when they translate. The properties of those units and the first translational responses are assumed to be indicative of the interlingual relationships and the memory of the translators. According to Malmkjær, the length of the translation unit "is limited by the amount of paired text a translator can hold in short-term memory" (2011: 91), and while there may be some variation across different language pairs and different translators, she assumes "that there will be limits on how different a first-response translation can be from its ST." Malmkjær supposes that it would be interesting to see what these limits are.

The first translational response is an early measure, which elucidates how a translator applies "the first meaning" that comes into their mind. Early measures in cognitive studies are indicative of automatic behavior, triggered through ST priming processes (see also Carl this volume, Chap. 14), while later processes, such as translation revisions, are indicative of conscious processing taking into account additional stimuli, such as TL grammatical or stylistic consideration. It is assumed

that a priming stimulus (e.g., an ST segment) has an impact on the activation of implicit memory mechanisms (e.g., the translation of that segment) and reduces effort to recall. In bilingualism and translation studies, priming studies investigate how a source language word, sentence, or structure has facilitating effects on a successive target language stimulus and/or production. Numerous priming studies suggest that bilinguals have access to shared interlingual mental representations which connect source and target language structures in a nonselective manner and that these representations are activated automatically (e.g., Tokowicz and Kroll 2007; Hartsuiker et al. 2008; see also Carl this volume, Chap. 14). Among other things, it has been shown that:

1. Priming effects exist for shared phonetic (e.g., cognates), semantic (Dimitropoulou and Andoni 2011; Schoonbaert et al. 2011), and syntactic structures (Bangalore et al. 2016; Maier et al. 2017). See also Heilmann and Llorca-Bofí this volume, Chap. 8.
2. Priming effects are stronger when ST-TT links are un-ambiguous. A number of studies (Tokowicz and Kroll 2007; Laxén and Lavaur 2010; Prior et al. 2013; Eddington and Tokowicz 2013) show that translation recognition, as well as translation production, is slowed down if a word has more translation alternatives. Schaeffer et al. (2016, 183) show that "the number of translation alternatives for a single word and differences between source and target text in terms of word order have an effect on very early and late eye movement measures."
3. In particular lexical priming effects decrease when items intervene between the trigger and the target: reversing word order is detrimental to priming effects. Effects are stronger if priming stimulus and target are adjacent, and not separated by intervening linguistic material (Hartsuiker et al. 2008).

These properties of priming effects predict that the first translational response is more automatic if the semantic overlap between the source and the target language is larger—which implies also less translation choice; if segments translate compositionally, ideally in a one-to-one fashion, as, for instance, exemplified in Example 3 and Fig. 2; and if the languages are syntactically closer to each other with no, or few long-distance reordering and only small alignment crossings. Conversely, nonliteral translations—as, e.g., discussed in Example 2—are likely to show weaker priming effects.

In this chapter, we assess *the first translational response universal* by means of an empirical investigation across six languages (English to Danish, German, Spanish, Hindi, Japanese, and Chinese) and the *multiLing* dataset. This dataset is particularly suited to investigating translation universals, due to the diversity of languages and different profiles of the translators involved, the multitude of translation modes, and the richness of the logged data.

We distinguish between three phases and a final translation product to investigate the first translational response:

1. An initial *translation act* in which the translator presumably mentally prepares the successive translation event of an ST segment and which immediately precedes the typing of a translation. This phase is referred to as *Pause1* during which we may have traces of the translator's gaze data on the source or/and target text.
2. An initial *translation event*, referred to as micro unit-1 (*Munit1)*, which immediately follows *Pause1* and in which the translator types the first draft translation for the ST segment. *Munit1* is, thus, characterized by keystroke activities, but we may also observe parallel gaze activities, as shown in Fig. 1.
3. *Munit1* can be followed by one or more optional revision phases in which the translation produced during *Munit1* may be revised. We refer to each successive revision as *Munit2 … n,* with the understanding that all *Munits 1* to *n* refer to the same ST segment.
4. The final translation of the ST segment for which we can determine the linguistic properties.

This study investigates the *first translational response universal* from two angles: (1) What are the parameters that determine the duration of *Pause1*? (2) What is the effect of translation literality on the number of revisions (i.e., number of *Munits*)? As suggested in Carl (this volume, Chap. 5), we define rendered translation literality as the degree to which translations are compositional, monotone, and entrenched, in line with TL grammatical constraints. Higher scores of rendered literality imply higher degree of cross-lingual semantic and syntactic similarity. Our results confirm that the possibility to render translations literally has a strong effect on the *Pause1* and the revision behavior. Our findings show that *Pause1* tends to be longer, and thus translation seems to be more challenged, if:

- AGs are longer (*SAGnbr*): longer non-compositional, phrasal translations require longer pre-processing than translations that are produced compositionally, i.e., word-for-word.
- The translated segment has higher self-information (*Prob1*): ST words that are less frequent are preceded by longer *Pause1*.
- Translations are produced in a more interrupted manner (*PUnbr*): there are more inter-keystroke pauses >1 s in the translation production.
- Information is mainly retrieved from the source text (*Pause1TrtSR*): the take-in of new information from the ST is more time-consuming than (re-) checking the TT what was (just) typed.

We also observed that the duration of *Munit1* (*Dur1*) has a positive effect on *Pause1*. This may be an effect of the literality and self-information of the produced translation, since translation equivalents involving longer ST segments and/or less frequent words will probably also take longer to produce and are likely to engender more hesitations. We also observed a task facilitation effect: *Pause1* tends to become shorter as the translator progresses in the text (*Id*). These effects are significant but slightly different across the six languages. In particular, the rendered literality effect on *Pause1* is less pronounced for Hindi.

Our second question relates to the literal translation hypothesis, which posits that translators "tend to start from a literal version of the target text, and then work towards a freer version" (Chesterman 2011, 23). Our findings support this literal translation hypothesis as we observe that the number of revisions (*Munits)* has a positive effect on translation literality, i.e., more revisions lead to higher lexical variation and more cross-linguistic syntactic reordering. Carl (this volume, Chap. 5) introduces a *joint ST-TT-crossing entropy* (*HSTC*) as a composite measure for syntactic and semantic variation, which was also used in this study. Taken together, our findings suggest that more challenged first translational responses are indicative of less literal final translations. However, we could not assess to what extent also the translators' short-term memory plays a role in this process.

References

Alves F, Gonçalves J (2003) A relevance theory approach to the investigation of inferential processes in translation. Benjamins Translation Library 45:3–24

Alves F, Vale D (2009) Probing the unit of translation in time: aspects of the design and development of a web application for storing, annotating, and querying translation process Data. Across Lang Cult 10(2):251–273

Alves F, Vale D (2011) On drafting and revision in translation: a Corpus linguistics oriented analysis of translation process data. TC3 1(1):105–122

Bangalore S, Behrens B, Carl M, Ghankot M, Heilmann A, Nitzke J, Sturm A (2016) Syntactic variance and priming effects in translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 211–238

Carl M (2012) Translog-II: a program for recording user activity data for empirical reading and writing research. LREC 12:4108–4411

Carl M (this volume) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Dragsted B (2012) Inside the monitor model: processes of default and challenged translation production. TC3 2(1):127–145

Carl M, Kay M (2011) Gazing and typing activities during translation: a comparative study of translation units of professional and student translators. Meta 56(4):952–975

Carl M, Schaeffer M (2017) Sketch of a noisy channel model for the translation process. Empirical modelling of translation and interpreting. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) , vol 2017. Language Science Press, Berlin, pp 71–116

Chesterman A (2011) Reflections on the literal translation hypothesis. In: Alvstad C, Hild A, Tiselius E (eds) Methods and strategies of process research. John Benjamins, Amsterdam, pp 23–35

Chesterman A (2015) Models of what processes? In: Ehrensberger-Dow M et al (eds) Describing cognitive processes in translation. Benjamins, Amsterdam, pp 7–20

Data-Bukowska E (2019) Priming as cognitive motivation for the 'first translational response'. Linguist Siles 40:185–204. https://doi.org/10.24425/linsi.2019.129409

Dimitropoulou M, Andoni JAD, Carreiras M (2011) Masked translation priming effects with low proficient bilinguals. Mem Cogn 39(2):260–275. https://doi.org/10.3758/s13421-010-0004-9

Dragsted B, Hansen IG (2008) Comprehension and production in translation: a pilot study on segmentation and the coordination of Reading and writing processes. Copenhagen Stud Lang 36:9–29

Eddington CM, Tokowicz N (2013) Examining English-German translation ambiguity using primed translation recognition. Biling Lang Congn 16(2):442–457
Franchak J, Adolph K (2014) Affordances as probabilistic functions: implications for development, perception, and decisions for action. Ecol Psychol 26(1–2):109–124. https://doi.org/10.1080/10407413.2014.874923
Gutt E-A (2000) Translation and relevance: cognition and context, 2nd edn. St Jerome, Manchester
Halverson SL (2015) Cognitive translation studies and the merging of empirical paradigms. The case of 'literal translation'. Transl Spaces 4(2):310–340
Halverson SL (2019) 'Default' translation: a construct for cognitive translation and interpreting studies'. Transl Cogn Behav 2(2):187–210
Hartsuiker RJ, Bernolet S, Schoonbaert S, Speybroeck S, Vanderelst D (2008) Syntactic priming persists while the lexical boost decays: evidence from written and spoken dialogue. J Mem Lang 58:214–238
Heilmann A, Llorca-Bofí C (this volume) Analysing the effects of lexical cognates on translation properties: a multi-variate product and process based approach. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Hvelplund KT (2011) Allocation of cognitive resources in translation: an eye-tracking and key-logging study. Ph.D. thesis, Copenhagen Business School. https://static-curis.ku.dk/portal/files/131448126/2011_Allocation_of_cognitive_resources_in_translation_Hvelplund.pdf
Jakobsen AL (2005) Instances of peak performance in translation. Lebende Sprachen 50(3):111–116
Jakobsen, AL. (2011). Tracking translators' keystrokes and eye movements with Translog. Methods and strategies of process research John Benjamins Publishing Company Amsterdam 37–55
Kumpulainen M (2015) On the operationalisation of 'pauses' in translation process research. Transl Interpreting 7(1):ti.106201.2015.a04
Lacruz I, Shreve GM (2014) Pauses and cognitive effort in post-editing. In: O'Brien S, Balling L, Carl M, Simard M, Specia L (eds) Post-editing: processes, technology and applications. Newcastle upon Tyne, Cambridge Scholars Publishing, pp 246–272
Laxén J, Lavaur J-M (2010) The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. Biling Lang Congn 13(2):157–183
Maier RM, Pickering MJ, Hartsuiker RJ (2017) Does translation involve structural priming? Q J Exp Psychol 70(8):1575–1589
Malmkjær, Kirsten. (2011). Translation universals. In The Oxford handbook of translation studies Kirsten Malmkjær and Kevin Windle, 83–93. Oxford: Oxford University Press
O'Brien S (2006) Pauses as indicators of cognitive effort in post-editing machine translation output. Across languages and cultures 7(1):1–21. https://doi.org/10.1556/Acr.7.2006.1.1
Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and source-text background information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Prior A, Kroll JF, Macwhinney B (2013) Translation ambiguity but not word class predicts translation performance. Computational Modeling of Bilingualism 16(2):458–474
Schaeffer MJ, Carl M (2017) Language processing and translation. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) Empirically modelling translation and interpreting. Language Science Press, Berlin, pp 117–154. (translation and multilingual natural language processing, no. 7)
Schaeffer M, Dragsted B, Hvelplund KT, Balling LW, Carl M (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, Cham, pp 183–210
Schoonbaert S, Holcomb PJ, Grainer J, Robert RJ, Hartsuiker J (2011) Testing asymmetries in noncognate translation priming: evidence from RTs and ERPs. Psychophysiology 48(1):74–81. https://doi.org/10.1111/j.1469-8986.2010.01048.x

Sekino K (2015) An investigation of the relevance-theoretical approach to cognitive effort in translation and the post-editing process. Transl Interpreting 7(1):375. http://www.trans-int.org/index.php/transint/article/view/375

Tirkkonen-Condit S (2004) Unique items – over- or under-represented in translated language? In: Mauranen A, Kujamäki P (eds) Translation universals: do they exist? John Benjamins, Philadelphia, PA, pp 177–184

Tirkkonen-Condit S (2005) The monitor model revised: evidence from process research. Meta 50(2):405–414

Tokowicz N, Kroll JF (2007) Number of meanings and concreteness: consequences of ambiguity within and across languages. Lang Cogn Process 22(5):727–779

Wilson D, Sperber D (1993) Linguistic form and relevance. Lingua 90:1–25. http://www.dan.sperber.fr/wp-content/uploads/1993_wilson_linguistic-form-and-relevance.pdf

Metrics of Syntactic Equivalence to Assess Translation Difficulty

Bram Vanroy, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken

Abstract We propose three linguistically motivated metrics to quantify syntactic equivalence between a source sentence and its translation. Firstly, syntactically aware cross (SACr) measures the degree of word group reordering by creating syntactically motivated groups of words that are aligned. Secondly, an intuitive approach is to compare the linguistic labels of the word-aligned source and target tokens. Finally, on a deeper linguistic level, aligned syntactic tree edit distance (ASTrED) compares the dependency structure of both sentences. To be able to compare source and target dependency labels, we make use of Universal Dependencies (UD). We provide an analysis of our metrics by comparing them with translation process data in mixed models. Even though our examples and analysis focus on English as the source language and Dutch as the target language, the proposed metrics can be applied to any language for which UD models are attainable. An open-source implementation is made available.

Keywords Translation studies · Computational linguistics · Tree edit distance · Syntax

1 Introduction

Readability prediction is a well-studied problem. Traditional readability formulas (e.g. Flesch–Kincaid Grade Level (Kincaid et al. 1975) and Gunning Fog Index (Gunning 1952)) typically use shallow source text features such as average word and sentence length and word frequency to assess the reading difficulty level of a given text. Recently, more complex lexical, syntactic, semantic and discourse text features have been used (see, for instance, Schwarm and Ostendorf (2005); Francois and

B. Vanroy (✉) · O. De Clercq · A. Tezcan · J. Daems · L. Macken
Ghent University, Ghent, Belgium
e-mail: bram.vanroy@ugent.be; orphee.declercq@ugent.be; arda.tezcan@ugent.be; joke.daems@ugent.be; lieve.macken@ugent.be

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_10

Miltsakaki (2012); De Clercq et al. (2014); De Clercq and Hoste (2016) and Collins-Thompson (2014) for an overview). The efforts in readability research contrast sharply with research into 'translatability': there are no well-established methods yet to assess the difficulty level of a translation task. That is not to say that translation difficulty itself has not been studied, though. In fact, defining translation difficulty has been approached from a number of different directions.

It has been shown that genre, registerial and even cultural factors influence the choices translators have to make (e.g. Borrillo (2000, Section 3) concerning literary translation and Steiner (2004) on registerial differences), which may introduce difficulties of its own. In addition, there is no doubt that individual translators may face different issues when translating the same text, and they may even choose to translate the same text differently (see, for instance, Dragsted (2012)). In this chapter, however, we will focus on the source and target text itself.

According to Campbell (1999) and Sun (2015), translation difficulty can be attributed to linguistic source text factors and translation-specific factors. For the source text factors, we can refer to the vast literature on readability research (see the survey by Collins-Thompson (2014) for an overview), though a few findings specific to translation should be highlighted. Liu et al. (2019) demonstrated that *source* text complexity plays an important role in perceived translation difficulty, which supports earlier findings by Mishra et al. (2013). Mishra et al. introduced a metric of translation difficulty that is based on source text features alone, namely sentence length, degree of polysemy and structural complexity. Campbell (1999) looked into translation difficulty from an empirical point of view and identified several source text elements that were difficult to translate across different target languages, such as multi-word units, complex noun phrases, abstract nouns and verbs. Campbell continued their research and developed the Choice Network Analysis (2000) in an attempt to model the mental process that underlies translation, particularly the multitude of choices that translators can choose from a given specific source text. Building on this, Carl and Schaeffer (2017) documented longer translation times when more elaborate choices were at the translators' disposal. This indicates that having more options available can increase the translation difficulty in terms of duration.

However, readability prediction and source text complexity alone do not suffice to adequately assess the *translation* complexity level of a given source text (Daems et al. 2013; Sun and Shreve 2014). This is not surprising because readability prediction is not designed to take into account co-activation of shared bilingual resources. Specifically, Sun and Shreve (2014) and Sun (2015) state that translation-specific difficulties can be ascribed, in part, to the lack of equivalence due to inherent differences between languages. Hence, this chapter will focus on the equivalence between the source and target text, specifically their syntactic similarity.

The notion of syntactic equivalence in a multilingual setting is not easy to define (see the next section) because syntax in itself is such a broad concept, so in this chapter we restrict *syntactic equivalence between a source and target segment* to mean three things:

(1) a. differences in word (group) order,
b. differences in dependency labels of aligned words (e.g. a subject (`nsubj`) is translated as an object (`obj`))
c. differences in syntactic structure (dependency tree).

In Sect. 2, we will first discuss background literature concerning the importance of syntactic equivalence with respect to translatability and previous research of equivalence. In Sect. 3, we then introduce three linguistically motivated metrics to quantify syntactic equivalence between a source sentence and its translation. Firstly, we introduce a metric to capture linguistic word group reordering (syntactically aware cross [SACr]). The next metric measures parse tree label changes between source and target sentences. Thirdly, we introduce a method to calculate tree edit distance between aligned dependency trees (aligned syntactic tree edit distance [ASTrED]). To illustrate the different proposed metrics, we will discuss two example sentence pairs in Sect. 4 to highlight how each metric accounts for different linguistic phenomena. As a proof of concept, we also apply our metrics to an existing dataset and measure the effect syntactic changes may have on the translation process by using mixed models (Sect. 5). Finally, we end with a conclusion and thoughts for future work concerning quantifying syntactic equivalence (Sect. 6).

2 Related Research

2.1 Background

In process-based translation studies, literal translation is conceived as the easiest way to translate a text and has been suggested as the default mode of translation, which is only interrupted by a monitor that alerts about imminent problems in the outcome (Tirkkonen-Condit 2005, and Carl, this volume, Chapter 5). In other words, translators will translate a source text literally into the target text, but as soon as an issue is encountered, translators stop working in the literal translation mode and try to find a more appropriate solution. Asadi and Séguinot (2005), for instance, observed that one group of translators processed the source text in short phrase-like segments. They translated while reading the text and followed the source language syntax and lexical items closely but then rearranged the completed text segments to create a more idiomatic target text. Literal translation, in this sense of translating word per word, is identical to the concept of *simple transfer* in transfer-based MT, which can occur when the lexical surface forms are the only required differences between the source and target segments for a successful translation. In other words, when the underlying structure of the segments is the same, a literal translation can happen and only the lexical values need to be changed (Andersen 1990; Chen and Chen 1995).

From a cognitive perspective, literal translation is often explained by priming (Hansen-Schirra et al. 2017), i.e. the process in which the production of an output (in the case of translation, the target sentence) is aided or altered by the presentation of a previously presented stimulus (in the case of translation, the source sentence). Priming can occur at different linguistic levels including the morphological, semantic and syntactic levels.

In Carl and Schaeffer (2017, 46), building on earlier work (Schaeffer and Carl 2014), 'literal translation' is defined by three criteria:

(2) a. each ST [source text] word has only one possible translated form in a given context;
b. word order is identical in the ST and TT [target text];
c. ST and TT items correspond one to one.

To quantify the first criterion 2a, they use word translation entropy, which indicates the degree of uncertainty to choose a particular translation from a set of target words based on the number and distribution of different translations that are available for a given word in a given context. To measure the second and third criteria, they use word crossings (Cross) calculated on word-aligned source–target sentences.

Criteria 2b and 2c for literal translation relate closely to what we consider syntactic equivalence as described in 1. 1a (differences in word (group) order) relates to criterion 2b (identical word order) above, and 2c is most similar to 1c: if ST and TT items do not correspond one to one, this must mean that the syntactic structure of the source and target sentences is different. In that respect, our interpretation for syntactic equivalence is closely linked, in part, to the definition of 'literal translation' by Carl and Schaeffer (2017).

The affinity between 'literal translation' on the one hand and equivalence on the other can also be seen in other research. Sun and Shreve (2014), repeated in Sun (2015), suggested that translation difficulties can be attributed to the lack of equivalence between the source and target text. Non-equivalence, one-to-several equivalence and one-to-part equivalence situations can be the root cause of translation difficulties. These situations can appear at both the lexical and syntactic levels. However, Carl and Schaeffer (2017) note that it is possible that a source text has viable ('equivalent') translation options available, but that a plethora of choices actually implies that there is not one single, obvious translation equivalent. In our current study, we will follow the definitions of *natural* equivalence (Pym 2014, Chapter 2), applied to syntax:

– equivalence is a relation of 'equal value' between a source–text segment and a target-text segment;
– equivalence can be established on any linguistic level, from form to function;
– natural equivalence should not be affected by directionality: it should be the same whether translated from language A into language B or the other way round.

Pym (2014) juxtaposes natural equivalence with directional equivalence, which assumes that the equivalency relationship between a source and target text is asymmetric. For a discussion between the two approaches, see the particularly interesting discussion sections (Pym 2014, Chapters 2.7, 3.9).

A similar idea to equivalence is that of translation shifts (Catford 1965), which dates back to an approach to translation that is based on formal linguistics. Catford distinguished two major types of shifts, namely level shifts (e.g. shifts from grammar to lexis in distant languages) and category shifts (e.g. changes in word order or word class). They also contrast obligatory and optional shifts; the former refer to shifts that are imposed as a result of differences in the language systems, whereas the latter term is used to indicate optional choices of the translator.

Bangalore et al. (2015) introduced syntactic entropy and as such expanded translation entropy to the syntactic level. Syntactic entropy measures the extent to which different translators produce the same structure for one source sentence. They analysed a corpus of six English source texts translated into German, Danish and Spanish by a number of translators (24 for German and Danish and 32 for Spanish) and manually coded the following three linguistic features for all translations: clause type (independent or dependent), voice (active or passive) and valency of the verb (transitive, intransitive, ditransitive and impersonal) to quantify the syntactic deviation between translations of the same source text, which is their implementation of syntactic entropy. They obtained lower syntactic entropy values for target sentences that had similar linguistic features as the source segments and obtained higher syntactic entropy values for the cases where they diverged. Moreover, syntactic entropy had a positive effect on behavioural measures such as total reading time on the source text and the duration of coherent typing activity. This study is, to the best of our knowledge, the only study in this field that uses linguistic knowledge to quantify syntactic differences between a source text and its human translation. As an alternative to their three manually annotated linguistic features, we will suggest metrics that can be automatically derived from comparing the syntactic structures of the source and target sentences (Sect. 3).

Carl and Schaeffer (2017) used word order distortion, measured by length of crossing links (called Cross) derived from word-aligned source–target sentences to measure the degree of monotonicity in translations. A bidirectional (symmetric) variant of Cross, which is applicable on either translation direction, was introduced by Vanroy et al. (2019b) (from now on referred to as `word_cross`). Using word alignment in this way provides a fine-grained (word-based) method to quantifying syntactic equivalence. An alternative, coarse-grained, approach was suggested by Vanroy et al. (2019b), who calculated cross on aligned word groups, or *sequences*, rather than single words to calculate syntactic equivalence between English source sentences and their Dutch translations (henceforth called sequence cross or `seq_cross`). These sequences, however, were not linguistically motivated but derived automatically adhering to a set of constraints. The lack of linguistic motivation in `seq_cross` prompted the creation of the three different metrics

described in this chapter. Each metric quantifies a different aspect of syntactic equivalence, but all are based on linguistic knowledge, specifically the syntactic structures of the source and target sentences.

There are two main different ways of annotating syntactic structures: by means of a phrase structure or using a dependency representation. The phrase structure representation sees sentences and clauses structured in terms of constituents. The dependency representation, on the other hand, assumes that sentence and clause structures result from dependency relationships between words (Matthews 1981). While the phrase structure representation is more suitable for analysing languages with fixed word order patterns and clear constituency structures, dependency representations, in contrast, are able to additionally deal with languages that are morphologically rich and have a relatively free word order (Skut et al. 1997; Jurafsky and Martin 2008). The dependency relation that each dependency label represents is relative to its root (with the exception of the root node itself) and is effectively a *to* relationship between the word and its root. For instance, in a sentence 'He eats the cookies', 'He' is an `nsubj` (subject) *to* its root 'eats', 'cookies' is an `obj` (object) to that root and 'the' is a `det` (determiner) to 'cookies'. The dependency labels, then, are actually nodes in a directed acyclic graph, starting from the root node of the sentence (in the example, 'eats') and recursively going down to its dependants. They can be represented as dependency *trees*. The dependency tree of the example sentence 'He eats the cookies' above can be visualised as in Fig. 1.

In recent years, research on automatic parsing methods has increased due to the availability of linguistically annotated corpora (treebanks) for many different languages (Hajič and Zeman 2017; Zeman et al. 2018; Peng et al. 2019). However, despite their availability, the annotation schemes in treebanks vary significantly across languages, such as between the Swedish Treebank (Nivre and Megyesi 2007), the Danish Dependency Treebank (Kromann 2003) and Stanford Typed Dependencies (de Marneffe and Manning 2008). Such differences, in turn, restrict multilingual research on and comparability of syntax and parsing (Nivre 2015; Nivre et al. 2016), as well as research on natural language processing (NLP) that relies on automatic parsers trained on treebanks. Universal Dependencies[1] (UD) is an initiative to mitigate this problem by developing a framework for cross-linguistically consistent morphosyntactic annotation (Nivre et al. 2016), which we will discuss further in Sect. 3.1.

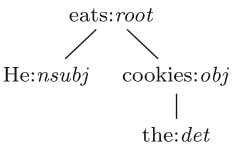


Fig. 1 Example of a dependency tree of the sentence 'He eats the cookies'

[1] See http://universaldependencies.org/ for label explanations, guidelines, and so on.

2.2 *Word Alignment*

The metrics suggested in this research aim to compare given source and target sentences to each other. As a starting point, the sentences need to be word aligned to be able to compare the source and target sides on the subsentential level. In word alignment, source words are aligned with target words as a way to find overlapping points of meaning and syntax. Aligned words should either carry meaning that is similar to their aligned counterpart or cover syntactic or morphological phenomena that are required to translate the aligned word into the desired language (Kay and Roscheisen 1993). In that sense, word alignment does not only involve semantic, conceptual agreement between a source and target sentence but also the (morpho-)syntactic connections between them. As shown in Example 4c, alignments are typically written as pairs of indices of the aligned source and target words separated by a dash, e.g. `0-0 1-1 2-3 3-2 4-4`. Such alignments are often visualised with alignment tables (e.g. Och and Ney 2000, Figure 1), but in this chapter, we opt for line diagrams such as Fig. 2.

In this chapter, we manually aligned the source and target sentences in the examples, but in the global scope of our research, we are interested in translatability, and we envisage to use large corpora to automatically detect and extract patterns that may be indicative of translation difficulties. Manually aligning those corpora is not feasible because of their size. Instead, we rely on automatic alignment systems. In previous research (Vanroy et al. 2019b), we justified using GIZA++ (Och and Ney 2003) in favour of another tool, `fast_align` (Dyer et al. 2013), because of its lower Alignment Error Rate (Och and Ney 2000; Mihalcea and Pedersen 2003).

Because word alignment occurs on the fine-grained word level, the connections between larger groups of words on each side (source and target) are not taken into account. Take, for example, a simple English noun phrase (Ex. 3) that has been translated into a Dutch noun phrase. The determiners ‘The’ and ‘De’ are aligned, and the nouns ‘dog’ and ‘hond’ are aligned to each other. The alignments are given in Example 3b.

(3) a. The dog
De hond
b. 0-0 1-1

In this example, the linguistic relationship between the determiner and its noun is not present in the word alignments; it is not clear that the determiner and the noun are somehow linguistically connected. Generally speaking, this means that metrics based on word-based representation focus on the position and movement into the target language of single words. As an alternative approach, for one of our metrics (syntactically aware cross [SACr]; Sect. 3.2), we want to capture the alignment of word groups. In previous research (Vanroy et al. 2019b), we suggested a naive sequence-based approach, but SACr expands on that by including linguistic information to adjust those sequences. The goal is, then, to have a metric that is based on alignment information, but where the alignment is done between

linguistically motivated groups instead of words or arbitrary sequences. In the example above, that would mean that 'The dog' is aligned, as a group, with 'De hond' rather than as single words. We will expand on aligning word groups rather than single words in the following sections.

2.3 Existing Word-Reordering Metrics

The translation process research database (TPR-DB; Carl et al. 2016) implements a word-based, direction-specific metric for reordering and calculates a cross value based on the movements of words relative to the previously translated word.[2] Vanroy et al. (2019b) take another approach by introducing a translation-direction agnostic variant that measures the number of times that translated words cross each other (`word_cross`). Example 4 (taken from Vanroy et al. 2019b, 104) is visualised in Fig. 2, where each cross is emphasised with a circle. The total number of these crossing links is normalised by the total number of alignments, which constitutes the `word_cross` value. The source and target segments can be aligned as shown in Example 4c. Note that 'me' in the source text is not aligned to an equivalent on the target side. If the source sentence had been translated differently as 'Soms vraagt ze mij waarom …', 'me' could have been aligned with 'mij'. However, in this specific translation, the indirect object is not made explicit, so the source word is not aligned.

(4) a. Sometimes she asks me why I used to call her father Harold .
0 1 2 3 4 5 6 7 8 9 10 11 12

b. Soms vraagt ze waarom ik haar vader Harold noemde .
Sometimes asks she why I her father Harold called .
0 1 2 3 4 5 6 7 8 9

c. 0-0 1-2 2-1 4-3 5-4 6-8 7-8 8-8 9-5 10-6 11-7 12-9

This approach is word-based, but as discussed in Sect. 2.2, an alternative option is to encode the aligned order of the source and target sentences with aligned word *groups* or *sequences*. For that reason, Vanroy et al. (2019b) suggested to group consecutive tokens that are word aligned to consecutive target tokens together to form a sequential cross metric (`seq_cross`). These sequences should be as large as possible while also adhering to the following constraints (Vanroy et al. 2019b, 104):

– each word in the source sequence (group) is aligned to at least one word in the target sequence and vice versa;

[2] We will not go into that version of Cross here but rather focus on our own implementations. See the original work for more details and Carl et al. (2019) for an analysis.

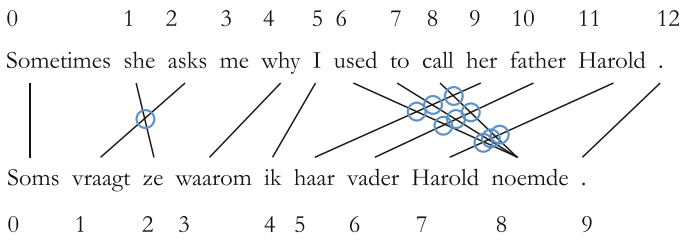


Fig. 2 Visualisation of `cross` in Ex. 4 with a `word_seq` value of $^{10}/_{12} = 0.83$ (modified from Vanroy et al. 2019b)

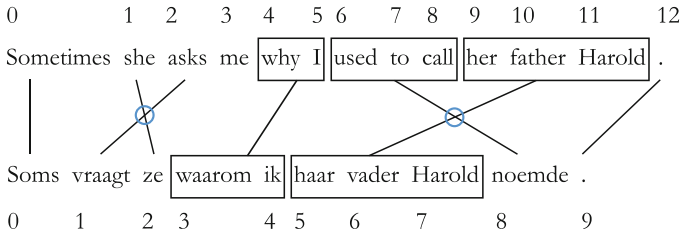


Fig. 3 Example of `seq_cross` in Ex. 4 with a total value of $^{2}/_{7} = 0.286$ (modified from Vanroy et al. 2019b)

– each word in the source word sequence is only aligned to word(s) in the aligned target word sequence (and not to words in other target sequences) and vice versa;
– none of the alignments between the source and target word sequences cross each other.

Similar to `word_cross`, normalisation takes place based on the number of alignments, only here it uses the alignments between the sequences rather than the word alignments. Following these requirements, the example in Fig. 2 can be modified so that instead of word movement, group movement is quantified (Fig. 3).

The problem with `seq_cross` is that, even though the metric works on the sequence level rather than the word level, its groups are linguistically arbitrary. Words are grouped together based on their relative reordering but irrespective of their linguistic properties (e.g. 'why I' and 'waarom ik' in the above examples). The need for grouping words founded on linguistic motivation gave rise to the current research. This specific issue involving word reordering is addressed in Sect. 3.2.

Motivated by the findings in previous studies, the main goal of this study is to introduce linguistically motivated, automatic, language-independent metrics to measure syntactic equivalence between source and target sentences in the context of translation.

3 Metrics

As discussed in Sect. 1, we restrict ourselves to three sub-components of syntactic equivalence,[3] namely word (group) order differences, changes in the dependency labels and structural differences with respect to the source and target dependency trees. To address these three individual differences, we introduce three corresponding metrics. First, we build on `seq_cross` and propose an improved version to quantify reordering of syntactic word groups (syntactically aware cross [SACr]; Sect. 3.2), then we discuss how label changes play a role (Sect. 3.3) and finally we introduce a method to calculate aligned syntactic tree edit distance (ASTrED; Sect. 3.4). A concise overview table of the metrics is given in Sect. 3.5. As all three metrics are based on comparing the syntactic structures of the source and target sentences using dependency representations, we start by explaining the chosen paradigm, Universal Dependencies, in closer detail.

3.1 Universal Dependencies

In all the metrics that we propose, we make use of UD annotation schemes (Nivre et al. 2016), which ensures comparable annotations across languages (see Sect. 2), such as the dependency labels of an English source text and its Dutch translation. To illustrate: the dependency trees of the source and target sentences of Example 4 are visualised in Figs. 4[4] and 5. In both figures, the nodes' labels are formatted as `word_index:dependency_label:token`. As can be seen, the dependency labels of both trees use the same scheme, which allows for straightforward comparison between the source and target trees without the need to convert one tagset into another. That would not be feasible if the source and target sentences were using different, language-specific annotation schemes.

To automate the parsing process, we depend on the recently introduced state-of-the-art `stanza` parser by the Stanford NLP group (Qi et al. 2020). In its annotation scheme, UD allows for language-specific extensions to the dependency relations to capture intricate properties of specific languages that may not generalise well to others languages. These extensions are also called *subtypes* because they always extend an existing UD dependency label. To minimise the effect of small

[3] An open-source implementation of our metrics is available at https://github.com/BramVanroy/astred.

[4] Note that dependency trees are different from phrase-based trees. For a more theoretical deep dive into the theory behind UD, we direct the reader to the work on Universal Dependencies (Nivre and Megyesi 2007; Nivre 2015; Nivre et al. 2016). Readers who are familiar with different dependency grammars may still disagree with the proposed trees, which may be due to the differences between UD and other grammars. For a critical comparison between UD and its alternatives, see Osborne and Gerdes (2019).

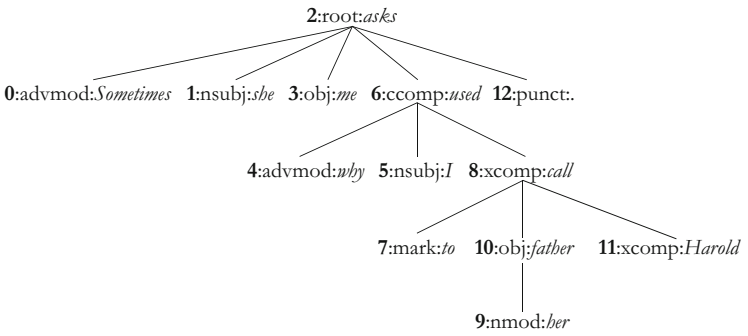


Fig. 4 Source dependency tree of Ex. 4: 'Sometimes she asks me why I used to call her father Harold '

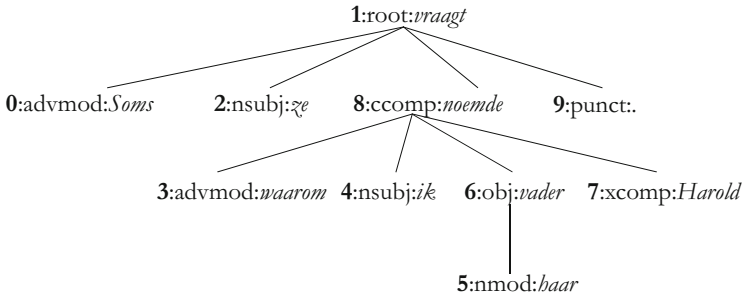


Fig. 5 Target dependency tree of Ex. 4: 'Soms vraagt ze waarom ik haar vader Harold noemde '

language or model-specific differences, we take a general approach and discard these UD subtypes, so a label such as `obl:tmod` (an oblique, nominal and temporal argument) will be reduced to `obl`.

3.2 Syntactically Aware Cross

In Sect. 2, we referred to `seq_cross`, in which reordering is quantified based on word sequences, i.e. consecutive words that are grouped together when they adhere to given constraints, also called *sequences*. Syntactically aware cross (SACr) expands on `seq_cross` by verifying that the words in generated `seq_cross` groups are linguistically motivated. Figure 6 shows an example of what we are trying to achieve. In this figure, the sequences as defined in `seq_cross` are shown as dotted boxes. In SACr, we verify whether these sequences are valid, linguistically motivated groups, and if this is not the case, we split the sequences up into smaller groups. The solid-line boxes in the figure represent those newly created, linguistically motivated groups. These groups (the initial `seq_cross` groups that were found to be valid SACr groups and the new SACr groups that

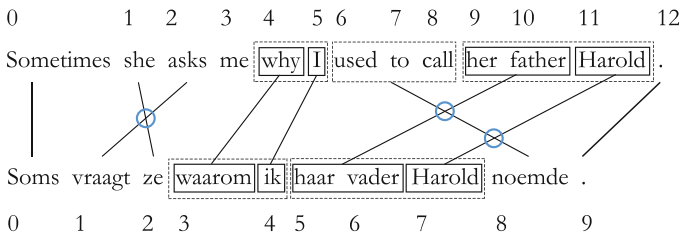


Fig. 6 Example of SACr with a total value of $3/9 = 0.33$. Dotted boxes indicate the initial groups of `seq_cross`. When required, these groups are split up into linguistically motivated SACr groups (solid boxes)

were created as a consequence of invalid `seq_cross` groups) are then used to calculate a syntactically aware cross value. Note that in this example, the number of crossing sequences has increased compared to the previous `seq_cross` value, as the sequence 'Her father Harold' is now split up into two groups 'Her father' and 'Harold'.[5]

The criterion for SACr to establish linguistically inspired word groups is that, in addition to the criteria of `seq_cross`, all words in a group need to be 'connected' to one another in the dependency tree: all nodes must exhibit one or more child–parent relationships with other nodes in the group. In practice, this means that siblings of a linguistic subtree can only be part of the same group if their parent is also in the group. More formally, we verify in a bottom-up, breadth-first fashion for each word that its parent in the dependency tree is also part of the same sequence group. The topmost node is excluded from the search because it cannot have a parent in this group. If all words in the group do not exhibit a child–parent relationship, the initial sequence group is not a valid SACr group. In such an event, in an iterative manner, a smaller subgroup of the initial sequence group is tested until a group is found for which the criterion above holds. We probe the largest subgroups first, and if no satisfying groups are obtained, smaller ones are tested (ultimately to the smallest size of two words) until no more groups can be found. This can mean that, for example, in an initial sequence group of four words, only a valid subgroup of two words is found. As a consequence, the other two words will both be singletons (separate SACr groups consisting of only one word each).

Figures 7 and 8 illustrate which of the proposed sequence groups (cf. dotted boxes in Fig. 3) are valid SACr groups in the dependency trees: when all items in a `seq_cross` group show a child–parent relation with other nodes in the group, the group is valid, but if not, new SACr subgroups will be created (e.g. 'haar vader Harold' is an invalid group, but 'haar vader' is a valid subgroup). In the following examples, square-cornered, blue groups are initial `seq_cross` groups that are also

[5]The sentence is ambiguous: 'her father Harold' *could* be interpreted as a single phrase ('…her father, who is named Harold'), but here we assume that the correct meaning of the sentence is '…call her father (by the name) Harold'.

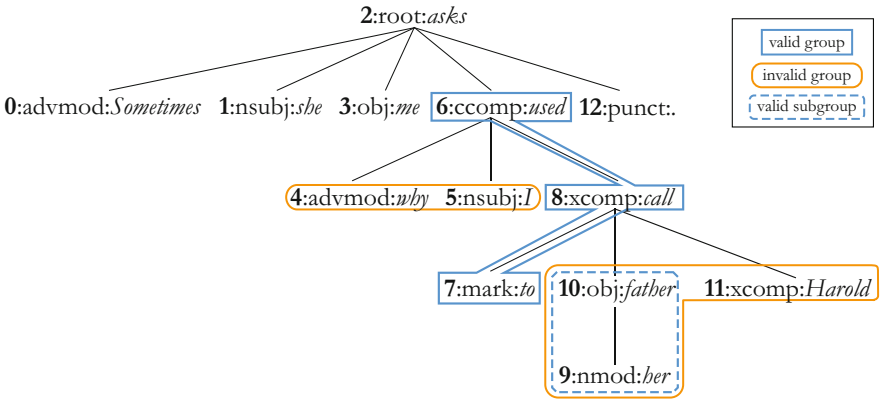


Fig. 7 Source dependency tree of Ex. 4 with highlighted groups: 'Sometimes she asks me why I used to call her father Harold'

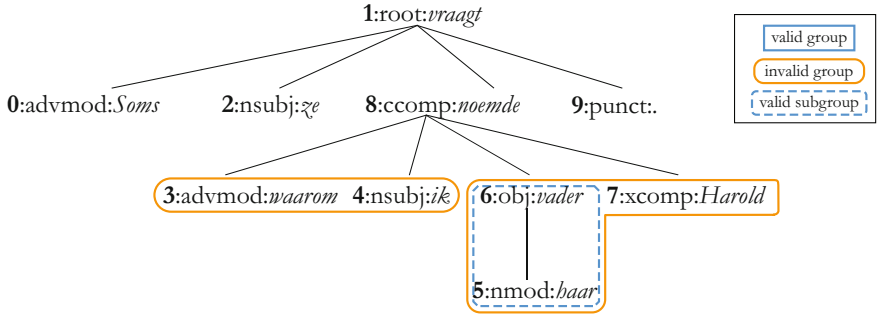


Fig. 8 Target dependency tree of Ex. 4 with highlighted groups: 'Soms vraagt ze waarom ik haar vader Harold noemde'

valid SACr groups. Round-cornered orange groups are initial `seq_cross` groups that are invalid SACr groups. Round-cornered blue and dashed groups are new SACr groups that are subgroups of invalid `seq_cross` groups.

Figure 6 above shows how the sequences from `seq_cross` have been adjusted according to the linguistic criteria derived from the dependency trees. This process can only increase the number of groups, not decrease them. In this particular case, the groups 'why I' and 'waarom ik' are split into two groups again, namely 'why' ('waarom') and 'I' ('ik') because these words are not connected to each other in the dependency tree. In both the source and target trees, the adverb and pronoun are siblings, but their root is not included in the group, causing them to not form a fully connected group. The group 'used to call' remains unchanged because all words are connected in the source dependency tree. The corresponding groups 'her father Harold' and 'haar vader Harold' are also split up, because in the dependency tree 'Harold' is not connected to 'her father'/'haar vader'. 'her father' and 'haar vader' are valid subgroups, though.

The final SACr value is the number of crossing alignment links between the source and target SACr groups, normalised by the number of these alignments. The example in Fig. 6 counts three crossing links and nine total alignment links, leading to a SACr value of $3/9 = 0.33$. This contrasts with the word-based `word_cross` value of the same example, which is $10/12 = 0.83$, and the `seq_cross` value of $2/7 = 0.29$ (cf. Sect. 2.3).

3.2.1 Cross Summary

The main distinction between our three proposed cross metrics (`word_cross`, `seq_cross` and SACr) is the size of the unit they use to calculate crossing links with. In `word_cross`, the reordering of single words is quantified. Alternatively, reordering can be counted when using sequences of words as alignment points by using `seq_cross`. Here, consecutive words are grouped together following given criteria so that crossing links can be counted on aligned groups of words rather than individual words. However, these groups are not linguistically motivated. To ensure that the word groups are linguistically motivated, SACr provides a linguistic correction of the groups of `seq_cross`. An initial group of `seq_cross` is maintained if it is linguistically valid according to our criteria (each item in a group must express a child–parent relationship to another item in the group). If it is not valid, new SACr subgroups are created inside that invalid group. This means that a sentence can have the same number of `seq_cross` and SACr groups or more SACr groups than `seq_cross` but never less.

Whereas SACr provides a way to quantify the reordering of phrase-like structures of a translation compared to its source text, counting the changes of the dependency labels of a source sentence after translation sheds light on linguistic differences of aligned words on the surface level.

3.3 Label Changes

An intuitive solution to syntactic equivalence is to assess how the dependency labels of translated words change from their aligned source text labels. To do so, we can simply count the alignment pairs where the source and target labels of an aligned word pair differ.

Formally, given a collection A of pairs of aligned source and target labels between a source sentence and its translation, the total number of label changes L is calculated as the number of alignment pairs in which the source label src is different from the target label tgt (Eq. 1).[6]

[6]Note that if a label, on either the source or target side, is aligned with multiple labels (one-to-many, many-to-one and many-to-many alignments), then all its alignments are counted separately.

$$L = \#\{(src, tgt) \in A : src \neq tgt\}, \tag{1}$$

where

A	=	the collection of pairs of aligned source and target labels,
src	=	the source label of a pair
tgt	=	the target label of a pair

For an illustrative example, consider the following active source sentence in Ex. 5a, which has been translated into a passive construction (Ex. 5b), and their word alignment (Ex. 5c).

(5) a. I saw him
`nsubj root obj`

b. Hij werd door mij gezien
He was by me seen
`nsubj aux case obl root`

c. 0-2 0-3 1-1 1-4 2-0

The word alignments can be visualised as in Fig. 9.

When counting the label changes, we look at each source word and compare its label to the labels of the words that it is aligned to. To exemplify this, consider the label changes of Ex. 5 in Table 1, leading to a total number of four label changes. These label changes are then normalised by the total number of alignments, leading to a value of $4/5 = 0.8$.

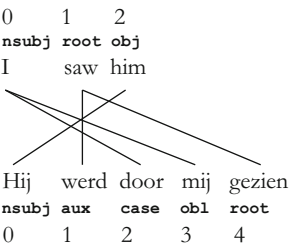


Fig. 9 Word alignment visualisation of Ex. 5

Table 1 Label changes for Ex. 5

Source (label)	Target (label)	Change
'I' (`nsubj`)	'door' (`case`)	1
'I' (`nsubj`)	'mij' (`obl`)	1
'saw' (`root`)	'werd' (`aux`)	1
'saw' (`root`)	'gezien' (`root`)	0
'him' (`obj`)	'Hij' (`nsubj`)	1
	Total: 4 (normalised: $4/5 = \mathbf{0.8}$)	

3.4 Aligned Syntactic Tree Edit Distance

Whereas SACr calculates a cross value on a shallow level (injected with a tree-based grouping) to quantify word order changes, it is also possible to determine deeper structural differences between the source and target sentences. To compare the actual source and target dependency *structures*, we propose ASTrED.

As the name implies, aligned syntactic tree edit distance (ASTrED) incorporates a source dependency tree and a target dependency tree with the word alignments between the source and target sentences. The goal is to modify the labels of the source and target dependency trees so that the labels of aligned words are identical. By doing so, we can ensure that the tree edit distance between these modified trees takes word alignment information into account.

Consider the example sentence and its translation in Ex. 6 and its word alignment (visualised in Fig. 10). This example will be used to explain ASTrED in the following subsections.

(6) a. Does he believe in love ?
`aux nsubj root case obl punct`

b. Gelooft hij in de liefde ?
Believes he in the love ?
`root nsubj case det obl punct`

c. 0-0 1-1 2-0 3-2 4-3 4-4 5-5

The metric can be summarised in the following steps, on which we elaborate in the next subsections.

1. Parse the source and target sentences into dependency trees (using UD labels).
2. Find grouped tokens between the source and target trees based on word alignment. A group is defined as the minimal group of tokens in the source and target sentences that are exclusively connected to each other through word alignment.
3. Modify the labels of the grouped tokens in their respective trees, so that the labels of tokens belonging to the same group get the same label. Nodes that were not aligned and thus do not belong to any group remain unchanged.
4. Calculate tree edit distance between the modified trees, which measures the structural difference between the aligned source and the target sentences. Normalise by the average number of source and target words.

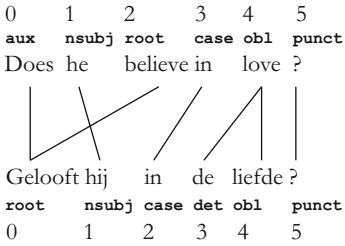


Fig. 10 Word alignment visualisation of Ex. 6

3.4.1 Constructing Dependency Trees

Identical to the previous metrics, we use dependency trees to represent the source and target sentences in a linguistically meaningful way (see Sect. 3.1). As an example, let us take the previously mentioned example, Ex. 6. The source and target sentences can each be represented as a dependency tree where each node is internally represented as the corresponding dependency label (Figs. 11 and 12).

3.4.2 Merge Grouped Tokens and Update Labels

In order to measure the structural difference between a source and target sentence, we use tree edit distance. The tree edit distance between two trees is the minimal number of operations that are needed to change one tree into the other. The three possible operations are deleting, inserting or substituting (also called 'renaming') a node in the tree.[7] We cannot simply take the edit distance between the source and target dependency trees, however, because that would disregard the word alignment information. Tree edit distance in itself is unaware of which source nodes are supposed to align with which target nodes. To be able to calculate alignment-aware tree edit distance (the distance between the source and target dependency structures while also taking word alignment information into account), we modify the source and target trees by merging their labels with respect to the word alignments. Unaligned words remain untouched. In practice, that means that all tokens that are connected to one another through word alignment are grouped together. Here, they are represented (serialised) as a mapping of source label(s) to target label(s), where source labels are separated by a pipe (|) and their corresponding target labels by a comma.

More specifically, if we consider the example in 6, we can distinguish five groups (Example 7), where the corresponding words are given between brackets:

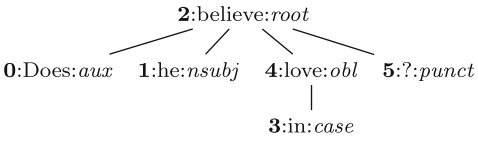


Fig. 11 Source dependency tree of Ex. 6a: 'Does he believe in love ?'

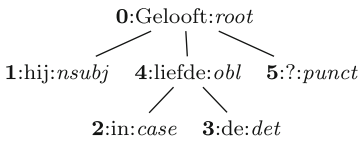


Fig. 12 Target dependency tree of Ex. 6b: 'Gelooft hij in de liefde?'

[7] To automate the tree edit distance calculation, we use a Python implementation (https://github.com/JoaoFelipe/apted) of the APTED algorithm (Pawlik and Augsten 2015, 2016).

(7) – `aux:root|root:root` (does:gelooft|believe:gelooft)
– `nsubj:nsubj` (he:hij)
– `case:case` (in:in)
– `obl:det,obl` (love:de,liefde)
– `punct:punct` (?:?)

3.4.3 Modify Dependency Trees

For all items involved in a group, their respective labels in their respective trees are updated to the serialised group. This implies that the nodes in the source and target trees that are aligned now have the same label. This is important, because the goal is to calculate tree edit distance on the *aligned* source and target trees.

The trees with modified labels are shown in Figs. 13 and 14 with a word's original position (index) placed before the serialised label. Note how the labels are now modified so that aligned nodes share the same label. Also, consider that if, for instance, two source nodes are aligned with one target node, then all three will share the same modified label, such as the label `aux:root|root:root`, which is the alignment of 'does . . . believe' to 'Gelooft'.

3.4.4 Calculate Tree Edit Distance

Finally, we calculate the tree edit distance between the modified trees shown above. To change the modified source tree in Fig. 13 to the modified target tree in Fig. 14, two operations are needed, as visualised in Fig. 15:

1. the source node `aux:root|root:root` (orange, solid line) must be deleted;
2. the target node `obl:det,obl` (blue, dashed line) must be inserted.

The ASTrED score is normalised by the average number of source and target words. This is different from the way that SACr and the label changes are

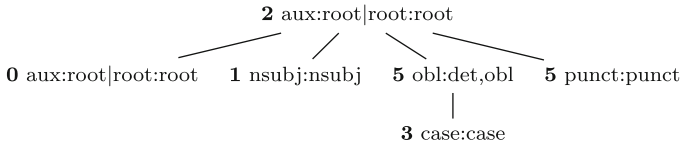
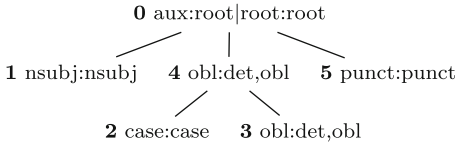


Fig. 13 Modified source dependency tree of Example 6a: 'Does he believe in love ?'

Fig. 14 Modified target dependency tree of Example 6b: 'Gelooft hij in de liefde ?'



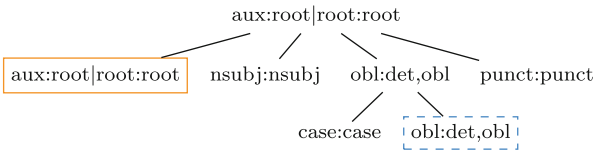


Fig. 15 A visualisation of the two needed edits to go from modified source tree in Fig. 13 to the modified target tree in Fig. 14. The orange solid box indicates the source node that needs to be deleted, and the dashed blue box highlights the target node that needs to be inserted

normalised: SACr is normalised by the number of alignment links between SACr groups because the crossing links originate from those alignments. Label changes are normalised by the number of word alignment link, because the differences in labels are calculated between aligned labels. ASTrED is calculated between tree representations of the source and target sentences, which means that each word's label in the source or target text is a node in the dependency tree. In other words, ASTrED takes unaligned words (null alignment) into account (see Sect. 4.2 for an example), whereas SACr and label changes only consider the alignments themselves. Therefore, ASTrED is normalised by the average number of source and target words. Applying that to this example, with a source sentence of six words and a target sentence of six words, we get an ASTrED score of $2/6 = 0.33$.

To reiterate: we calculate tree edit distance on the modified trees where node labels are replaced by a serialised representation of the aligned source and target nodes. This is done to ensure that tree edit distance takes word alignment information into account (Table 2).

3.5 Metrics Overview

Table 2 summarises the proposed metrics and how they are normalized on the sentence level.

Table 2 An overview of the metrics introduced in this chapter

Metric	Captures	Normalisation by
Label changes	Changes in dependency labels in the surface form based on word alignment	Number of alignments
SACr	Reordering of linguistically motivated groups by measuring crossing links	Number of alignments
ASTrED	Structural difference between the source and target dependency trees while also taking word alignment into account	Average number of source and target words

4 Discussion with Examples

As discussed before, syntactic equivalence is an ill-defined concept because it entails different linguistic aspects: from word reordering at the surface level to deep structural differences. For that reason, we proposed three linguistically motivated metrics (which can be used and calculated independently) that all tackle a different part of the problem. In this section, we will discuss further what the differences between the metrics are by going over two examples that illustrate other typical linguistic differences between English and Dutch, in addition to the previously given examples (active–passive, indirect speech, English *do*). In the following two examples, we discuss subject–verb word order and the future tense and the translation of the English gerund to Dutch and null alignments.

4.1 Subject–Verb Word Order and the Future Tense

English is typically classified as a language with subject–verb–object (SVO) word order, but there is no consensus on Dutch. One approach suggests that Dutch uses the subject–object–verb (SOV) with V2 (verb second) word order (Koster 1975), where in the main clause, the finite verb must be placed second with one constituent preceding it, and where subordinate clauses adhere to the SOV word order. Alternatively, Zwart (1994) suggested that Dutch is SVO, by dissecting the verb phrase (VP) structure of a subordinate clause in detail.

Even though that discussion exceeds the scope of this chapter, the practical implication is that in many cases (e.g. topicalisation, left dislocation and subordinate clauses), the word order of English and Dutch differs.

Consider Ex. 8 where the word order of the main verb and the subject differs between Dutch and English because of the dislocated adverb, which leads to inversion in Dutch. The example also shows how the simple future tense can be presented in the present tense in Dutch, which leads to the source auxiliary 'will' and its root 'go' to be aligned with the present tense root 'ga'.

(8) a. Tomorrow I will go home .
`advmod nsubj aux root obj punct`

b. Morgen ga ik naar huis .
Tomorrow go I to home .
`advmod root nsubj case obl punct`

c. 0-0 1-2 2-1 3-1 4-3 4-4 5-5

The alignments and word crosses can be visualised as follows in Fig. 16. The `word_cross` value is $2/7 = 0.29$.

Vanroy et al. (2019b) suggested a sequential approach to word reordering where consecutive words are grouped together following a given set of criteria (cf. Sect. 2.3). In the example above, this can be visualised as in Fig. 17, showing a `seq_cross` value of $1/4 = 0.25$.

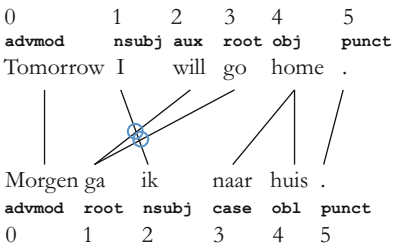


Fig. 16 Visualisation of word alignment of Ex. 8 and a `word_cross` value of $2/7 = 0.29$

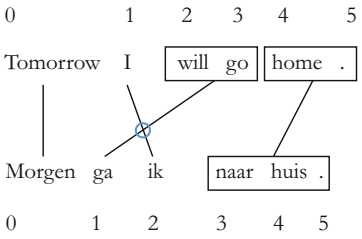


Fig. 17 `seq_cross` representation of Ex. 8 with a value of $1/4 = 0.25$

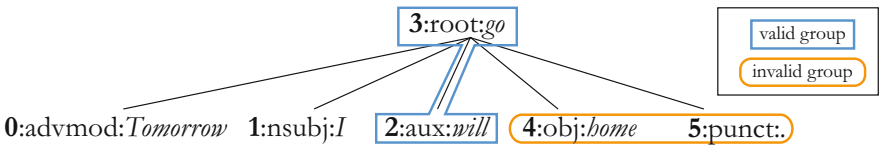


Fig. 18 Source dependency tree of Ex. 8, highlighting valid and invalid groups

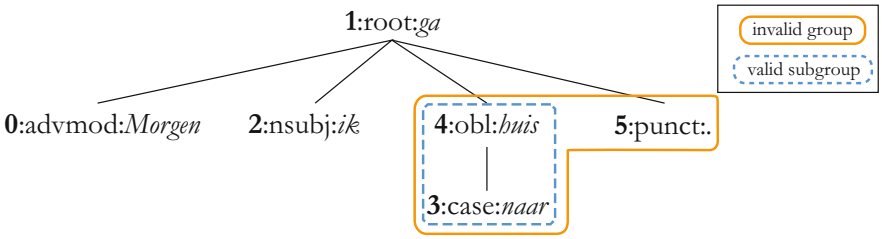


Fig. 19 Target dependency tree of Ex. 8, highlighting an invalid group and a valid SACr subgroup

In this chapter, we have proposed an improved version of `seq_cross` named SACr. Whereas `seq_cross` is not aware of linguistic information and naively groups word sequences together, SACr ensures that these groups are linguistically motivated: all items in a SACr group must exhibit a child–parent relationship to at least one word in the group. The valid and invalid groups are shown for both the source and target dependency trees in Figs. 18 and 19.

The initial groups of `seq_cross` are not linguistically motivated, but by means of the dependency trees, we can correct these groups to ensure that all groups are indeed linguistically valid. The alignment between these groups can be used

to quantify the reordering of syntactic word groups. In this example, there is one crossing link that is then normalised by the total number of alignments (five). The SACr value, then, is $1/5 = 0.2$.

In addition to word reordering, the label changes are indicative of diverging linguistic properties. Looking at the label changes going from the source to the target sentence in Fig. 16, we find three alignments where the labels of the source word have changed (Table 3), which when normalised gives a value of $3/6 = \mathbf{0.5}$.

With ASTrED, we also provide a means to compare the underlying structure of aligned dependency trees. This is done by grouping aligned words together in the source and target trees, changing their labels according to this grouping in both trees and calculating tree edit distance between the modified trees. In Ex. 8, we can distinguish five groups (Ex. 9).

(9)
- `advmod:advmod` (Tomorrow:Morgen)
- `nsubj:nsubj` (I:ik)
- `aux:root|root:root` (will:galgo:ga)
- `obj:case,obl` (home:naar,huis)
- `punct:punct` (.:.)

We can then modify the original dependency trees (see Figs. 18 and 19) by changing the label of each node to the serialised group that it belongs to. The modified trees are given in Figs. 21 and 22:

These modified trees can then finally be used to calculate tree edit distance. Figure 23 shows the two edit operations that are needed to change the modified source tree to the modified target tree. This value is normalised with the average number of source (six) and target words (six), which leads to an ASTrED score of $2/6 = 0.33$.

Table 3 Label changes for Ex. 8

Source (label)	Target (label)	Change
'Tomorrow' (`advmod`)	'Morgen' (`advmod`)	0
'will' (`aux`)	'ga' (`root`)	1
'go' (`root`)	'ga' (`root`)	0
'home' (`obj`)	'naar' (`case`)	1
'home' (`obj`)	'huis' (`obl`)	1
'.' (`punct`)	'.' (`punct`)	0
Total: 3 (normalised: $3/6 = \mathbf{0.5}$)		

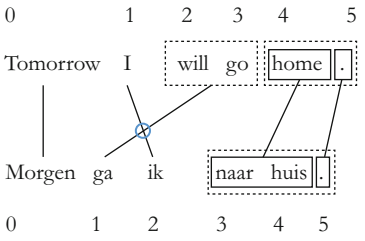


Fig. 20 SACr representation of Ex. 8 with a value of $1/5 = 0.2$. Dotted boxes indicate the groups of `seq_cross`, which, when required, are split up into linguistically motivated SACr groups (solid boxes)

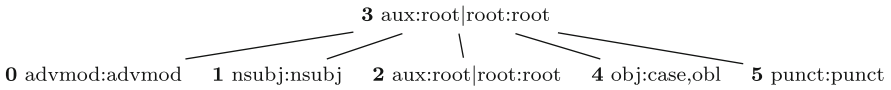


Fig. 21 Modified source dependency tree of Ex. 8: 'Tomorrow I will go home'

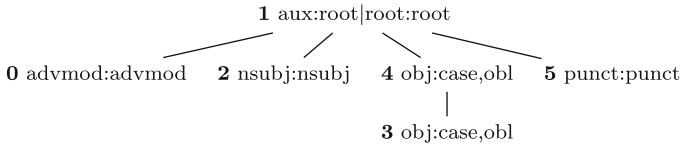


Fig. 22 Modified target dependency tree of Ex. 8: 'Morgen ga ik naar huis'

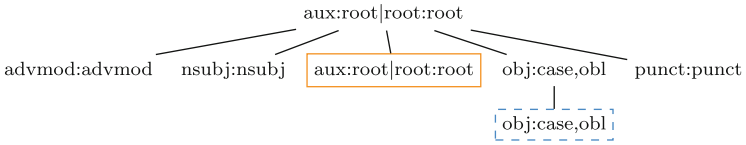


Fig. 23 A visualisation of the two needed edits to go from the modified source tree in Fig. 21 to the modified target tree in Fig. 22. The orange solid box indicates the source node that needs to be deleted, and the dashed blue box highlights the target node that needs to be inserted

In this example, which involves a different subject–verb order in English and Dutch, SACr clearly models how the word order of the verb with respect to the subject has changed (Fig. 20). Label changes, on the other hand, do not catch the word group reordering aspect because they solely compare aligned words, disregarding their position relative to each other. In this example, it does catch how the auxiliary verb 'will' has a different label than the present tense of its Dutch translation 'ga' (root). It also finds that whereas English allows for a 'go `obj`' construction, Dutch requires a case marker in such case, in the form of 'ga `case obl`'.

The edit operations of ASTrED (e.g. Fig. 23) highlight that tree edit distance does not account for word reordering in some cases. That is due to the nature of dependency trees: even though our implementation of a dependency tree ensures that the order of *sibling* nodes is identical to their word order, there is no way in the tree to know the word order position of a parent node vis-à-vis its children. So, two tree structures may be identical, but the word order of a parent node with respect to its descendants can still differ. In this case, the subtree structure of the subjects ('I' and 'ik') and their main verb ('go' and 'ga') are identical (it is a child–parent relationship), so the tree edit distance for that subtree is 0, even though the word order of the source and target sentences is different: in the English sentence, the subject precedes the verb, whereas in the Dutch translation, the verb comes first. That order difference is not visible in the trees. As such, it is clear that the reordering metrics capture different information than ASTrED. In this case, ASTrED catches the same differences that the label changes find, concerning the future tense that is translated as a present tense, and the English object following 'go' that needs to be case-marked in Dutch. As a consequence, the node of the future auxiliary verb

Table 4 Summary of the results of all metrics for Ex. 8 (rounded to two decimals)

word_cross	0.29
seq_cross	0.25
SACr	0.2
Label changes	0.5
ASTrED	0.34

(aux:root | root:root) needs to be removed from the English source, and the case marker of the Dutch translation must be added (obj:case,obl), to arrive at the same tree structure (see Fig. 23). The results of all metrics for this example are summarised in Table 4.

4.2 English Gerund, Verb Order and Null Alignment

In English, gerunds are verb forms that typically end with -ing and that most often take a nominal function. In Dutch, however, this construction is frequently translated as an infinitive, but just as often a complete rewrite of the original constituent seems appropriate. In the following example, an English gerund ('Shouting') is translated as an infinitive ('roepen'). Both their dependency relations to their root are csubj, meaning that they are clausal subjects, i.e. they are the subject of a clause and they are themselves a clause. Similar to the previous example, the word order of the object ('for help' and 'om hulp') with respect to its verb ('Shouting' and 'roepen') is a noteworthy difference in the source and target sentences. Finally, in this example, 'seemed' is translated by adding a pronoun as an object[8] to the verb 'leek' *seemed*, namely 'mij' *to me*. Because of this explicitation, 'mij' cannot be aligned with a source word.

(10) a. Shouting for help seemed appropriate .
csubj case obl root xcomp punct

b. Om hulp roepen leek mij gepast .
For help call seemed me appropriate .
case obl csubj root obj xcomp punct

c. 0-2 1-0 2-1 3-3 4-5 5-6

The alignments in Example 10c can be visualised in Fig. 24, which also shows the crossing links on the word level. In this case, there are two crossing links that indicate the different word order of objects relative to their verb in English compared to Dutch, as discussed before. After normalisation, the word_cross value is $^2/_6 = 0.33$.

[8]Following the conventions of UD, we label 'mij' as an obj. The annotation guidelines suggest that when a verb has only one object, it should be labelled as an obj and not an iobj, regardless of the morphological case or semantic role of that word. (See https://universaldependencies.org/u/dep/iobj.html).

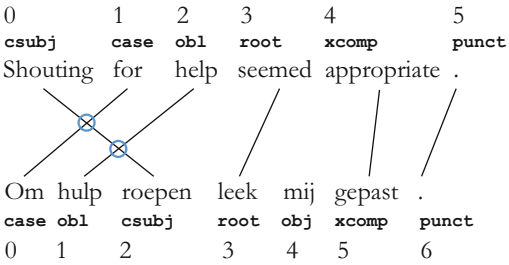


Fig. 24 Visualisation of word alignment in Ex. 10 and a `word_cross` value of $2/6 = 0.33$

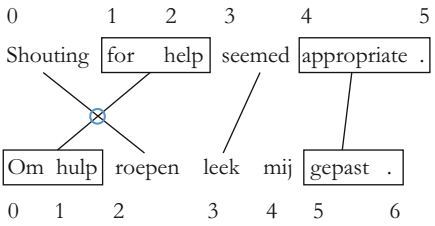


Fig. 25 `seq_cross` representation of Ex. 10 with a value of $1/4 = 0.25$

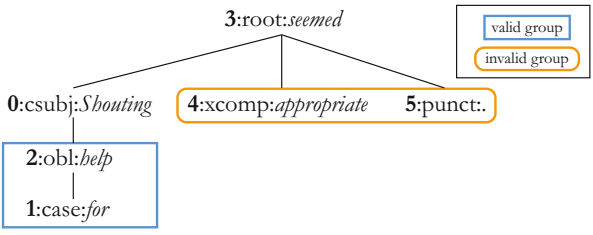


Fig. 26 Source dependency tree of Ex. 10, highlighting an invalid group and a valid SACr subgroup

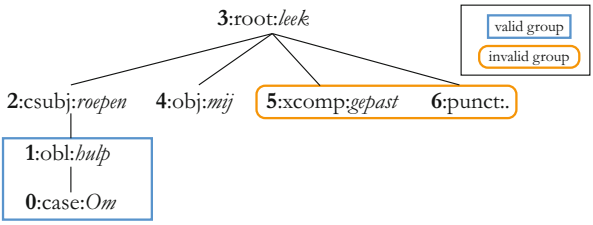


Fig. 27 Target dependency tree of Ex. 10, highlighting an invalid group and a valid SACr subgroup

When grouping consecutive words, as discussed in Sect. 2.3, we find that 'for help' and 'Om hulp' each constitute a group, as well as 'appropriate .' and 'gepast .'. This is visualised in Fig. 25. Grouping 'for help' and 'Om hulp' leads to a reduction in crossing links: now, there is only one crossing. The `seq_cross` value is $1/4 = 0.25$.

However, as discussed in Sect. 3.2, the groups of `seq_cross` are not linguistically motivated. To create groups that take the linguistic structure into account, we verify that all items in a group share a child–parent relationship with another word in that group. For this example, we can investigate the source and target dependency trees in Figs. 26 and 27, respectively.

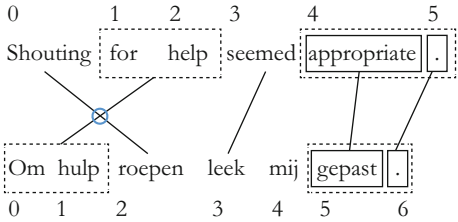


Fig. 28 SACr representation of Ex. 10 with a value of $1/5 = 0.2$. Dotted boxes indicate the groups of `seq_cross`, which, when required, are split up into linguistically motivated SACr groups (solid boxes)

The visualisations of the dependency trees make clear that the groups 'for help' and 'Om hulp' are valid because the prepositions ('for' and 'om', respectively) are children of their root ('help' and 'hulp', respectively) and that child–parent relationships constitute a valid SACr group. The other groups 'appropriate .' and 'gepast .' are not valid because the two words in each group share a sibling relationship rather than a child–parent relationship, which is not sufficient to form a valid SACr group. These linguistically corrected groups have been visualised in Fig. 28. The number of crossing links is still one, but because the invalid groups are corrected ('appropriate .' and 'gepast .'), the normalised value has now changed from `seq_cross` 0.25 to SACr 0.2.

The label changes in this example are quite self-explanatory: looking at the word alignments in Fig. 24, it is evident that all the labels of aligned words are identical on the source and target sides. Therefore, there are zero label changes in this example. Nevertheless, that does not mean that there are no structural difference, as ASTrED will illustrate.

To calculate ASTrED, first, the labels of the source and target trees need to be grouped according to the word alignments. Each group should contain all the labels of words that are connected to one another through word alignment. In Example 11, we can find six groups and also one unaligned word ('mij' *me*).

(11)
- `csubj:csubj` (Shouting:roepen)
- `case:case` (for:Om)
- `obl:obl` (help:hulp)
- `root:root` (seemed:leek)
- `xcomp:xcomp` (appropriate:gepast)
- `punct:punct` (.:.)
- **null alignment** (in target): `obj` (mij)

As a next step, the labels of each node in a group must be updated to the serialised group's label. In this example, the groups always consist of only one source and one target item. The unaligned `obj` node in the target sentence is still present after changing the labels (Figs. 29 and 30).

Now, the tree edit distance between these modified trees can be calculated. The structure of the source sentence is in fact exactly the same as the one in the target sentence, with the exception of one unaligned `obj` node ('mij'). The only operation that is needed to change the source structure to the target structure is inserting the unaligned target node (Fig. 31). This illustrates that ASTrED is the only one of the

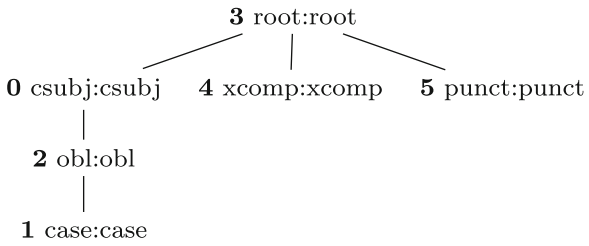


Fig. 29 Modified source dependency tree of Ex. 10: 'Shouting for help seemed appropriate'

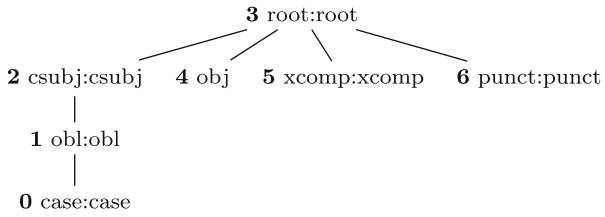


Fig. 30 Modified target dependency tree of Ex. 10: 'Om hulp roepen leek me gepast.' Note the unaligned `obj` node

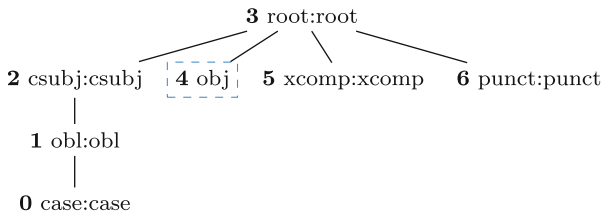


Fig. 31 A visualisation of the edit (insertion, the dashed blue box) to go from the modified source tree in Fig. 29 to the modified target tree in Fig. 30

tree metrics that is able to take into account null alignments. The edit operations are normalised by the average number of source (6) and target (7) tokens, so the ASTrED value is $1/6.5 = 0.15$.

In this example, it became clear how SACr again accurately quantifies the reordering of linguistically motivated word groups. In particular, it showed how the subject–verb order of English and Dutch can be quantified with a single crossing link because of the syntactically aware word grouping of 'for help' and 'Om hulp'. Because the examples were quite closely related in this example, we did not observe any label changes. However, on a deeper structural level, we found that the structure of both sentences does differ slightly because of a null alignment on the target side: 'mij' *me* was inserted in the translation even though there is no source word to align it with. The results are summarised in Table 5.

Generally speaking, the three metrics model three different things: SACr specifically quantifies the reordering of linguistically inspired word groups. When the surface word order of languages differs in specific structures, SACr catches up on

Table 5 Summary of the results of all metrics for Ex. 10 (rounded to two decimals)

`word_cross`	0.34
`seq_cross`	0.25
SACr	0.2
Label changes	0.0
ASTrED	0.15

that. This is particularly evident in Example 6 where a different word order is found twice in the same sentence ('Sometimes **she asks** me why I **used to call her father Harold** .' vs. 'Soms **vraagt ze** waarom ik **haar vader Harold noemde** .'). Also, based on the surface forms, label changes compare the labels of the aligned words on the source and target sides. By doing so, it can quickly become evident when a source sentence and its translation have been translated completely differently (think, for instance, about the active–passive example in Example 5 where a `nsubj` became an `obj`). ASTrED serves a similar function, but it compares the actual tree structures of the source and target sentences while at the same time also taking the word alignments into account. Whereas SACr and label changes work on the surface forms, ASTrED does a deeper linguistic comparison between a source sentence and its translation, as the last example clearly shows.

5 Proof of Concept

To investigate how syntactic differences between a source text and its translation relate to difficulty, we can measure the effect that our syntactic measures have on translation process features that may be indicative of cognitive effort, which in turn points to translation difficulty (also see our previous research for details and a literature overview concerning cognitive effort and translation; Vanroy et al. 2019a).[9] We built mixed-effect models in R (R Core Team 2019), using the lme4 package (Bates et al. 2015) with lmerTest (Kuznetsova et al. 2017) to obtain p-values and perform automatic backward elimination of effects.

We used part of the ROBOT dataset (Daems 2016) for this analysis. The full ROBOT dataset contains translation process data of ten student translators and twelve professional translators working from English into Dutch. Each participant translated eight texts, four by means of post-editing (starting from MT output) and four as a human translation task (starting from scratch). Task and text order effects were reduced by using a balanced Latin square design. The texts were newspaper articles of 150–160 words in length, with an average sentence length between

[9]Other chapters in this volume also discuss new advances in cognitive effort research. See, for instance, the work by Huang and Carl in Chapter 2 and Chapter 3 by Cumbreño and Aranberri regarding cognitive effort during post-editing, and Lacruz et al. on cognitive effort in JA-EN and JA-ES translation (Chapter 11).

15 and 20 words. As the goal of the original ROBOT study was to compare the differences between post-editing and manual translation, the texts were selected to be as comparable to one another as possible, based on complexity and readability scores, word frequency, number of proper nouns and MT quality. For the present study, however, only the process data for the human translation task was used. This dataset was manually sentence and word aligned. Dependency labelling was done automatically by using the aforementioned `stanza` parser (Qi et al. 2020).

We followed exclusion criteria suggested by Bangalore et al. (2015) before analysing our data: exclude cases where two ST (source text) segments were fused into one, exclude the first segment of each text, exclude segments with average normalised total reading time values below 200ms (total reading time; the time (in ms) that participants have their eyes fixated on the source or target side, measured by eye tracking) and exclude data points differing by 2.5 standard deviations or more from the mean. After filtering, the dataset consists of 537 data points, i.e. translated segments. All plots were made using the effects package (Fox and Weisberg 2019). In parallel with Bangalore et al. (2015), dependent variables from the TPR-DB (Carl et al. 2016) were chosen, specifically total reading time on the target (TrtT) and source (TrtS) side, and duration of coherent typing behaviour (total duration of coherent keyboard activity excluding keystroke pauses of more than 5s; Kdur), normalised by the number of words per segment and centred around the grand mean (hence the negative values in the graphs).[10] The predictor variables were our three proposed metrics: SACr, label changes and ASTrED. In the full model, all three variables were included with interaction. We performed backward elimination of effects to build the best model for each dependent variable. Participant codes and item codes were included as random effects.

For coherent typing behaviour (Kdur), the only predictor variable that was retained in the best performing model was the number of label changes. An increase in label changes had a highly significant ($p < 0.001$) positive effect on Kdur (estimate = 969.1, SE = 232 and $t = 4.18$). This effect can be seen in Fig. 32. This indicates that translators needed more time to translate those source segments that required more label changes when translating.

Source reading time (TrtS) was best predicted by SACr only, although the model that included both participants and items as random effects gave rise to convergence warnings. The main effect of SACr on TrtS was positive (estimate = 69.82, SE = 28.39 and $t = 2.46$) and significant ($p = 0.01$). The effect can be seen in Fig. 33. The model without participants as random effect did converge and showed a similar main effect (estimate = 95.11, SE = 33.85, $t = 2.81$ and $p = 0.005$). This means that those segments that were translated by moving more word groups or move word groups further away required more reading time on the source side.

[10]Even though our experimental set-up is similar, our results cannot be compared to those of Bangalore et al. (2015) because we use a different dataset and do not use entropy but absolute values per segment.

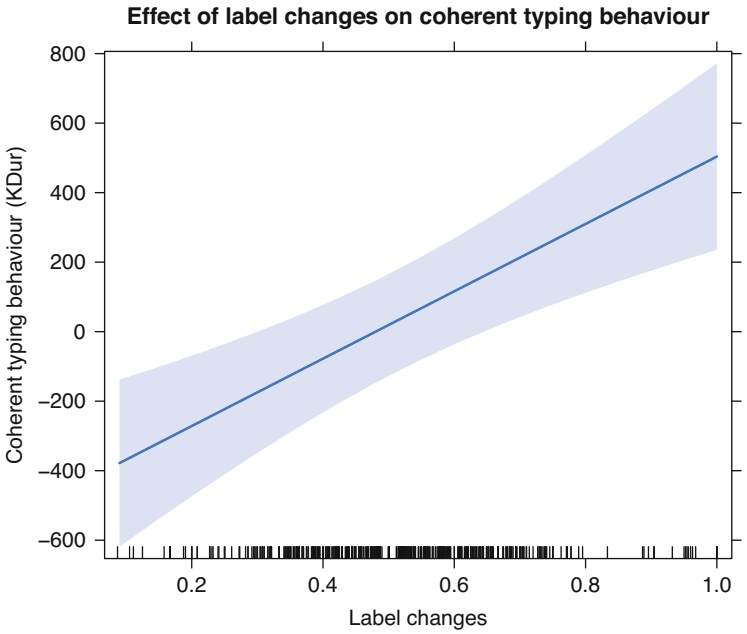


Fig. 32 Effect plot for the main effect of label changes on coherent typing behaviour

Target reading time (TrtT), on the other hand, was best predicted by a combination of all three predictor variables with interaction. The three-way interaction effect was significant (estimate = 3383.2, SE = 1173.6, $t = 2.88$ and $p = 0.004$). All effects included in the model are summarised in Table 6. The interaction effect is visualised in Fig. 34. The figure shows the effect of ASTrED values on target reading time, given a certain SACr value and the number of label changes. Only the minimum and maximum values of SACr and label changes are included as reference points (0 and 9.7 for SACr and 0.09 and 1 for label changes, respectively). What this indicates is that, if SACr is low, an increase of ASTrED or an increase in the number of label changes does not really have that much of an impact on target reading time. However, if SACr values are high and there is a low number of label changes, target reading time goes down for higher ASTrED values, whereas target reading time goes up for higher ASTrED values when SACr values are high and there are a high number of label changes. Looking at the graph on the right (high SACr value), it would seem that when a lot of word group reordering is required without many label changes (blue line with negative slope), structurally similar source and target sentences (low ASTrED) lead to a higher TrtT. Conversely, when a lot of word group reordering is needed alongside many label changes (orange line with positive slope), dissimilar syntactic structures (high ASTrED) positively affect the time that translators read the target text. This conclusion should be taken with a grain of salt,

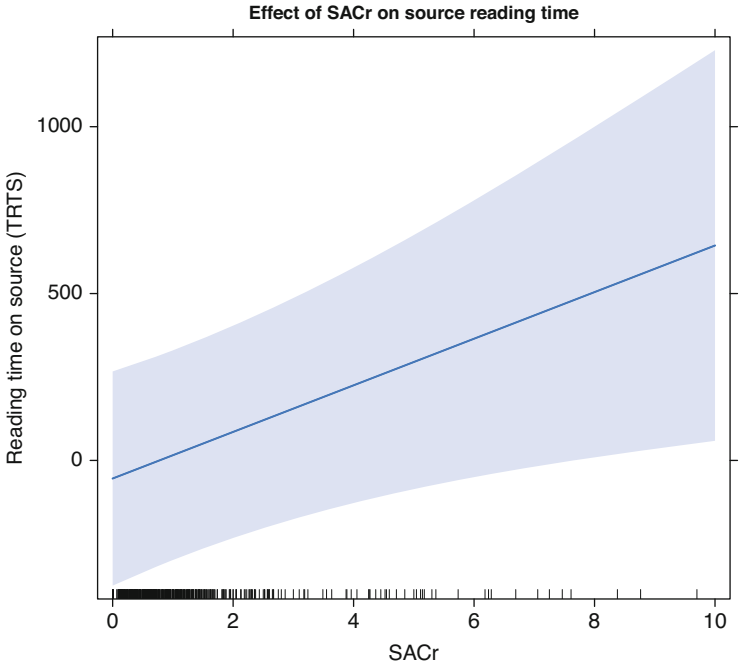


Fig. 33 Effect plot for the main effect of SACr on source text reading time

though, and additional experiments with other data sets are required to draw more certain conclusions.

Unsurprisingly, the metrics are only weakly to moderately correlated, as seen in Table 7. This is likely due to a single common factor of all metrics: they are, at their core, all based on the same dependency labels. Different dependency trees lead to different SACr groups, a change in the merged ASTrED trees, as well as the label changes themselves. However, because each metric uses the dependency labels in its own way, a change in dependency structures affect specific metrics differently. The metrics are therefore mildly correlated, but they have a different effect on the translation process, as shown above.

In this section, we have calculated the effect of our proposed syntactic metrics on translation process features to show that our interpretation of syntactic equivalence has an effect on the translation process. Even though our dataset was rather small, and more elaborate experiments are needed, these findings already confirm that, as the literature indicates (cf. Sect. 2), (syntactic) equivalence does affect some translation process features such as reading time and typing duration, which serve as a proxy for the translation difficulty. Generally speaking, this experiment arrives to the same conclusion as Bangalore et al. (2015), namely that syntactically diverging source and target segments impose difficulty on the translator. In addition, this

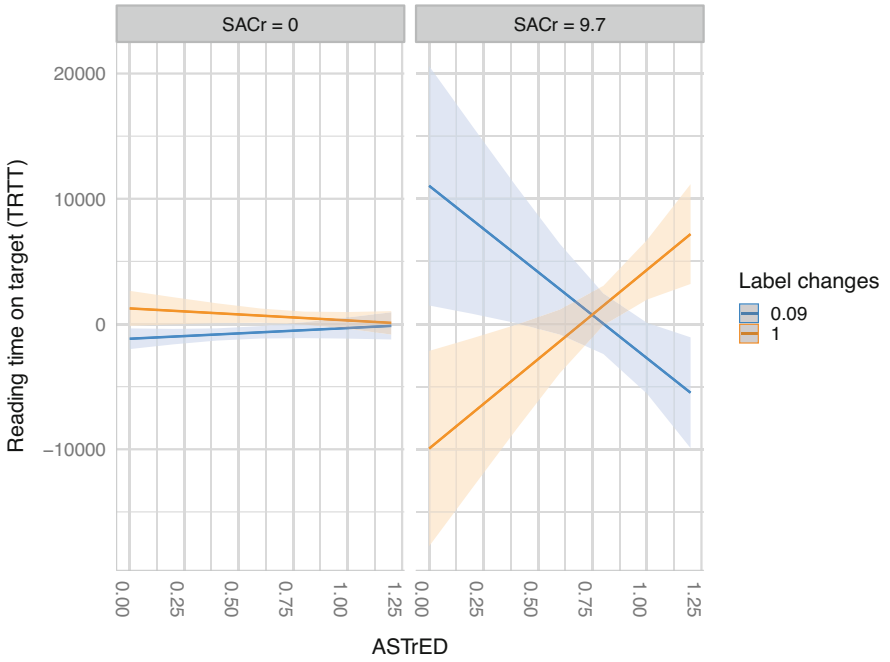


Fig. 34 Effect plot for the three-way interaction effect of ASTrED, label changes and SACr on target reading time

Table 6 Effect summary of three-way interaction effect between ASTrED, label changes and SACr on target reading time

Fixed effect	Estimate	SE	t	p
ASTrED	1034.4	819.1	1.26	.207
Label changes	2662.5	1103	2.41	.016 *
SACr	1498.3	602.3	2.49	.013 *
ASTrED:label changes	−1994.7	1514.1	−1.32	.188
ASTrED:SACr	−1812.6	692.3	−2.62	.009 **
Label changes:SACr	−2652.4	989.5	−2.68	.008 **
ASTrED:label changes:SACr	3383.2	1173.6	2.88	.004 **

$*p < .05$
$**p < .01$

Table 7 Kendall correlation between normalised metrics: ASTrED, label changes and SACr ($p < .01$)

	ASTrED	Label changes
ASTrED		
Label changes	.41	
SACr	.40	.35

experiment also confirms that all three metrics seem to affect the translation process differently, which motivates further research into this topic.

6 Conclusion and Future Work

In this work, we have introduced three new metrics to measure syntactic equivalence between a sentence and its translation. The three metrics serve different purposes, which is also revealed in Sect. 5. Keeping track of dependency label changes is an intuitive approach to see how the relation of each word to its root has changed in the translation. Syntactically aware cross (SACr) offers a linguistically motivated method to calculate word group reordering. Finally, aligned syntactic tree edit distance (ASTrED) compares the deep linguistic structure of the source and target sentences while taking word alignment into account. We open-source the implementation of the metrics as a Python package.

Broadly speaking, we are interested in ways to quantify translation difficulty. Syntactic equivalence is one part of that, as we have discussed in previous research (Vanroy et al. 2019a,b). In future work, we want to investigate whether we can distil typical word group reordering patterns, label changes or structural divergence and categorise them into Catford's obligatory and optional shifts (Catford 1965). The hypothesis is that in language pair-specific contexts, some word group orders, labels and structures are simply incompatible between two languages, in which case the translator is forced to make an obligatory shift and cannot rely on a literal translation. In addition, we want to perform more analyses using our metrics and compare them to translation process data. As a proof of concept, we presented one such analysis in Sect. 5, but since the used dataset is relatively small, similar experiments should be done to confirm, and expand on, these results. Moreover, we intend to run equivalent experiments on different language pairs to investigate (the difficulties between) syntactically divergent languages.

Finally, rather than calculating syntactic entropy based on the features Valency, Voice and Clause type (Bangalore et al. 2015), we are interested in investigating the feasibility of calculating syntactic entropy based on our metrics. Syntactic entropy can be simplified as the agreement between the translators of the same source text with respect to the syntax of their translations. Put differently, how similar or divergent in syntax are the different translations of the translators? Because our proposed metrics aim to quantify syntactic equivalence between a source sentence and its translation, they are good candidates to be used in an entropy setting to see how well translators agree on structural or syntactic changes when translating. This information, in turn, can be used in modelling the translatability of specific linguistic phenomena.

References

Andersen P (1990) How close can we get to the ideal of simple transfer in multi-lingual machine translation (MT)? In: Proceedings of the 7th Nordic conference of computational linguistics (NODALIDA 1989), Institute of Lexicography, Institute of Linguistics, University of Iceland. Reykjavík, Iceland, pp 103–113

Asadi P, Séguinot C (2005) Shortcuts, strategies and general patterns in a process study of nine professionals. Meta Trans J 50(2):522–547

Bangalore S, Behrens B, Carl M, Ghankot M, Heilmann A, Nitzke J, Schaeffer M, Sturm A (2015) The role of syntactic variation in translation and post-editing. Translation Spaces 4(1):119–144

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Borrillo JM (2000) Register analysis in literary translation: a functional approach. Babel 46(1):1–19. https://doi.org/10.1075/babel.46.1.02bor

Campbell S (1999) A cognitive approach to source text difficulty in translation. Target 11(1):33–63

Campbell S (2000) Choice network analysis in translation research. In: Olohan M (ed) Intercultural faultlines: research models in translation studies. St. Jerome, Manchester, pp 29–42

Carl M, Schaeffer MJ (2017) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. J Lang Commun Bus (56):43–57. https://doi.org/10.7146/hjlcb.v0i56.97201

Carl M, Schaeffer MJ, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer MJ (eds) New directions in empirical translation process research, New frontiers in translation studies. Springer, Cham, pp 13–54

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process. Translat Cognit Behav 2(2):211–232. https://doi.org/10.1075/tcb.00026.car

Catford JC (1965) A linguistic theory of translation: an essay in applied linguistics. Oxford University Press, Oxford

Chen KH, Chen HH (1995) Machine translation: an integrated approach. In: Proceedings of the sixth international conference on theoretical and methodological issues in machine translation, Leuven, pp 287–294

Collins-Thompson K (2014) Computational assessment of text readability: a survey of current and future research. Int J Appl Linguist 165(2):97–135. https://doi.org/10.1075/itl.165.2.01col

Daems J (2016) A translation robot for each translator. PhD thesis, Ghent University, Ghent

Daems J, Macken L, Vandepitte S (2013) Quality as the sum of its parts: a two-step approach for the identification of translation problems and translation quality assessment for ht and mt+pe. In: O'Brien S, Simard M, Specia L (eds) MT summit XIV workshop on post-editing technology and practice, proceedings, European association for machine translation, pp 63–71

De Clercq O, Hoste V (2016) All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. Comput Linguist 42(3):457–490. http://dx.doi.org/10.1162/COLI_a_00255

De Clercq O, Hoste V, Desmet B, van Oosten P, De Cock M, Macken L (2014) Using the crowd for readability prediction. Nat Lang Eng 20(3):293–325. http://dx.doi.org/10.1017/S1351324912000344

de Marneffe MC, Manning CD (2008) The Stanford typed dependencies representation. In: Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation, Coling 2008 Organizing Committee, Manchester, pp 1–8. https://www.aclweb.org/anthology/W08-1301

Dragsted B (2012) Indicators of difficulty in translation: correlating product and process data. Across Lang Cult 13(1):81–98. http://dx.doi.org/10.1556/Acr.13.2012.1.5

Dyer C, Chahuneau V, Smith NA (2013) A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of NAACL-HLT 2013, Association for Computational Linguistics, Atlanta, pp 644–648

Fox J, Weisberg S (2019) An R companion to applied regression, 3rd edn. Sage, Thousand Oaks

Francois T, Miltsakaki E (2012) Do NLP and machine learning improve traditional readability formulas? In: Proceedings of the workshop on predicting and improving text readability (PITR 2012), Montréal, Québec, pp 49–57

Gunning R (1952) The technique of clear writing. McGraw-Hill, New York

Hajič J, Zeman D (eds) (2017) Proceedings of the CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies, Association for Computational Linguistics, Vancouver. https://doi.org/10.18653/v1/K17-3. https://www.aclweb.org/anthology/K17-3000

Hansen-Schirra S, Nitzke J, Oster K (2017) Predicting cognate translation. Empirical Modelling of Translation and Interpreting 7:3

Jurafsky D, Martin JH (2008) Speech and language processing: an introduction to speech recognition, computational linguistics and natural language processing. Prentice Hall, Upper Saddle River

Kay M, Roscheisen M (1993) Text-translation alignment. Comput Ling 19(1):121–142

Kincaid JP, Fishburne RP, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research branch report RBR-8-75, Naval Technical Training Command Millington Tenn Research Branch, Springfield

Koster J (1975) Dutch as an SOV language. Ling Anal, 111–136

Kromann M (2003) The Danish dependency treebank and the DTAG treebank tool. In: Proceedings of the 2nd international workshop on treebanks and linguistic theories

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: tests in linear mixed effects models. J Stat Softw 82(13):1–26. https://doi.org/10.18637/jss.v082.i13

Liu Y, Zheng B, Zhou H (2019) Measuring the difficulty of text translation: the combination of text-focused and translator-oriented approaches. Target 31(1):125–149. https://doi.org/10.1075/target.18036.zhe

Matthews P (1981) Syntax. Cambridge textbooks in linguistics, Cambridge University Press, Cambridge

Mihalcea R, Pedersen T (2003) An evaluation exercise for word alignment. In: Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts data driven machine translation and beyond, Association for Computational Linguistics, Edmonton, vol 3, pp 1–10. https://doi.org/10.3115/1118905.1118906

Mishra A, Bhattacharyya P, Carl M (2013) Automatically predicting sentence translation difficulty. In: Proceedings of the 51st annual meeting on association for computational linguistics (ACL 2013), Sofia, pp 346–351

Nivre J (2015) Towards a universal grammar for natural language processing. In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, pp 3–16. https://link.springer.com/chapter/10.1007/978-3-319-18111-0_1

Nivre J, Megyesi B (2007) Bootstrapping a Swedish treebank using cross-corpus harmonization and annotation projection. In: Proceedings of the 6th international workshop on treebanks and linguistic theories, pp 97–102

Nivre J, De Marneffe MC, Ginter F, Goldberg Y, Hajic J, Manning CD, McDonald R, Petrov S, Pyysalo S, Silveira N, et al. (2016) Universal dependencies v1: a multilingual treebank collection. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 1659–1666

Och FJ, Ney H (2000) A comparison of alignment models for statistical machine translation. In: Proceedings of the 18th conference on computational linguistics, Association for Computational Linguistics, Saarbrücken, vol 2, pp 1086–1090. https://doi.org/10.3115/992730.992810

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51. https://doi.org/10.1162/089120103321337421

Osborne T, Gerdes K (2019) The status of function words in dependency grammar: a critique of Universal Dependencies (UD). Glossa J General Linguist 4(1):17. https://doi.org/10.5334/gjgl.537

Pawlik M, Augsten N (2015) Efficient computation of the tree edit distance. ACM Trans Database Syst 40(1). https://doi.org/10.1145/2699485. http://dl.acm.org/citation.cfm?doid=2751312.2699485

Pawlik M, Augsten N (2016) Tree edit distance: Robust and memory-efficient. Inf Syst 56:157–173. https://doi.org/10.1016/j.is.2015.08.004. https://linkinghub.elsevier.com/retrieve/pii/S0306437915001611

Peng X, Li Z, Zhang M, Wang R, Zhang Y, Si L (2019) Overview of the NLPCC 2019 shared task: cross-domain dependency parsing. In: CCF international conference on natural language processing and Chinese computing, Springer, Berlin, pp 760–771

Pym A (2014) Exploring translation theories, 2nd edn. Routledge, London

Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: a Python natural language processing toolkit for many human languages. arXiv:2003.07082

R Core Team (2019) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. https://www.R-project.org/

Schaeffer M, Carl M (2014) Measuring the cognitive effort of literal translation processes. In: Proceedings of the EACL 2014 workshop on humans and computer-assisted translation, Association for Computational Linguistics, Gothenburg, pp 29–37. https://doi.org/10.3115/v1/W14-0306

Schwarm SE, Ostendorf M (2005) Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd annual meeting on association for computational linguistics (ACL 2005), Ann Arbor, pp 523–530. https://doi.org/10.3115/1219840.1219905

Skut W, Krenn B, Brants T, Uszkoreit H (1997) An annotation scheme for free word order languages. In: Fifth conference on applied natural language processing, Association for Computational Linguistics, pp 88–95. https://doi.org/10.3115/974557.974571. https://www.aclweb.org/anthology/A97-1014

Steiner E (2004) Ideational grammatical metaphor: exploring some implications for the overall model. Lang Cont 4(1):137–164. https://doi.org/10.1075/lic.4.1.07ste

Sun S (2015) Measuring translation difficulty: theoretical and methodological considerations. Across Lang Cult 16(1):29–54. https://doi.org/10.1556/084.2015.16.1.2. http://www.akademiai.com/doi/abs/10.1556/084.2015.16.1.2

Sun S, Shreve GM (2014) Measuring translation difficulty: an empirical study. Target 26(1):98–127. https://doi.org/10.1075/target.26.1.04sun. https://benjamins.com/online/target/articles/target.26.1.04sun

Tirkkonen-Condit S (2005) The monitor model revisited: evidence from process research. Meta Transl J 50(2):405–414

Vanroy B, De Clercq O, Macken L (2019a) Correlating process and product data to get an insight into translation difficulty. Perspectives 27(6):924–941. https://doi.org/10.1080/0907676X.2019.1594319

Vanroy B, Tezcan A, Macken L (2019b) Predicting syntactic equivalence between source and target sentences. Comput Linguist Neth J 101–116. https://www.clinjournal.org/clinj/article/view/95

Zeman D, Hajič J, Popel M, Potthast M, Straka M, Ginter F, Nivre J, Petrov S (2018) CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2018 shared task: multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics, Brussels, pp 1–21. https://doi.org/10.18653/v1/K18-2001. https://www.aclweb.org/anthology/K18-2001

Zwart CJW (1994) Dutch is head-initial. Ling Rev 11(3–4), https://doi.org/10.1515/tlir.1994.11.3-4.377

Using a Product Metric to Identify Differential Cognitive Effort in Translation from Japanese to English and Spanish

Isabel Lacruz, Haruka Ogawa, Rika Yoshida, Masaru Yamada, and Daniel Ruiz Martinez

Abstract We examine the variability of Japanese-English and Japanese-Spanish translations at the level of *bunsetsu* (文節), the smallest coherent linguistic units that sound natural as part of Japanese sentences. These are equivalents of chunks or phrases in English, linguistic units generally larger than a word but smaller than a sentence. We measure variability by adapting the widely studied word translation entropy metric HTra to the context of *bunsetsu*. Word translation entropy has been shown to correlate with various behavioral measures of cognitive effort during translation between several language pairs. Word translation entropy values also correlate for translations of the same English source texts into several languages. Here, we extend the range of prior findings to translations from Japanese, a very different source language to English. We exhibit significant correlations of word translation entropy values in Japanese-English and Japanese-Spanish translations of *bunsetsu* from the same source texts. In line with prior observations on comparability of cognitive effort exerted in translations from English to closely related European languages, we also find comparable average word translation entropy values at the *bunsetsu* level for translations from Japanese to English and to Spanish. Nevertheless, we exhibit examples where there are large differences between entropy values for translations of specific types of *bunsetsu* into English

I. Lacruz (✉) · H. Ogawa
Kent State University, Kent, OH, USA
e-mail: ilacruz@kent.edu; hogawa@kent.edu

R. Yoshida
Rikkyo University, Tokyo, Japan

M. Yamada
Kansai University, Suita, Japan
e-mail: yamada@apple-eye.com

D. Ruiz Martinez
Universidad de Salamanca, Salamanca, Spain
e-mail: druiz@usal.es

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_11

and Spanish, relating these differences to general characteristics of the languages, such as the degree of dependence on context to infer meaning. We propose that in appropriate circumstances, different levels of cognitive effort during the translation process can be identified through differences in the variability of the translation product.

1 Introduction

Written translation from one language to another involves complex cognitive processes including reading in one language and production in the other language, mediated by transfer between languages. The investigation of these largely subconscious cognitive processes relies heavily on highly developed methodologies used in cognitive psychology, particularly psycholinguistics, and relies crucially on linguistic theories and classifications.

Much work has been done to identify good measures of mental effort and to investigate sources of heightened mental effort in translation tasks. A principal goal of this chapter is to further our understanding of how and why effort levels vary for translation into different languages. In line with previous research for translation from English to a variety of languages, we anticipate clear correlation of effort levels for Japanese-to-English and Japanese-to-Spanish translations of the same texts (see also Ogawa et al., Chap. 6), but we intend to probe more deeply, anticipating discrete discrepancies stemming from linguistic and cultural contrasts associated with the three languages.

Mental or cognitive effort is a function of the load imposed on working memory. Specifically, cognitive effort is "the amount of the available processing capacity of the limited-capacity central processor utilized in performing an information-processing task" (Tyler et al. 1979). It must be measured indirectly. One common measure is subjective self-evaluation of effort recorded once the translation is completed. Another widespread measure is overall time spent, sometimes known as temporal effort (Krings 2001). To facilitate comparisons between effort spent on different texts or at different points in a text, temporal effort is often normalized as time spent per word or character in the text to be translated (source text, ST) or in the translated text (target text, TT). However, temporal effort captures more than the cognitive effort exerted, since part of the time is spent on the mechanical effort exerted in actually writing or typing the TT (Krings 2001).

More sophisticated measures of cognitive effort can use direct measurements of electrical activity or blood flow in the brain during translation, but these are very labor- and resource-intensive and still only provide indirect information about the mind of the translator (Lachaud 2011). Translation process researchers more commonly use eye movement data collected with an eye tracker or typing data collected with a keystroke logger. Eye trackers can record a translator's eye movements during reading: the eyes jump from one spot to another in the text, stopping to fixate on a word or group of words before moving on to the next or

moving backward (regressing) to gather more information. The general assumption by cognitive psychologists (Just and Carpenter 1976) is that cognitive processing is carried out during fixations and that longer or more fixations at a particular point in a text or in the text as a whole indicate expenditure of greater cognitive effort at that point or in the text. Common eye tracking metrics are first fixation duration (the time spent the first time the eye stops at a particular point in the text), gaze duration (the total time spent during all fixations at a particular point or in the text), and fixation count (the total number of fixations at a particular point or in the text) (Rayner and Pollatsek 1989).

Translation involves hands (or voice) as well as eyes. Keystroke logging offers a further window into cognitive effort in translation. Just as pauses in eye movements indicate cognitive effort, so do pauses in typing (Schilperoord 1996; Krings 2001). The number and duration of typing pauses, normalized in various ways, are reliable indicators of cognitive effort in translation (Lacruz and Shreve 2014; Lacruz 2017).

Reading the source text activates linguistic representations in both source and target languages, and then a selection process results in production of the target text. All of these are recursive processes that interact with each other but proceed with general forward momentum (Schaeffer and Carl 2013; Carl and Schaeffer 2017a). Simultaneous recording of different cognitive effort metrics permits comparisons, or triangulation, between them over the time course of a translation, and this triangulation enables researchers to refine models of the translation process (Alves 2003). Separation of behavioral metrics for the source text and the target text facilitates studies of interactivity (Carl et al. 2016a, b).

The development of the large CRITT database, which records metrics computed from many empirical studies across different languages and translation modalities (e.g., from-scratch translation, post-editing, revision, dictation, and others), has allowed the large-scale comparison of translations between different language pairs. In this context, the concept of entropy is useful as a proxy for behavioral measurements of cognitive effort (see also Carl Chap. 5). Entropy essentially measures the variability in multiple translations of the same text. If there is no variability and all the translations are identical, the entropy is zero; if there is maximum variability and all translations are different, the entropy is the highest it can be. Word translation entropy is high if the same word is translated in several different ways. In situations where translators make highly variable decisions on the translation of a word, where the word translation entropy is high, it can be expected that they are expending high levels of cognitive effort in considering several alternatives and selecting one of them. This would be less likely to be the case when there is little variability in the translation decisions, when the word translation entropy is low.

The word translation entropy metric HTra (Carl et al. 2016a, b) was borrowed from information theory as developed by Shannon in 1948, but see (Shannon 2001) for an accessible reprint. Consistent with the discussion above, HTra turns out to correlate significantly with behavioral metrics of cognitive effort based on eye tracking and keystroke logging (Carl and Schaeffer 2017b). Although the computation of HTra is a little involved, it can easily be programmed in a

spreadsheet. If several translators offer translations $t_1, t_2, \ldots, t_n$ of the same word w with relative frequencies $p_1, p_2, \ldots, p_n$, the word translation entropy of w is

$$\mathrm{HTra} = -\left(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \cdots + p_n \log_2 p_n\right)$$

It is remarkable that when the same text is translated by multiple groups of translators from English to several other languages, the word translation entropies for the different language pairs correlate significantly (Tokowicz 2014; Schaeffer et al. 2018; Carl and Baez 2019; Carl (Chap. 5); Ogawa et al. Chap. 11). This finding even applies when one group of "translators" consists of several MT programs (Almazroei et al. 2019). This is perhaps not surprising since MT programs use human translations as starting points for training. The finding also applies regardless of the nature of the languages in a pair: the correlation still holds true for remote language pairs, such as English and Japanese, as well as for pairs of more closely related languages, such as English and Spanish.

What this finding says is that if certain source text features demand elevated cognitive effort to translate into one language, they will also tend to demand elevated cognitive effort to translate into another language. It does not say anything about the degree of elevation; this will vary from one target language to another. For example, using keystroke logging measures of cognitive effort, Lacruz et al. (2016) found that translation from English to remote languages (Japanese and Hindi) tended to require more cognitive effort than translation from English to closer languages (Spanish and Danish.)

Word translation entropy would appear at first sight to reflect cognitive effort expended in the production of a target text. However, there are other parts of the translation process where translators can be expected to expend substantial cognitive effort, in particular during the reading and comprehension of the source text (see also Wei, Chap. 7). There could be various influences on this locus of cognitive effort. The L1 of the translator can be expected to play a role here, particularly when the languages in the translation pair have very different cultural conventions, in particular when there are differences in expectations about reliance on context to convey information. There are interesting differences in reliance on context in English, Japanese, and Spanish that will allow us to use word translation entropy in the different pairs as a tool to distinguish situations where cognitive effort arises during source text comprehension as opposed to target text production.

2 Rationale

In this paper, we investigate translations of the same source materials from Japanese into two remote languages (English and Spanish) that are relatively close to each other. We expect to find that the word translation entropy (and, by extension, cognitive effort) correlations described above for out-of-English translations (Tokowicz

2014; Schaeffer et al. 2018; Carl and Baez 2019) will also appear in the new context of out-of-Japanese translations into English and Spanish. Since English and Spanish are relatively close languages, we do not expect to find significant overall differences in cognitive effort between the Japanese-to-English and the Japanese-to-Spanish translations, in line with the findings of Lacruz et al. (2016).

If these expectations are confirmed, we plan to dig deeper, triangulating between the language pairs to identify situations where the tendency underlying the correlation breaks down. If we are able to identify such situations, then we will have identified Japanese source text features that are associated with heightened cognitive effort in translation into *both* English and Spanish, which could be interpreted as due to difficulties in reading and comprehension of the Japanese source text. We will also have identified Japanese source text features that are associated with heightened cognitive effort in translation into *one, but not both* target languages, which we can interpret as likely due to difficulties in making the necessary selections for target text production. In other words, solely on the basis of examination of completed translation products, we will have identified source text features that cause heightened cognitive effort at different stages of the translation process.

3 Participants and Materials

We selected two Japanese texts of the equivalent of approximately 100 words each. They were extracted from general texts and were of similar difficulty. These texts were translated into English or Spanish by translators or students affiliated with American, Japanese, and Spanish universities.

The two texts were translated into Spanish (S) by 14 participants:

- A group (S-L1J) of seven Japanese participants (L1 Japanese, L2 Spanish): six were undergraduate students of Spanish at a Japanese university, and one was a professional Spanish-Japanese translator.
- A group (S-L2J) of seven Spanish participants (L1 Spanish, L2 Japanese): six were undergraduate students of Japanese at a Spanish university, and one was a professional Japanese-Spanish translator.

The two texts were translated into English (E) by 13 participants:

- A group (E-L1J) of seven Japanese participants (L1 Japanese, L2 English): five were MA students of English at a Japanese university, and two were professional English-Japanese translators.
- A group (E-L2J) of six American participants (L1 English, L2 Japanese): five were students in an MA in Translation program at an American university, and one was a professional Japanese-English translator.

4 Alignment Process

We aimed to use product variability at the word level, measured by HTra, as a way to trace the cognitive effort required to translate from Japanese to English and to Spanish. In order to compute HTra, the source and target text had first to be aligned to allow counting of translation variants.

Despite some structural differences between English and Spanish, alignments for this language pair are relatively straightforward and are facilitated by clear word boundaries, even when the structures differ somewhat. As an example, consider the English phrase:

he will drink an iced coffee.

A possible translation into Spanish would be:

tomará un café con hielo.

Optimal alignment of the English and Spanish phrases requires a segmentation of the source phrase and the target phrase into semantic units of a size that is natural for both languages, followed by an alignment mapping from the source text units to the target text units. In this case, the segmentation and alignment could be:

[*he will drink*] [*an*] [*iced*] [*coffee*].

[*tomará*] [*un*] [*café*] [*con hielo*].

This pairing highlights several structural differences between English and Spanish that generally rule out simple word-by-word alignment.

- The subject of a verb is often omitted in Spanish and is instead inferred from the form of the verb. In English, subjects of verbs cannot normally be omitted.
  - The English *he* can only align here in combination with other words.
- In contrast to English, Spanish verb tenses are often indicated by inflection.
  - The English *will* that is used to form the future tense can only align here in combination with other words.
- The default placement of Spanish adjectives is after the noun, rather than before the noun, as in English.
  - Alignment in this example requires a change in the word order.

The alignment process becomes much more complex for translations between Japanese and English or Spanish, since the structures and orthographies are very different. One immediate issue is the agglutinative nature of Japanese, where word boundaries are not indicated. Another is the much greater need for context to understand a Japanese sentence. The Japanese sentence 彼はアイスコーヒーを飲みますcan be translated in a variety of ways (including *he will drink an iced coffee, he drinks an iced coffee, he drinks the iced coffee, he drinks iced coffee*) where the

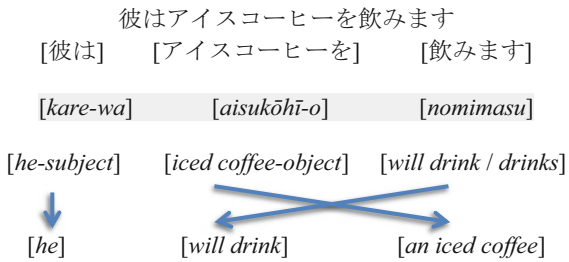
actual tense translation and the choice of article require additional context in the Japanese text.

In order to align the Japanese text with the English translation, there is a need to systematically tokenize the Japanese text to account for the lack of spaces between Japanese words. However, morphological analysis is too granular, since Japanese makes extensive use of function words, such as subject/object markers, which would need to be left out in alignment. We chose to base our alignment on segmentation of the Japanese text into *bunsetsu* (文節), smallest coherent linguistic units that sound natural as part of a sentence in Japanese. They are equivalents of chunks or phrases in English, linguistic units larger than a word but smaller than a sentence, and have also been used as smallest coherent linguistic units when conducting psycholinguistic experiments on Japanese reading processes. *Bunsetsu* always include a content word, sometimes followed by a (possibly empty) string consisting of an auxiliary verb and/or a particle.

Other complicating factors for alignment are that Japanese uses no articles and that word order in Japanese is very different from word order in English or in Spanish. In particular, the default structure in Japanese is subject-object-verb, as opposed to subject-verb-object in English or Spanish.

To actually carry out alignments, we first processed the Japanese STs through a Japanese dependency analyzer CaboCha[1] to divide sentences into *bunsetsu* (Kudo and Matsumoto 2002). We then manually mapped TTs onto the *bunsetsu* produced by CaboCha using Excel sheets. We focused on meaning when finding an equivalent of a *bunsetsu* in TTs, without pre-processing TTs through a dependency analyzer. This meant that one *bunsetsu* could have multiple TT words from different parts of the sentence or even from another sentence, for example, in a case where a participant translated one ST sentence into two or more TT sentences. For more creative translations (often called free translations) or translations that deviated from the pure ST meaning (such as through explicitation), we attempted to find a source text word or *bunsetsu* that triggered such a translation and map it from the *bunsetsu* in question. Rarely, we merged two *bunsetsu* in order to create a coherent mapping between the ST and TT. This procedure enabled us to consistently count the number of alternative translations based on the ST and so to calculate the HTra value.

We illustrate the alignment process with the example used above in the context of English-Spanish alignment. We tokenize the Japanese text using *bunsetsu* and then render those in *romaji* to help clarify the alignment.



[1]Available at https://taku910.github.io/cabocha/

5 Method and Analysis

The two Japanese source texts were translated into Spanish by the 14 participants in the S-L1J and S-L2J groups and into English by the 13 participants in the E-L1J and E-L2J groups. No time limits were set for the translations. Participants were free to use dictionaries.

We divided each Japanese source text into *bunsetsu*, and these were manually aligned, as described above, with the target texts produced by each participant. In total, there were 57 *bunsetsu*. For each participant group, we computed the mean HTra value for each *bunsetsu.*

We planned to compare HTra values for the four groups. For the S-L1J, S-L2J, and E-L1J groups, the possible HTra values ranged from 0 to $\log_2(7) = 2.81$, while the possible HTra values for the E-L2J group ranged from 0 to $\log_2(6) = 2.58$. In order to be able to make direct comparisons of HTra values for all four groups, we scaled the E-L2J HTra values linearly to vary between 0 and $\log_2(7)$. In other words, we multiplied all the E-L2J group HTra values by $\log_2(7)/\log_2(6) = 1.09$. Scaling was necessary for the analysis of variance (ANOVA) below. It could also have been achieved, for example, by normalizing the maximum possible entropy to 1 in each group.

6 Results

Correlations are summarized in Table 1. HTra values for *bunsetsu* translated by the S-L1J group correlated strongly and positively with the HTra values for the same *bunsetsu* translated by the E-L1J group, $r(55) = 0.58$, $p < 0.001$. In addition, HTra values for *bunsetsu* translated by the S-L2J group correlated strongly and positively with the HTra values for the same *bunsetsu* translated by the E-L2J group, $r(55) = 0.51$, $p < 0.001$. This result is consistent with the expectation that heightened cognitive effort for *bunsetsu* translation into Spanish is associated with heightened cognitive effort for *bunsetsu* translation into English, regardless of whether the translations are into the participants' first or second language.

Table 1 Pearson *r* correlations between HTra values for the four translation groups

	E-L1J	E-L2J	S-L1J	S-L2J
E-L1J	1	0.31*	0.58***	0.54***
E-L2J		1	0.52***	0.51***
S-L1J			1	0.61***
S-L2J				1

*Correlation significant at the 0.05 level (two-tailed), ***correlation significant at the 0.001 level (two-tailed)

We also found significant positive correlations for translations into L1 and into L2. HTra values for *bunsetsu* translated by the S-L1J group correlated strongly and positively with the HTra values for the same *bunsetsu* translated by the S-L2J group, $r(55) = 0.61$, $p < 0.001$. In addition, HTra values for *bunsetsu* translated by the E-L1J group correlated moderately and positively with the HTra values for the same *bunsetsu* translated by the E-L2J group, $r(55) = 0.31$, $p = 0.018$.

HTra values were also submitted to a factorial analysis of variance (ANOVA) with two independent variables manipulated between subjects: language with two levels (into Spanish and into English) and directionality with two levels (into L1 and into L2). Neither the main effects nor the interaction was significant ($p > 0.05$.)

7 Discussion

Various previous studies (Tokowicz 2014; Schaeffer et al. 2018; Carl and Baez 2019) found significant positive correlations of cognitive effort for translations of the same source text from English to a variety of other languages (see also Ogawa et al., Chap. 6). The present study, using HTra as a metric for cognitive effort, provides support for the hypothesis that such correlations persist for translations from other languages, in this case from Japanese to English and Spanish.

Previous studies (e.g., Lacruz et al. 2016) had found comparable cognitive effort, measured by HTra, for translations of the same source texts from English to other relatively close European languages (e.g., German and Danish). The present study provides support for the hypothesis that this finding should continue to hold true when the source language is remote from the target languages. In the present case, there is comparable cognitive effort for translations from Japanese to the remote but relatively closely related languages, English and Spanish.

Naturally these findings raise the question of whether they can be replicated in different language pairings. But they also prompt questions of whether there are predictable circumstances when there is differential cognitive effort for translation into other languages that deviates from the general correlation patterns. In our specific case, we ask, for example, whether specific characteristics of the languages might explain situations where there is high HTra for a *bunsetsu* translated into English, but low HTra for the same *bunsetsu* translated into Spanish, or vice versa.

To visualize the discrepant *bunsetsu*, we use 100% stacked column charts to compare into-English and into-Spanish HTra for L1-Japanese participants (Fig. 1) and separately for L2-Japanese participants (Fig. 2). Each vertical column in Fig. 1 and Fig. 2 corresponds to a single *bunsetsu*. For each *bunsetsu*, we computed the proportion of the total HTra attributable to the translations into English (red bar) and into Spanish (blue bar).

We consider *bunsetsu* 26 to explain the relationship between the lengths of the vertical bars. For this *bunsetsu*, the S-L1J HTra value was 0.59 (a low value), and the E-L1J HTra value was 2.13 (a high value). This large discrepancy is apparent from Fig. 1, where the blue S-L1J bar is much shorter than the red E-L1J bar.

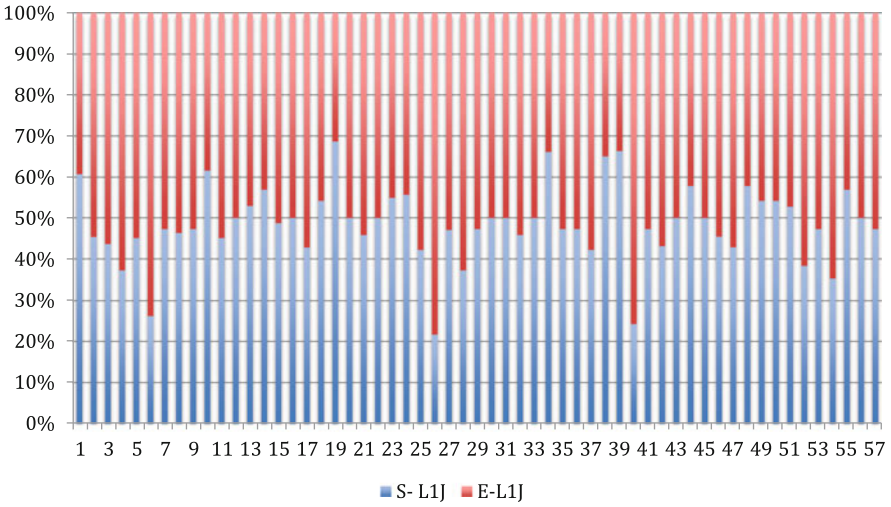


Fig. 1 Proportions of total HTra for L1J translations into Spanish and into English for each *bunsetsu*

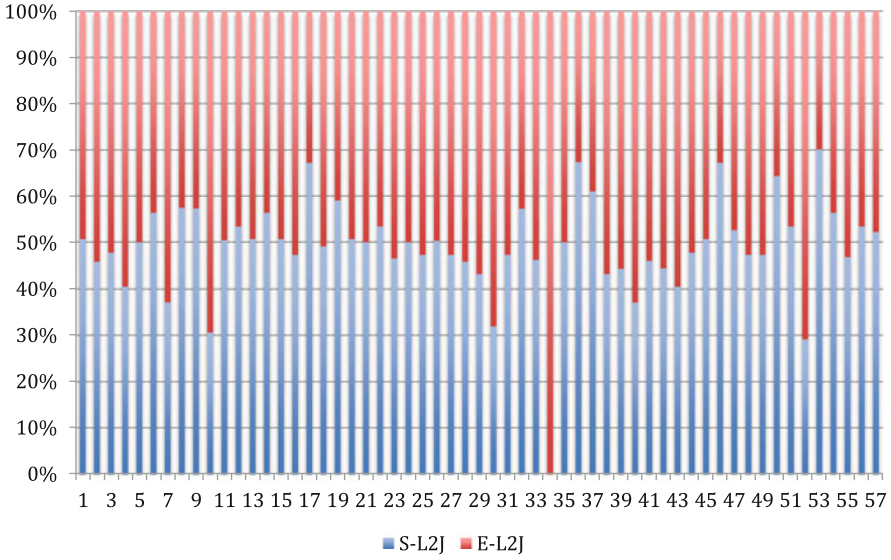


Fig. 2 Proportions of total HTra for L2J translations into Spanish and into English for each *bunsetsu*

The total HTra recorded for this *bunsetsu* is $0.59 + 2.13 = 2.72$ (the sum of the S-L1J HTra and the E-1 J HTra). The proportion of the total HTra coming from the S-L1J HTra is $0.59/2.72 = 0.217$ or 21.7%. Figure 1 shows this is the length

of the blue bar. The proportion of the total HTra coming from the E-L1J HTra is 2.13/2.72 = 0.783 or 78.3%, which is the length of the red bar.

We consider this particular case in Example 1, an example of proportionally high into-English HTra and proportionally low into-Spanish HTra for L1J translators.

Differences in Japanese, Spanish, and American cultures are reflected in language structure, use, and conventions, which influence the translation process. This influence can manifest itself throughout the process. Angelone (2010) proposed a stratificational model according to which translating implies at least three distinct types of processes. These are early processes of reading and comprehension of the source text, intermediate processes of transfer to the target language, and late processes of written production in the target language. Naturally, there can be interactivity between these processes. A recursive model (e.g., Schaeffer and Carl 2013) is better positioned to capture the complexity and interactions of the various processes involved in translation.

Emerging theories, as proposed by Carl et al. (2019), build on recent views of bilingual language processing (e.g., Dijkstra et al. 2018), which explicitly posit non-language-selective word identification processes. When a source language word is read, orthographically similar words or word fragments are automatically activated in both the source and target languages, and these orthographic representations immediately activate semantic representations in both languages. Such patterns of activation are initially highly unstructured. Resulting high levels of ambiguity in the information being processed (high information entropy) must rapidly be resolved through processes of excitation and suppression. These give access to the meaning of the source text and set up a decision process that results in the selection of a single appropriate translation. Appropriateness must be achieved both at the granular lexical level and at the more global syntactic and contextual levels. Cognitive effort must be exerted for these processes to reduce the information entropy sufficiently to reach a conclusion. The degree of effort will depend on various factors, including the need to suppress inappropriate lexical and semantic activations in the source and target languages. When there are many translation candidates to exclude, different translators are likely to make a range of choices, so HTra is also likely to be higher.

In the realm of physics, work (or effort) results when a physical force acts to move a physical body over a distance. One could consider the mental effort made in selecting translation solutions in an analogous way. Mental forces act to move one of the possible translation solutions to the forefront as the final selection. Translation process researchers (Halverson 2003) have suggested that the selection of translation solutions is guided by a metaphorical "gravitational pull," whereby the most salient translation solutions attract the most attention, require the least effort from translators, and are most likely to be selected.

Another perspective might be that mental work or effort is expended as the result of metaphorical mental forces acting to create order (reduce entropy) in the highly unstructured information automatically generated by nonselective lexical and semantic activation of both languages (Dijkstra et al. 2018) during the reading of the source text. Wei (Chap. 7), gives a detailed account of such processes. These mental forces can be considered as "entropic gravity" forces (Carl et al. 2019). For

each translator, the initial activation chaos gives way to emerging order as excitatory and inhibitory processes engage and allow various translation candidates to become available. Intensifying cognitive effort is expended as decision processes guide the selection of translation candidates and evaluate their possible integration into the final translation product, acting to move one of them to the status of translation solution. Notice that differences in language organization and cultural constructs in L1 and L2 may be detectable through analysis of decisions made by different groups of translators.

In terms of a mental analogy to the physical world, cognitive effort is exerted as a result of entropic gravity forces that pull the target text together. The more translation candidates there are to choose between, the more work entropic gravity will have to perform (the more cognitive effort will be exerted on average), and the more likely it will be that different translators select alternative translation solutions (resulting in higher word translation entropy).

Differences in HTra (and so cognitive effort) patterns for Japanese-to-English and Japanese-to-Spanish translation can point to expenditure of cognitive effort in different aspects of the translation process. For example, high HTra in one language pair but not the other suggests that the discrepancy in cognitive effort may occur due to differences in production and revision processes associated with the target languages, possibly due to more entrenched cross-linguistic equivalents in one pair but not the other. In contrast, when there is high HTra in both language pairs, the elevated cognitive effort could be due to source text reading processes or to factors related to the source text that make it difficult to translate into both languages. For example, source text non-compositionality, as in metaphorical expressions where the meaning cannot be generated through a literal reading, or culture-specific words or expressions (including metaphors) where there are no direct equivalents in the target language are likely sources of elevated cognitive effort. Further information might also be gleaned when there are differences in HTra for translation into the first and second languages.

One factor that highlights differences between Japanese, English, and Spanish cultures is the extent of their reliance on context to achieve understanding. According to Byrne (2012), "context is the amount of explicit information we need to include in a given communicative act so that the recipient can understand us." Languages fall on a spectrum from high context, where understanding is reliably achieved even when information is expressed implicitly, to low context, where information must be expressed explicitly to ensure understanding (Hall 1976.) Among the languages considered by Katan (1999), German is very low context, and Japanese is extremely high context. English and Spanish are intermediate: Spanish is high context, but English is low context.

Example 1. High HTra into English, Low HTra into Spanish We examine the translations of the *bunsetsu* **あるのは** *aru-no-wa* (number 40 in Fig. 1 and Fig. 2). This appears in the sentence:

その背後に**あるのは**、大人のことばのほうが普通であって、
子どものことばのほうが普通でないということでしょう。

Google translates this into English as:

The reason *behind that is that the language of adults is more common, and the language of children is less common.*

In Spanish, the Google translation is:

Detrás está que las palabras adultas son más comunes y las palabras infantiles son menos comunes.

The HTra values for translation into English (1.84 for L1J; 1.95 for L2J) are substantially higher than the HTra values for translation into Spanish (0.59 for L1J; 1.15 for L2J). The underlying translations are shown in Table 2. This systematic difference in HTra values points to differential cognitive effort involved in the production of the target texts. As an isolate, あるのは*aru-no-wa* translates in English as *there is* or in Spanish as *hay.* Most of the translations into Spanish omit any translation of this *bunsetsu*, consistent with the high-context nature of Spanish, where information is often conveyed implicitly. On the other hand, there is a much higher need to convey information explicitly in low-context English, which leads to a wider availability of possible translations and so to higher entropy.

When the source text *bunsetsu* conveys little specific information, it should tend to be much less effortful to render in high-context Spanish, where omission may be a reasonable option, than in high-context English, where the translation demands that choices be generated and narrowed down to a selection. This difference impacts the

Table 2 Translations of *bunsetsu* 40 with HTra values

あるのは、***aru-no-wa,*****(*there is*)**	
L1J into English	**L1J into Spanish**
–	–
The idea	–
The reason	–
What lies	–
The reason	Lo que hay
The reason	–
–	–
HTra 1.84	**HTra 0.59**
L2J into English	**L2J into Spanish**
–	–
–	–
This way of thinking is	–
–	–
The reasoning	la razón
This assumption	Subyace la idea de que
	–
HTra 1.95	**HTra 1.15**

cognitive effort required to produce the target text, which should be greater when the target language is low context, regardless of the translation direction.

Example 2. Low HTra into English, High HTra into Spanish We next examine the translations of the *bunsetsu* 代表する *daihyousuru* (number 19 in Fig. 1 and Fig. 2). This appears in the sentence:

食はその国を代表する文化であり

Google translates this into English as:

*Food is the cultural **representative** of the country.*

The Spanish versión is:

*La comida es una cultura **que representa** al país.*

The HTra pattern is flipped from the one in Example 1. The HTra values for translation into English (1.15 for L1J; 1.95 for L2J) are substantially *lower* than the HTra values for translation into Spanish (2.52 for L1J; 2.81 for L2J). The underlying translations are shown in Table 3. This systematic difference in HTra values points to differential cognitive effort involved in the production of the target texts.

However, in this example, the underlying reason is not an issue of high- or low-context languages. Instead, the differences can be explained by the fact that Spanish tends to use more complex structure and less direct expression than English. This can be seen, for example, in the fact that text tends to grow in length when it is translated from English to Spanish, and to decrease in length when the translation

Table 3 Translations of *bunsetsu* 19 with HTra values

代表する ***daihyousuru*(represent)**	
L1J into English	**L1J into Spanish**
Represents	Que representa
Represents	a representar
Represents	Expresa
Representing	Que puede representar
Represents	Que representa
Represents	Representativa
Representative	Representa
HTra 1.15	**HTra 2.52**
L2J into English	**L2J into Spanish**
–	Representa
Representative	Una parte representativa
Acts as an ambassador for	Representativa de
Representative	Que representa
Representative	Emblematica
As _ represents _	Representativa
	Parte integrante
HTra 1.95	**HTra 2.81**

direction is the opposite. Meaning is expressed in multiple words more frequently in Spanish than in English, long sentences are more common in Spanish than in English, and subordinate clauses are used more frequently in Spanish than in English (López Guix and Minett Wilkinson 2006). These language characteristics tend to result in simpler sentence structure in English than in Spanish, and when they are present, the cognitive effort of production should be lower in English than in Spanish.

The translations of *bunsetsu* 19 illustrate these comments. The English renderings of 代表する*daihyousuru* (represent) are mostly single words and rarely relative clauses, which restricts the opportunities for variation and so keeps the entropy low. However, Spanish translations frequently use multiple words, and there are several instances of relative clauses, which invite wider variation as is evident in the higher HTra.

Accordingly, it appears that there are predictable structural differences between English and Spanish that, when present, will tend to result in higher cognitive effort at the production stage of English translations than of Spanish translations.

Example 3. High HTra for L1J Participants, but Only When Translating into English The translations of *bunsetsu* 日本を ***nihon-wo*** (number 26 in Fig. 1 and Fig. 2) follow a different HTra pattern. This *bunsetsu* appears in the segment:

日本を 体験し、理解するための旅行の大切な要素であり、楽しみになっています。

Google translates this into English as:

It is an important element of travel to experience and understand ***Japan*** *and I am looking forward to it.*

The translation into Spanish is:

Es un elemento importante del viaje experimentar y comprender ***Japón****, y estoy deseando que llegue.*

For the L1 Japanese participants, the HTra level was much higher for translations into English than for translations into Spanish, which suggests differential cognitive effort during production of the target text. However, for the L2 Japanese participants, the HTra levels were intermediate and comparable for translations into both languages. This might be taken to indicate that cognitive effort arose during source text reading. The underlying translations are shown in Table 4.

Close examination allows us to resolve this apparent contradiction. First, it is important to note that the high frequency of the *bunsetsu* would appear to rule out high expenditure of cognitive effort during reading and comprehension. The *bunsetsu* is composed of two distinct parts, namely, 日本*nihon* (Japan) and the grammatical particle を*wo* (object marker). It is not credible that either of these components would cause significant problems with reading and comprehension or language transfer. Accordingly, we turn to cognitive issues surrounding target text production.

Table 4 Translations of *bunsetsu* 26 with HTra values

日本を*nihon-wo*(Japan)	
L1J into English	**L1J into Spanish**
The country of Japan	Japón
The country	–
The Japanese culture	Japón
Japan	Japón
Japan	Japón
"Japan"	Japón
Japan	Japón
HTra 2.13	**HTra 0.59**
L2J into English	**L2J into Spanish**
Japan	–
Japan	Japón
Japanese culture	el país
Japan	el país
Japan	Japón
The country	Japón
	Japón
HTra 1.36	**HTra 1.38**

The key lies once again in the contrasts between the high-context and low-context characteristics of the three languages. Recall that Japanese is very high context and relies heavily on context to generate meaning. On the other hand, while Spanish is moderately high context, American English is low context, and so readers of American English make less use of implicit cues to generate meaning, relying more on explicit references.

L1 Japanese participants are likely to be highly attuned to the need to be explicit in order to be well understood when translating into English. This is likely to result in over-explicitation in some instances, which will lead to relatively high variation in the target language, in other words to the observed high value of HTra. On the other hand, since both Japanese and Spanish are high-context languages, although to different degrees, L1 Japanese participants translating into Spanish are likely to feel comfortable with the most obvious translation of 日本を*nihon-wo* as simply *Japón,* resulting in low HTra.

This possible explanation for the discrepancy in HTra values for L1J-to-English and L1J-to-Spanish participants no longer applies to L2J participants. These participants' first languages are much lower context than Japanese, and 日本を*nihon-wo* is a high-frequency concrete word, so should not require elevated cognitive effort to translate. However, this word had already appeared more than once in the source text. The default in English and Spanish to avoid repetition likely caused some participants to select an alternative to the literal translation. As a result, the into-English and into-Spanish HTra values were neither elevated nor low.

For our final example, we examine a *bunsetsu* where the HTra values are not highly unbalanced in the two target languages.

Example 4. HTra Values Balanced Across Languages, but Different for L1 Japanese and L2 Japanese The translations bunsetsu 日常の ことばというのは、nichijoo-no-kotoba-to-iu-no-wa (number 53 in Fig. 1 and Fig. 2) follow a more unusual HTra pattern. This *bunsetsu* appears in the sentence:

つまり、日常のことばというのは、普通であって、
詩のことばは何か特殊なことばであるという

Google translates this into English as:

In other words, ***everyday language*** *is an idea that is ordinary and that poetry is something special.*

In Spanish, the Google translation is:

En otras palabras, el ***lenguaje cotidiano*** *es una idea que es común y que la poesía es algo especial.*

Here the HTra values for L1 Japanese participants are much higher than those for L2 Japanese participants, regardless of the target language. In other words, L1 Japanese participants are expending significantly more cognitive effort than the English or Spanish speakers.

It is notable in this example that the *bunsetsu* has a proliferation of particles (というのは、*to-iu-no-wa*) modifying the basic concept (日常の ことば*nichijoo-no-kotoba*), literally *words of everyday.* The final sequence of particles does not have a direct equivalent in English or in Spanish, and all of the participants chose to omit them in their translations (Table 5).

It appears that the Japanese speakers struggled much more with this decision to omit than the English and Spanish speakers. The particles contribute to the meaning of the Japanese sentence in ways that have no parallel in English or Spanish. The English and Spanish speakers appear to reach the conclusion to omit the components that are superfluous in their native language, but it is not surprising that this same decision to omit is much more cognitively effortful for Japanese speakers, for whom the omitted components have critical significance in defining relationships between elements of a sentence, which are instantiated differently in English and Spanish.

8 General Conclusions and Future Directions

Prior research (Tokowicz 2014; Schaeffer et al. 2018; Carl and Baez 2019) has found that, when differing word translation entropy (HTra) levels are taken to indicate changes in cognitive effort, there are robust correlations of cognitive effort expended in translating the same source text from English to a variety of other languages, even languages that are quite remote from English. In this paper, we modified the HTra

Table 5 Translations of *bunsetsu* 53 with HTra values

日常の ことばというのは、nichijoo-no-kotoba-to-iu-no-wa **(everyday language)**	
L1J into English	**L1J into Spanish**
The language we use everyday	el lenguaje cotidiano
The words we use in daily conversation	Las palabras cotidianas
Daily use of language	Las palabras diarias
Words used in everyday life	el vocabulario de Vida cotidiana
Daily conversation	Las palabras cotidianas
Daily language	el lenguaje diario
Words in daily life	la Lengua diaria
HTra 2.81	**HTra 2.52**
L2J into English	**L2J into Spanish**
Everyday language	el Registro que usamos en la Vida diaria
Everyday language	el lenguaje cotidiano
Everyday language	el lenguaje cotidiano
Everyday language	el lenguaje diario
Everyday language	el lenguaje coloquial,
The words of everyday speech	el lenguaje cotidiano
	el lenguaje cotidiano
HTra 0.71	**HTra 1.66**

concept, substituting words by the more appropriate *bunsetsu* in Japanese, and, as hypothesized, found similar significant correlations persisted for translations from Japanese to English and to Spanish and that these were little affected by the first language of the translators. In an ANOVA, we also found that average expenditure of cognitive effort in *bunsetsu* translation was not significantly different across the target languages or the first languages of the translators. Since this was a small-scale study, it is important to replicate these findings in larger-scale studies and to seek to extend them to different language pairs.

We also found notable exceptions to the general patterns of correlation that could be attributed to target cultural differences, such as expressing information explicitly rather than making extensive inferences from context (Examples 1 and 3), or structural differences, such as the degree of reliance on grammatical inflections or the prevalence of subordinate clauses (Example 2). Example 4 also illustrated how structural differences can lead to differential expenditure of cognitive effort for translations into L1 and into L2. The greater structural complexity of Japanese seems to have caused L1 Japanese translators to expend more effort in simplifying to the English and Spanish translations, when compared to the effort expended by L1 English and L1 Spanish translators. Significantly, such variations in HTra, and so presumably cognitive effort, reflect differences in cognitive effort expended during the translation process manifested in the final translation product. This offers a novel way to investigate different stages of the translation process (in a stratificational view) or differences in the conceptual systems that are broadly, but not completely, shared across languages and cultures (in the nonselective activation view). It would

be interesting to undertake more tightly controlled investigations where such effects might be demonstrated experimentally and to seek a broader range of cultural and linguistic characteristics that trigger them. Ideally, such investigations should be conducted in a variety of language pairs to probe the robustness of this approach to studying cognitive effort in translation.

References

Almazroei SA, Ogawa H, Gilbert D (2019) Investigating correlations between human translation and MT output. Proceedings of the second MEMENTO workshop on modelling parameters of cognitive effort in translation production. European Association for Machine Translation, Dublin

Alves F (2003) Triangulation in process oriented research in translation. In: Alves F (ed) Triangulating translation: Perspective in process oriented research. Benjamins, Amsterdam

Angelone E (2010) Uncertainty, uncertainty management, and metacognitive problem solving in the translation task. In: Shreve GM, Angelone E Translation and cognition. Benjamins, Amsterdam.

Byrne J (2012) Scientific and technical translation explained. St. Jerome, Manchester

Carl M, Aizawa A, Yamada M (2016a) English-to-Japanese translation vs. dictation vs. post-editing: comparing translation modes in a multilingual setting. Proceedings of the NLP 2016 (language processing society 22nd annual meeting), Sendai (Japan), 7–10 Mar

Carl M, Schaeffer M, Bangalore S (2016b) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, Cham

Carl M, Schaeffer M (2017a) Models of the translation process. In: Schwieter J, Ferreira A (eds) The handbook of translation and cognition. Wiley Blackwell, Malden MA

Carl M, Schaeffer M (2017b) Sketch of a noisy channel model for the translation process. In: Hansen-Schirra S, Czulo O, Sascha H (eds) Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 71–116

Carl M, Baez M (2019) Machine translation errors and their relation to post-editing and from-scratch translation across different languages. J Spec Transl 31:107–132

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process. Transl Cogn Behav 2(2):211–232

Dijkstra T, Wahl A, Buytenhuijs F, Van Halem N, Al-Jibouri Z, De Korte M, Rekke S (2018) Multilink: a computational model for bilingual word recognition and word translation. Biling Lang Congn 13(3):1–23. https://doi.org/10.1017/S1366728918000287

Hall ET (1976) Beyond culture. Anchor Books, New York, NY

Halverson SL (2003) The cognitive basis of translation universals. Targets 15(2):197–241. https://doi.org/10.1075/target.15.2.02hal

Just MA, Carpenter PA (1976) Eye fixations and cognitive processes. Cognitive Psychol 8(4):441–480

Katan D (1999) Translating cultures: an introduction for translators, interpreters, and mediators. St. Jerome Publishing, Manchester

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. Kent State University Press, Kent, OH

Kudo T, Matsumoto Y (2002) Japanese dependency analysis using cascaded chunking in *CoNLL 2002: proceedings of the 6th conference on natural language learning 2002* (COLING 2002 post-conference workshops), pp 63–69

Lachaud CM (2011) EEG, EYE and KEY: three simultaneous streams of data for investigating the cognitive mechanisms of translation. In: O'Brien S (ed) Cognitive explorations of translation. Bloomsbury Publishing, London, pp 131–153

Lacruz I (2017) Cognitive effort in translation, editing and post-editing. In: Schwieter J, Ferreira A (eds) Handbook of translation and cognition. John Wiley & Sons, Malden, MA, pp 386–401

Lacruz I, Carl M, Yamada M, Aizawa A (2016) Pause metrics and machine translation utility. Proceedings of the NLP 2016 (language processing society 22nd annual meeting), Sendai (Japan), 7–10 Mar, pp 1213–1216

Lacruz I, Shreve GM (2014) Pauses and cognitive effort in post-editing. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation. Cambridge Scholars Publishing, Newcastle upon Tyne

López Guix JG, Minett Wilkinson J (2006) Manual de traducción. Gedisa Editorial, Barcelona

Rayner K, Pollatsek A (1989) The psychology of reading. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ

Schaeffer M, Carl M (2013) Shared representations and the translation process: a recursive model. Transl Interpreting Stud 8(2):169–190

Schaeffer M, Oster K, Nitzke J, Tardel A, Gros A, Gutermuth S, Hansen-Schirra S, Carl M (2018) Cross-linguistic (dis)similarities in translation: process and product. In: Granger S, Lefer M, Aguiar de Souza Penha Marion L (eds) Book of abstracts of the 5th edition of using corpora in contrastive and translation studies conference. Louvain-la-Neuve, 12–14 Sep 2018

Schilperoord J (1996) It's about time: temporal aspects of cognitive processes in text production, vol 6. Brill Rodopi, Amsterdam

Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mob Comput Commun Rev 5(1):3–55

Tokowicz N (2014) Translation ambiguity affects language processing, learning, and representation. In: Miller RT et al (eds) Selected proceedings of the 2012 second language research forum. Cascadilla Proceedings Project, Somerville, MA, pp 170–180

Tyler SW, Hertel PT, McCallum MC, Ellis HC (1979) Cognitive effort and memory. J Exp Psychol Hum Learn Mem 5:607–617

Translating Chinese Neologisms Without Knowledge of Context: An Exploratory Analysis of an Eye-Tracking and Key-Logging Experiment

Jinjin Chen

Abstract As would be intuitively expected, knowledge of context has a positive impact on the effort involved in translation. However, studies regarding how knowledge of context affects the grasp of meaning of words and how it influences the effort, translation strategy, and translation quality are still scarce. Our study seeks to explore how the absence of knowledge of context can be compensated for Chinese neologism translation utilizing eye-tracking and key-logging techniques along with a retrospective interview and holistic translation quality assessment. A pilot study was conducted among three groups of participants including one beginning translation student, one advanced translation student, and one professional translator. They were asked to perform three written from-scratch translations from Chinese to English, after which a retrospective interview was conducted to check their knowledge of context and their translation strategy for the neologisms. Various indicators of effort including ST and TT gaze measures and keystroke measures were analyzed and compared to the subjects' self-assessment. Our study is expected to help get an understanding of the following issues: (1) Does compensation for the absence of knowledge of context induce an increased effort in Chinese neologism translation? Is translation expertise related to effort? (2) What translation process, vertical or horizontal, is more triggered for this compensation? (3) What strategies do translators use for compensation in terms of different categories of Chinese neologisms? The recursive model of translation proposed by Schaeffer and Carl (Transl Interpreting Stud 8:169–190, 2013) was used to help explain our findings.

Keywords Knowledge of context · Effort · Neologism translation · Eye-tracking · Key-logging

J. Chen (✉)
Center for Studies of Translation, Interpreting and Cognition, University of Macau, Macau, SAR, China

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_12

1 Introduction

Neologisms, either rendered as "newly coined lexical units" or "existing lexical units that acquire a new sense," are "perhaps the non-literary and the professional translator's biggest problem" (Newmark 1988, 140). To date, most of the previous studies within the discipline of translation studies have attempted to explore the translation of neologisms predominantly from the perspective of translatability and translation output, while few empirical studies have been completed within a process-oriented framework. Shreve et al. (1993) is one of the process-oriented pioneers investigating particular linguistic-translation problems including neologisms, unusual collocations, and problematic phrasal units (idioms, figurative phrases) aiming to gain an insight into the way translators read for translation. However, still very few have attempted to explore the cognitive process of neologism translation.

The context, which neologism depends on, cannot be ignored in the study of it. The notion of context has always been regarded as central in many disciplines such as linguistics, pragmatics, or philosophy of language. In linguistics, mainly in textual linguistics, a distinction is made between linguistic context and extra-linguistic context. According to Newmark (1991), words exist not only in the context of their collocations, grammatical functions, or their positions in the sentence (linguistic context) but also in the context of the topic, real situation, or cultural background (extra-linguistic context). In pragmatics, context is divided into three focuses by Givón (1989), the generic focus, the deictic focus, and the discourse focus, referring to shared world and culture, shared speech situation, and shared prior text, respectively. From a cognitive perspective, van Dijk (2001a, b) argues that contexts are not social situations but mental constructs of participants. This view is advocated by a few scholars from relevance theory who claim that "[a] context is a psychological construct, a subset of the hearer's assumptions about the world. It is these assumptions . . . rather than the actual state of the world, that affect the interpretation of an utterance" (Sperber and Wilson 1986, 15).

Various approaches have been adopted to relate different perspectives on context to translation and interpreting research and practice. When Gutt (2000) applies the cognitive conception of context in the study of written translation, he insists that the relevance theory-based translation does not focus on the reproduction of words, linguistic constructions, or textual features, but on the comparison of interpretations. Setton (2006) and Mason (2006) also draw on relevance theory and apply it in interpreting studies, the former relating to conference interpreting and the later to dialogue interpreting. Setton (2006) believes that the context of simultaneous interpreting requires a significant modification of Gutt's model which applies to written translation, claiming that the core cognitive activities of simultaneous interpreting are not only less effortful than their counterparts in written translation but can also share resources, making their fusion into a unified cognitive activity. Mason (2006) provides vivid evidence of how interpreters engage in a process of joint negotiation of contextual assumptions. Diriker (2004), addressing context from a sociological perspective, states that simultaneous interpreters are

constrained by but, at the same time, constitutive of several interacting contexts ranging from the immediate context of utterance to the broad sociocultural context. These contexts are in a "mutually reflexive relationship" (Diriker 2004, 14). Baker (2006), viewing context from a dynamic perspective, discusses the active process of contextualization. Various examples of written translation, court interpreting, media interpreting, and subtitling were used in her discussion. Interestingly, she also shows how power shapes the context of interpretation in subtle ways.

Previous studies have noted that context is engaged in translation practice. However, the active processes of engagement have been studied rarely. Moreover, what these processes are like when context is absent or insufficient also deserves further study. Our study focuses on how the absence of knowledge of context can be compensated in the process of neologism translation. In the present study, the context we refer to is not the linguistic situation surrounding neologisms, but rather the cultural context in which the neologisms are embedded.

2 Knowledge of Context in Translation and Interpreting Studies

There has been an increasing concern on the role that knowledge of context plays in translation and interpreting studies. Knowledge of context, which refers to the extra-linguistic context, has been studied in the form of speech transcripts, summaries, briefing, PPT slides, background information checking, and the like.

In the field of translation research, different forms of background information accessibility are usually compared to explore the influence of knowledge of context on the process of translation or its results. Griffin (1995), carrying out a within-subject experiment, examines the translation performance of ten professional translators by providing background information under two conditions: the first is with two related and the other two unrelated background texts. Production times, correctness, and appropriateness are quantified in this study showing that background information can effectively improve the quality of translation, but result in longer production duration. Kim (2006), using a between-subject design, analyzes the impact of quantity and quality of background information by asking 16 undergraduate students to research the background on the translation topic prior to the translation tasks and the other 16 undergraduate students to only check the dictionary during translation to complete the identical task. This study finds that translation quality is markedly impacted by the quality of the background information but is hardly influenced by its quantity.

In interpreting practice, the availability and the acquisition of contextual knowledge are recognized by the interpreters as an important part of their working conditions (Gile 1995, 2002; Diriker 2004). Despite the importance of preparation addressed by scholars, research on how preparation affects interpreting has been impeded to a large extent by high variability and sensitivity of measures and tasks

(Gile 2005). However, still a few studies in this regard have been carried out to date. The construct of contextual knowledge is often operationalized in various kinds of in-advance preparation. Díaz-Galaz (2011), targeting 14 advanced undergraduate students, explores the effect of in-advance preparation in SI (simultaneous interpreting) of specialized speeches. Two comparable speeches were used in preparation and non-preparation conditions. EVS (ear-voice span), translation accuracy, and percentage of omission were analyzed in both conditions, finding that although EVS was slightly longer in the preparation than non-preparation condition, translation accuracy was improved, and translators were failing less easily for difficult segments after preparation. Díaz-Galaz et al. (2015) extend their previous study by comparing the behavior of seven professional interpreters and 16 interpreting students in SI with/without preparation. Both "neutral" and "difficult" speech segments were inserted in the speeches. At last, the improved accuracy and a shorter EVS reveal that both groups perform significantly better in the in-advance preparation condition: for the inexperienced translators, this was evident in all the difficult segments with terminology, complex syntactic structures, and nonredundant elements; for experienced translators, only nonredundant elements were processed better in the preparation condition.

The aforementioned observation was not consistent with the results of some empirical studies from Anderson (1979) in which no observable effect of contextual knowledge was found on the performance of professional interpreters. In a between-subject experiment, 12 professional English-French conference interpreters participated in a simultaneous interpreting task under three conditions: a "written text" condition, a "summary" condition, and a "no information" control condition. The intelligibility and informativeness of interpreting output were measured as dependent variables in relation to interpreting performance. The results indicated no significant effect of prior information about the content of the speech on either performance measures. The author was somewhat surprised about the "no effect" result giving the possible explanation that the great subject and inter-passage variability, as well as the small sample size, may overshadow the real effect. Moreover, the author also explained that subjects with no provided information were not completely uninformed since they already knew the topic of speech beforehand, and the speech was not so complexed that the background knowledge may contribute little to their performance. Lamberger-Felber and Schneider (2008) address the effect of transcript availability on linguistic interference. Linguistic interferences are "those instances of deviation from the norms of the language which occur in the speech of bilinguals as a result of language contact" (Weinreich 1953, 15). The authors made two hypotheses: one is that linguistic interference is more frequent in SI with text than without, and the other is that a "prepared manuscript" condition generates fewer instances of linguistic interference than an "unprepared manuscript" condition. It is very interesting that results do not verify the first hypothesis unequivocally and even show the exact opposite trend as the second hypothesis suggests.

Particular attention is paid to the cognitive impact of cultural background knowledge (CBK) on processing metaphors. Zheng and Xiang (2013, 2014) and

Xiang and Zheng (2011, 2015) have been immersed in a series of English-Chinese sight translation experiments, directly touching upon the cognitive effort required for metaphorical and non-metaphorical expressions in CBK and no CBK-provided condition. These studies conclude that cultural background knowledge, to a great extent, lowers the cognitive load imposed by metaphorical expressions and therefore betters the translation process indicated by reduced processing time and improved translation quality. They also correlate cultural background knowledge with translation strategies, finding a negative correlation between the frequency of using omission as a coping strategy and the acquisition of cultural background knowledge.

Think-aloud protocols have been used as a method to investigate the translation process of metaphorical expression, metonymic expression, idioms, and the like, as well as the cultural knowledge required in this process. Jensen, in identifying how professional translators and non-professional translators cope with metaphors and metonymic expressions as a problem, concludes that "translating metaphor and metonymic expressions requires knowledge of source domains and target domains of two cultures" (2005, 189).

Despite the fruitful results of empirical research on the above topics, still, several issues centering on these topics are short of discussion. Firstly, as would be intuitively expected, knowledge of context has a positive effect on the effort involved in translation. However, to what degree knowledge of context affects the effort is under-researched. Secondly, even though some research has been published to explore the influence of knowledge of context on word translation, there is little evidence of the role of context playing on the effort in neologism translation. Neologisms are unique in that they are closely connected with new things and the changes in the society; therefore, there is no immediate translation equivalent available. Thirdly, the primary focus of existing studies on this topic is European-languages driven, but Chinese and English combination, a pair of two typologically distant languages, lacks this exploration.

In terms of research design or methodology, most of the previous experiments, especially on written translation, are common in that the knowledge of context or background information is provided to translators beforehand in the form of a summary or a related text, which is not naturalistic in a real written translation scenario. Moreover, normal texts, or certain kinds of specialized texts, scientific and technical, are often used as experimental materials, which may result in a biased finding not applying to other text genres. Additionally, although sophisticated technology such as eye-tracking and key-logging have been used to explore the translators' behavior, a little exploration of the role of knowledge of context by this technology is found.

3 A Recursive Model of Translation

It has been a controversial discussion whether translation is a vertical or a horizontal process (De Groot 1997). In the vertical perspective, translation is composed of two monolingual systems: one for understanding the source text and the other for reformulating the captured meaning in the target language. In contrast to the vertical view, the horizontal view proposes that the target language reformulation commences during the source text comprehension. These two languages are linked via shared representations.

Based on these theories, Schaeffer and Carl (2013) propose a different kind of model, a recursive model, to describe the process of translation. They argue that translation involves both vertical and horizontal processes which are always active at the same time. In other words, they don't assume separate input and output lexicons for source and target language. The vertical process serves as the monitor for target text reformulation by the horizontal process.

Further, Carl et al. (2019) point out that translation can be understood as a process, composed of interacting word and phrase translation systems, which function as dissipative structures. Based on evidence from bilingualism research (Dijkstra 2019), when a word in the source language is read, similar words are automatically activated in both the source and target languages. This activation is a nonselective subliminal process, which involves stimulation of shared semantic representations between two languages. Gazing activities and pause analysis of keystrokes can be used to measure the effort a translator spends to activate and integrate the word and phrase translation systems during the translation process. Wei (this volume, Chap. 7) provides a detailed analysis of such an activation and integration process in the translation of a difficult metaphorical expression.

Based on this translation model, we would like to assess whether more effort is spent on CNEO (neologisms with context) translation or NNEO (neologisms without context) translation, as measured by gaze and keystrokes as well as translators' subjective rating for their effort. Moreover, we hope to generate more insights on the translation process (vertical or horizontal) involved in CNEO and NNEO translation. In addition, we expect to find clues of translation strategies used for the compensation for the absence of knowledge of context in terms of different categories of Chinese neologisms.

4 Methodology

4.1 Participants

In the pilot study, three groups of people were invited to the experiment, including one beginning translation student (P02), one advanced translation student (P03), and one professional translator (P01). The beginner is a third-year undergraduate

Table 1 Text profiling

Parameters/stimuli	Text 1	Text 2	Text 3
No. of characters	168	173	172
Average sentence length	32.60	34.60	33.40
No. of neologisms	7	5	7

majoring in English and receiving half a year translator training, while the advanced is a second-year master's student with 1.5 years of formal translator training and 1.5 years of part-time translation experience. The professional translator has received 5 years of formal translator training in his bachelor and master studies and has a translation experience of 3 years and 500,000 words. All of them have Chinese as their L1 and translated primarily into English as their L2 and have a similar level of language proficiency in English which equates the IELTS score of 7–7.5. They have received initial training in order to be familiar with the experimental environment.

4.2 Stimuli

Three Chinese texts were selected as stimuli of the experiment, each consisting of around 170 characters and 5–7 neologisms. The neologisms were chosen according to the following criteria: they are either (1) "newly coined lexical units" or (2) "existing lexical units that acquire a new sense." For example, "雄安新区" [Xiongan New Area] belongs to "newly coined lexical units," which was made of "雄安" [Xiongan] and "新区" [New Area]. "放水养鱼" [using accommodative measures] belongs to "existing lexical units that acquire a new sense." The term is a popular idiom in China but has a new meaning, which refers to an economic policy.

They fell into three domains including news, economics, and sci and tech. Different domains are taken in our study trying to minimize a biased finding that might be caused by a certain kind of genre. Text 1 was a news report from http://www.chinanews.com/ on 18 July 2017, which introduced the top ten neologisms of Chinese media in 2016. Text 2 was about economics and Text 3 about sci and tech, both of which came from different parts of *Report on the Work of the Government* in 2018 introducing the government's work for the past 5 years and laying out the proposal for the present year's work in the area of economics and sci and tech, respectively. An online tool for testing the readability of international Chinese texts, https://www.languagedata.net/editor/, was used to calculate their total number of characters and average sentence length (see Table 1).

These texts were presented statically in the source window of Translog-II (Carl 2012a, b) with 20-point Song font size and 2-line spacing on a 23" LCD monitor at 1920 × 1080 pixels. To ensure the accuracy of the data, each text only covered one page in Translog-II so that participants didn't need to scroll the interface while reading and translating. The source window of Translog-II at the left side

was presented with the source text, while the target window at the right side was left blank for participants to type their translation. Tobii TX300 was connected to Translog-II so that we had automatic gaze-to-word mapping.

4.3 Procedure

Before the experiment, participants were asked to sign a consent form and fill in their personal information and were informed of the procedure of the experiment. In this part, the researcher also led the participants to get familiar with the equipment and gave clear instructions to them.

Participants conducted a written translation of four texts from Chinese to English without time constraint, including one warm-up text and three main texts. These texts were presented in the following the order: the warm-up text, Text 1, Text 2, and Text 3. Participants sat in front of the screen at an around 60 cm distance throughout the experiment and went through the calibration with a standard 9-point grid before each of the four translation tasks.

After their translation, translators were instructed in a retrospective interview to explicate whether they had a knowledge of the context of each neologism. "Knowledge of context" in our study refers to background knowledge or the prior knowledge acquired before the experiment.

In principle, translators may encounter four possibilities including:

1. Translators know the meaning and have knowledge of context of the neologisms.
2. Translators know the meaning but do not have knowledge of context of the neologisms.
3. Translators have knowledge of context but do not know the meaning of the neologisms.
4. Translators neither know the meaning nor do they have knowledge of context of the neologisms.

To fulfill the objectives of our study, we only compare possibilities (1) and (2). (1) was marked as "neologism with context" (CNEO), while (2) was labeled as "neologism without context" (NNEO) for the sake of our future analysis.

In marking each neologism, we followed the following criteria of "knowing the meaning" and "having knowledge of context." Knowing the meaning means knowing the linguistic knowledge, such as the word form, that is, how each unit is formed into a word. Having knowledge of context means knowing the origin or the usage of the word. For example, knowing the meaning of "雄安新区" [Xiongan New Area] means knowing that "雄安" [Xiongan] is the modifier and "新区" [New Area] is the head. Having knowledge of context also means knowing that "雄安新区" [Xiongan New Area] was established in April 2017. The area has the main function to serve as a development hub for the Beijing-Tianjin-Hebei (Jingjinji) economic triangle.

Participants were also asked to rate the effort from 0 to 10 (0 means the lowest effort and 10 means the highest effort) in conducting the translation task of each neologism. In addition, they were asked to recall their translation strategy in dealing with these neologisms.

4.4 Data Processing

This study used the CRITT TPR-DB method to process the eye-tracking and key-logging data recorded. CRITT TPR-DB is a publicly available database of recorded text production (copying, editing, post-editing, and translation) sessions for TPR, containing UAD (user activity data) of both the process and product components recorded with Translog-II and the CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation) workbench (Carl 2012a, b). The raw logging data generated from the recordings can be further annotated and processed into 11 tables that can be easily processed by various visualization and analysis toolkits (Carl et al. 2016).

We went through the following steps to convert and align the data:

1. Uploading the original logging file generated from Tranglog-II to the TPR-DB.
2. Manually aligning the target text with the source text at the word or phrase level on YAWAT (Yet Another Word Alignment Tool), a browser-based TPR-DB management tool for the visualization and creation of word-or-phrase-level alignments (Germann 2008).
3. Downloading the tables containing gaze and typing data processed by the TPR-DB for analysis. In our analysis, we focus on the ST tables.

5 Effort for CNEO Translation and NNEO Translation

5.1 Objective Measures of Effort

The objective measures of effort for word translation production include ST gaze measures such as TrtS, FixS, FPDurS, and FFDurS, TT gaze measures such as TrtT and FixT, and keystroke measures such as Dur, Ins, and Del (see Table 2). A basic assumption (the so-called eye-mind assumption) in eye movement research is that "the eye remains fixated on a word as long as the word is being processed" (Just and Carpenter 1980, 329). Therefore, we used ST gaze measures and TT gaze measures as indicators of processing effort. Moreover, first fixations are considered to be indicative of early (lexical) processing (Rayner 1998). Therefore, we used FPDurS and FFDurS in ST gaze measures as indicators of early processing effort. In addition, we used keystroke measures as indicators of performance effort.

Table 2 Objective measures of effort for word translation production

Measures	Explanation
TrtS	Total reading time/fixation duration on ST
FixS	Fixation count on ST
FPDurS	First pass duration on ST
FFDurS	First fixation duration on ST
TrtT	Total reading time/fixation duration on TT
FixT	Fixation count on TT
Dur	Translation production duration
Ins	Number of insertions
Del	Number of deletions

The values of the measures in Table 2 were normalized in the following way. The ST gaze measures were divided by the number of Chinese characters. TT gaze measures and keystroke measures were divided by the number of English letters. TrtSNor, FixSNor, FPDurSNor, and FFDurSNor represented the normalized ST gaze measures, TrtTNor and FixTNor stood for the normalized TT gaze measures, and DurNor, InsNor, and DelNor were used as the normalized keystroke measures. In the following analysis, we present results for the three participants individually and report trends with respect to the measures and the two conditions (CNEO and NNEO). Due to the limited amount of data, we do not present a detailed statistical analysis.

5.1.1 ST Gaze Measures

The ST gaze measures for CNEO and NNEO for three translators are illustrated in Fig. 1. All of them spent much more time and fixed many more times reading NNEO compared to reading CNEO on the ST as indicated by TrtSNor and FixSNor. The results reveal that these three translators allocated much more processing effort on the ST for NNEO than for CNEO.

Moreover, the gap between NNEO and CNEO in TrtSNor and FixSNor is bigger for the beginner than for the advanced and the professional. The results show that knowledge of context helped the beginner reduce the effort more than it helped the advanced and the professional.

5.1.2 TT Gaze Measures

Figure 2 displays the TT gaze measures for CNEO and NNEO for three translators. An opposite trend was observed in the TT gaze measures compared to the ST gaze measures. All the TT gaze measures for these three translators show longer fixation duration and more fixations on the TT of CNEO than on NNEO. These results indicate that these three translators experienced more cognitive load in the TT of CNEO than in NNEO.

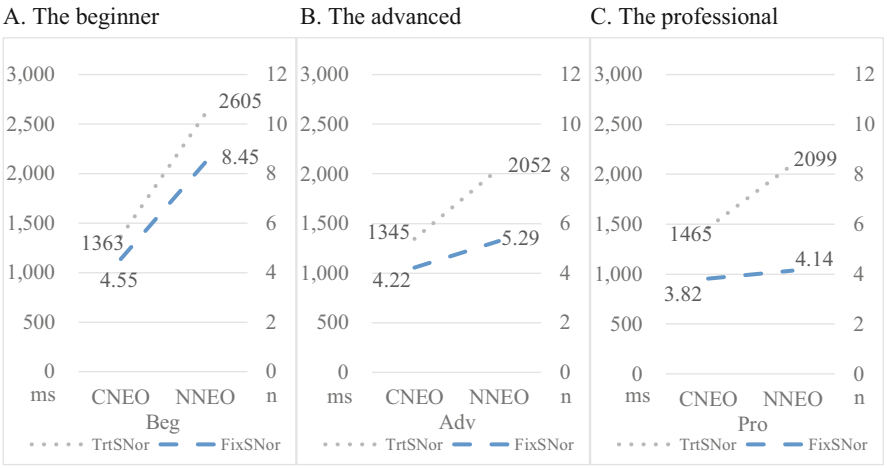


Fig. 1 ST gaze measures for CNEO and NNEO for the three translators. (**a**) The beginner. (**b**) The advanced. (**c**) The professional

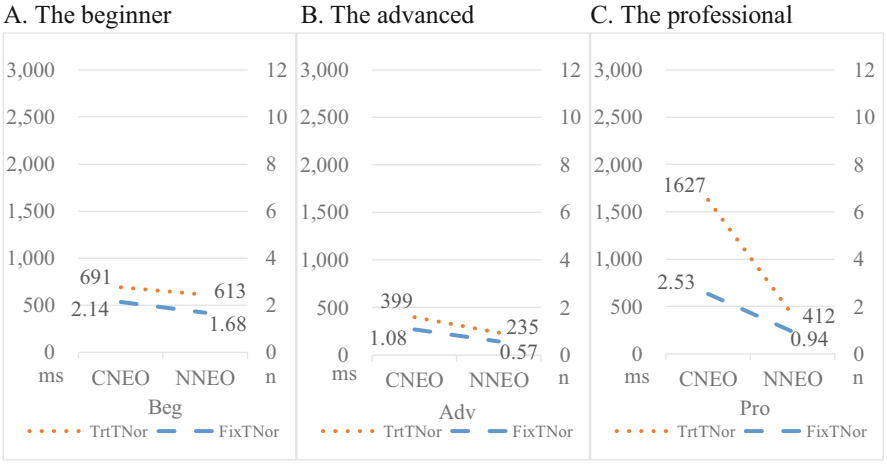


Fig. 2 TT gaze measures for CNEO and NNEO for the three translators. (**a**) The beginner. (**b**) The advanced. (**c**) The professional

Moreover, the gap between NNEO and CNEO in TrtTNor and FixTNor is bigger for the professional than for the advanced and the beginner. The results show that for the professional, knowledge of context led to more TT processing effort than for the advanced and the beginner.

5.1.3 Keystroke Measures

The keystroke measures for CNEO and NNEO for the three translators are presented in Fig. 3. Keystroke measures DurNor, InsNor, and DelNor show a different trend compared to ST gaze measures TrtSNor and FixSNor. Overall, longer duration, more insertions, and deletions were found on CNEO than on NNEO, indicating that more performance effort was allocated on CNEO than on NNEO.

Moreover, the gap between NNEO and CNEO in DurNor, InsNor, and DelNor is bigger for the professional than for the beginner and the advanced. The results show that for the professional, knowledge of context resulted in more performance effort than for the advanced and the beginner.

5.2 *Subjective Assessment of Effort*

Participants were also asked to rate the perceived effort from 0 to 10 (0 means the lowest perceived effort and 10 means the highest perceived effort) in conducting the translation task of each neologism. We used 0 representing CNEO condition and 1 representing NNEO condition in the following analysis. Table 3 shows the correlation between knowledge of context and subjective assessment of effort.

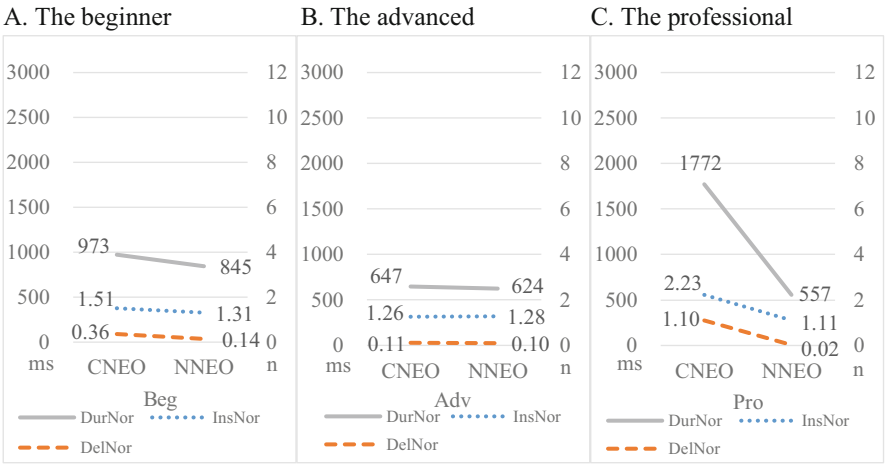


Fig. 3 Keystroke measures for CNEO and NNEO for the three translators. (**a**) The beginner. (**b**) The advanced. (**c**) The professional

Table 3 Correlation between knowledge of context and subjective assessment of cognitive effort

	Knowledge of context (beginner)	Knowledge of context (advanced)	Knowledge of context (professional)
Perceived effort	−0.44	−0.10	−0.07

The negative correlation indicated by the Spearman coefficient shows that knowledge of context helped three participants reduce the effort. However, the influence of knowledge of context on perceived effort was larger for the beginner than for the advanced and the professional. For the beginner, the perceived effort was greatly influenced by the familiarity of context, indicated by a relatively high and negative correlation (Spearman coefficient = −0.44) between effort and context. For the advanced and the professional, the effort had little to do with knowledge of context indicated by a very low value in coefficients of Spearman (−0.10 for the advanced; −0.07 for the professional).

5.3 *Early Processing Effort and Late Processing Effort for CNEO Translation and NNEO Translation*

Here, we used FPDurS and FFDurS as indicators of early processing effort and TrtSNor and FixSNor as indicators of late processing effort. Measures of early processing effort and late processing effort for CNEO and NNEO for the three translators are displayed in Fig. 4.

Longer fixation duration and more fixations were observed on the ST for NNEO than for CNEO, while shorter first ST pass duration and first ST fixation duration were found on NNEO than on CNEO. The results reveal that these three translators allocated more late processing effort on NNEO than on CNEO but allocated less early processing effort on NNEO than on CNEO.

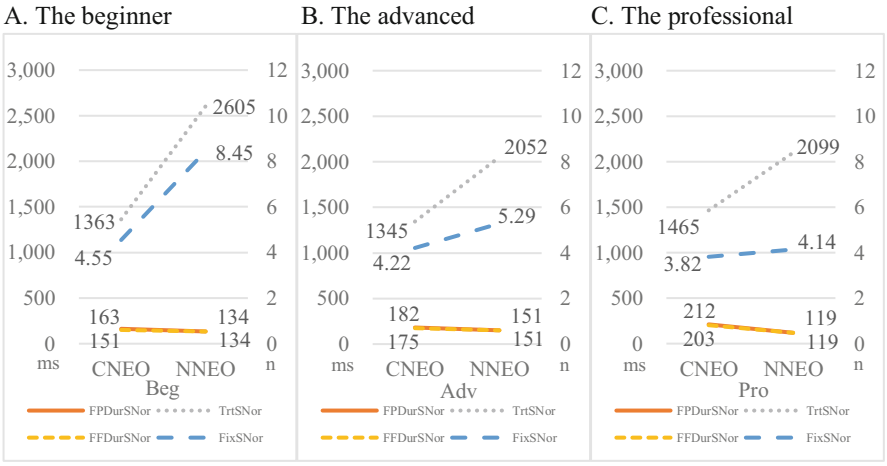


Fig. 4 Measures of early processing effort and late processing effort for CNEO and NNEO for the three translators. (**a**) The beginner. (**b**) The advanced. (**c**) The professional

6 Translation Strategies in Dealing with Different Categories of NNEO

In this section, we present a qualitative analysis regarding how translators can compensate if they do not have knowledge of context.

After their translation, these three translators were asked to explicate what strategy they used in coping with each neologism. For the sake of analysis, roughly two strategies were identified from their retrospect including literal translation and paraphrase. Here we used Kielar's (2013) definition of literal translation. According to this definition, in literal translation, the rules of syntax of the target language are used to combine the words calqued from the source language as separate lexical units. Non-literal translations were marked as paraphrase in our study. After the experiment, the experimenter judged the translation quality of the three translators, categorizing their renderings into successful and unsuccessful translation.

Moreover, different categories of NNEO were sorted out for the sake of exploring the translation strategy made by the translators when lacking knowledge of context (see Table 4). The criteria of categorization of NNEO were based on the meaning of the NNEO.

6.1 Translating S-NNEO

S-NNEO refers to the neologism whose meaning is very straightforward and therefore is easily known to the participant, but background knowledge is unknown to the participant.

Take "雄安新区" [Xiongan New Area] as an example. The participant may know the meaning of the newly coined lexical units, "雄安新区" [Xiongan New Area], very well because "雄安" [Xiongan] is the modifier and "新区" [New Area] is the head. However, the participant may not have heard of the new area itself, which means he/she lacks the background knowledge of it, such as how and why it was

Table 4 Category of NNEO

Category	NNEO
Straightforward (S-NNEO)	山寨社团[fake social organization] 品质革命[quality revolution] 最多跑一次[without the need for a second trip] 雄安新区[Xiongan new area] 中国制造2025 [Made in China 2025]
Event-related (ER-NNEO)	闺蜜门[Choi-gate scandal]
Idiomatic (I-NNEO)	放水养鱼[using accommodative measures]
Ambiguous (A-NNEO)	区间调控[range-based regulation] 融通创新[collaborative innovation]
Ellipse (E-NNEO)	双创[business startup and innovation]

Table 5 Successful (+) and unsuccessful (−) translations (Column T) following a literal (L) and paraphrase (P) translation strategy (column S) for S-NNEO from the professional (Pro), advanced (Adv), and beginner (Beg) translators (column TS)

ST	TT	S	T	TS
山寨社团[fake social organization]	Faked starts group	L	+	Pro
	Fake social clubs	L	+	Beg
	Copying society	L	−	Adv
品质革命[quality revolution]	Revolutional progress on our quality	L	+	Beg
最多跑一次[without the need for a second trip]	Prevent you from coming twice	L	+	Beg
雄安新区[Xiongan new area]	Xiong an new area	L	+	Adv
中国制造2025[Made in China 2025]	Made in China 2025	L	+	Adv

built and where it is. Table 5 shows the translation quality, translation strategy, and translation output for S-NNEO.

For S-NNEO, participants tended to adopt literal translation. This kind of translation strategy in dealing with S-NNEO often resulted in successful translation. There is only one exception, which is the TT "copying society" of the ST "山寨社团" [fake social organization] made by the advanced translator. She tried to translate the coined lexical units unit by unit, but chose the wrong target words, resulting in unsuccessful translation.

Seeing the translation output, translation quality, and translation strategy, knowledge of context seems not to have an obvious effect on S-NNEO translation.

6.2 Translating ER-NNEO

ER-NNEO refers to the neologism whose meaning is event-related. The participant may partly know the lexical meaning of the linguistic sign, which was related to an event, but not know the referential meaning of the event. Moreover, the background knowledge was unknown to the participant.

Take "闺蜜门" [Choi-gate scandal] as an example. It refers to an event in which Choi Soon-sil, a confidante of President Park Geun-hye, was interfering in politics. The participant may partly know the meaning of these newly coined lexical units because "门" [scandal] was often used to describe a scandal or an event. Table 6 shows the translation quality, translation strategy, and translation output for ER-NNEO.

For ER-NNEO, participants tended to adopt literal translation. In the interview, participants claimed that when encountering these kinds of words, literal translation may be a "safer" strategy than paraphrase, since paraphrase may induce wrong interpretation of the meaning of the word and result in unsuccessful translation. However, these kinds of renderings, such as "Likable Sisters" and "the affair of so-

Table 6 Successful (+) and unsuccessful (−) translations (column T) following a literal (L) and paraphrase (P) translation strategy (column S) for ER-NNEO from the professional (Pro), advanced (Adv), and beginner (Beg) translators (column TS)

ST	TT	S	T	TS
闺蜜门[Choi-gate scandal]	Likable sisters	L	–	Pro
	The affair of so-called best friend	L	–	Adv

called best friend," were regarded as unsuccessful translations, since they cannot be understood by the target reader.

The participants also claimed that if they knew the background knowledge of the event, they would have chosen to paraphrase to better convey the meaning of the neologism. For example, a participant who knew the background knowledge of the "闺蜜门" [Choi-gate scandal] would translate it into "Korean president's scandal," which was regarded as successful translation.

Knowledge of context for ER-NNEO helps translators grasp the referential meaning of the event and therefore has a better interpretation of neologisms. Seeing the translation output, translation quality, and translation strategy, knowledge of context may have a positive effect on ER-NNEO translation.

6.3 Translating A-NNEO

A-NNEO refers to a neologism whose meaning is ambiguous because of the parsing of the term (structure-based) or the polysemy of the term (meaning-based). The participant may vaguely understand the meaning and may face difficulty in filtering out the inappropriate meaning because background knowledge was unknown to the participant.

Take "区间调控" [range-based regulation] as an example. The term is ambiguous in terms of parsing due to a lack of connectives, which can mean either "the regulation of range" or "regulation within a range." The participant may vaguely know the meaning of these newly coined lexical units, but lack the background knowledge of it. The background knowledge of it may include the tenet of this economic term, which is to stimulate vitality, strengthen weak links and the real economy, and focus on key areas and weak links in economic and social development.

"融通创新" [collaborative innovation] is also an A-NNEO because part of the lexical unit "融通" may have several meanings, "consilience," "accommodation," or "financing." Table 7 shows the translation quality, translation strategy, and translation output for A-NNEO.

For A-NNEO, participants tended to adopt literal translation. Since the meaning was ambiguous for this category of neologisms, translators easily chose the wrong meaning of the term resulting in an unsuccessful translation. For example, the advanced translator gave the TT "control it section by section" for the ST "区间

Table 7 Successful (+) and unsuccessful (−) translations (column T) following a literal (L) and paraphrase (P) translation strategy (column S) for A-NNEO from the professional (Pro), advanced (Adv), and beginner (Beg) translators (column TS)

ST	TT	S	T	TS
区间调控[range-based regulation]	Phrase-coordinating	L	−	Pro
	Partly adjust the economic system with the help of the government	P	−	Beg
	Control it section by section	L	−	Adv
融通创新[collaborative innovation]	Combine the new finding of	L	+	Beg
	The innovation	L	−	Adv

调控" [range-based regulation] because she chose the wrong interpretation of the ST "the regulation of range" instead of the correct one "regulation within a range." There is only one exception, which is the TT "Combine the new finding of" of ST "融通创新" [collaborative innovation] made by the beginner. In the interview, he claimed that he happened to choose the correct interpretation of ST.

Knowledge of context for A-NNEO helps the translator to filter out the inappropriate meaning of neologisms. Seeing the translation output, translation quality, and translation strategy, knowledge of context may have a positive effect on A-NNEO translation.

6.4 Translating I-NNEO

I-NNEO refers to the neologism that is often used idiomatically. The literal meaning and idiomatic meaning were often well-known to the participant because the idiom is popular. However, the actual meaning of the idiom under the new condition was not fully understood by the participant because the background knowledge was unknown.

Take "放水养鱼" [using accommodative measures] as an example. The term was often used idiomatically, which has the literal meaning of "add water and raise fish" and the idiomatic meaning of "keep a long-term point of view." Under the new condition, the neologism refers to an economic policy. When the state implements the existing tax and fee reduction policies, it strives to improve the tax system and studies new measures to further reduce the burden on enterprises. The participant may partly know the meaning of the existing lexical units, but lack the background knowledge of the new sense. Table 8 shows the translation quality, translation strategy, and translation output for I-NNEO.

For I-NNEO, one participant adopted literal translation based on the originally literal meaning of the term, while the other chose paraphrase based on the originally idiomatic meaning of the term. Both of their renderings are unsuccessful. Neither of them knew the background knowledge of the term which has the new meaning and may help them better understand the new sense of the term.

Table 8 Successful (+) and unsuccessful (−) translations (column T) following a literal (L) and paraphrase (P) translation strategy (column S) for I-NNEO from the professional (Pro), advanced (Adv), and beginner (Beg) translators (column TS)

ST	TT	S	T	TS
放水养鱼[using accommodative measures]	Avoiding damages to the potential of our future economic growth	P	–	Beg
	Raising fish by giving away water	L	–	Adv

Table 9 Successful (+) and unsuccessful (−) translations (column T) following a literal (L) and paraphrase (P) translation strategy (column S) for E-NNEO from the professional (Pro), advanced (Adv), and beginner (Beg) translators (column TS)

ST	TT	S	T	TS
双创[business startup and innovation]	Creating new scientific developments and new business	P	+	Beg
	Innovation and start-up	P	+	Adv

Knowledge of context for I-NNEO helps the translator grasp the new sense of the idiomatic neologisms. Seeing the translation output, translation quality, and translation strategy, knowledge of context may have a positive effect on I-NNEO translation.

6.5 Translating E-NNEO

E-NNEO refers to the neologism that has the ellipsis. The meaning of the neologism was partly known to the participant, but the background knowledge of it was unknown to the participant.

Take "双创" [business startup and innovation] as an example. The term was literally regarded as "double chuang," meaning "business startup" [chuang ye] and "Innovation" [chuang xin]. The participant may vaguely know that this term is related to two kinds of "创" [chuang], but may not know what the exact two "创" [chuang] are because they lack the background knowledge of the term. Table 9 shows the translation quality, translation strategy, and translation output for E-NNEO.

For E-NNEO, participants tended to adopt paraphrase. In the interview, it is interesting to find that even the participants did not know the background knowledge of the new term "双创" [business startup and innovation]; they successfully guessed the meaning of the new term from its co-text (the linguistic situation surrounding a word). This part, which "双创" was located in, was about the collaboration between enterprises, universities, and research institutes, so they successfully guessed the meaning of "双创" related to business startup [chuang ye] and innovation [chuang xin].

It is known from the interview that the intra-linguistic context serves as the function of background knowledge on this occasion, which helped participants fully understand the "ellipse neologisms." Seeing the translation output, translation quality, and translation strategy, knowledge of context, in some sense, seems to have a positive effect on E-NNEO translation.

7 Discussion

The ST gaze measures, including fixation duration and fixation count, show that NNEO translation was cognitively more effortful than CNEO translation. According to Schaeffer and Carl (2013), translation involves the activation of both source and target items which share one single cognitive representation. More importantly, early during source text reading, the shared representation is activated which then serves as a basis for regeneration in the target language. Therefore, the ST gaze measures suggest that it is early during source text reading that NNEO requires more effort than CNEO. Moreover, our finding may show that the activation of source and target languages is more rapid in CNEO translation than NNEO translation. This finding could seek theoretical support from Kintsch's construction-integration model (1988) and Gernsbacher's structure-building framework (1996) in which they put forward that the rapid activation of knowledge of context stored in long-term memory can facilitate the interpretation of linguistic cues and thus can save effort for translation.

This finding is confirmed in the subjective assessment of effort, which shows that knowledge of context helped translators reduce the effort of neologism translation.

Interestingly, this facilitation effect is different for the three participants in our study. Knowledge of context helped the beginner reduce effort more than it helped the advanced and the professional. One explanation is that in NNEO translation, the beginner tended to dwell on the words for a long time, while the advanced and the professional did not linger on these words for a long time. The two more experienced translators made decisions faster than the beginner, even when they lack knowledge of context.

The TT gaze and keystroke measures may offer a different perspective in understanding how translators spend their effort in TT processing and performance phase. It seems that NNEO attracted less effort as compared to CNEO in this regard. This may be due to translators' distribution of effort. They tended to spend more effort on ST processing and relatively less effort on TT processing and TT production. In line with Paradis (1994, 321), the recursive model predicts that during concurrent reading and writing, the activation threshold for both source and target language is similarly high, while the activation threshold for the non-active language is higher than for the active language when reading and writing do not occur simultaneously. This prediction gives us hints to explore the activation in more details by looking into the AUs (Schaeffer et al. 2016) in the future, which could provide abundant information for different translation activities, such as "translation

typing while reading the source text," "translation typing while reading the target text," and "translation typing while reading the source and the target text."

Overall, in relation to our research question (1), objective measures of effort and subjective assessment of effort for CNEO and NNEO translation indicate that compensation for the absence of knowledge of context induces an increased effort in Chinese neologism translation. The compensation triggers more effort for the beginner than for the advanced and the professional.

Measures of early processing effort and late processing effort reveal that these three translators allocated more late processing effort on NNEO than on CNEO, but allocated less early processing effort on NNEO than on CNEO. According to the recursive model, translation involves both vertical and horizontal processes. The horizontal process is an early process, while the vertical process often advances later. The above results may indicate that CNEO translation triggers horizontal processes more, while NNEO translation triggers vertical translation processes more. Knowledge of context comes in play in the early processing stage.

As for our research question (2), these indicators reflect that the compensation for the absence of knowledge of context triggers a more vertical process than a horizontal process.

Regarding our research question (3), it can be seen from the translation output and retrospective interview that in NNEO translation, different translation strategies come in play in terms of different categories of Chinese neologisms in order to compensate for the absence of knowledge of context. The recursive model sees the horizontal process as a default mode and the vertical process as a monitor accessing the output from the horizontal automatic process, especially when translation departs from formal correspondence. The monitor role is often related to conscious processing, which is meaning-based. This theory could give support to our findings that translators made the compensation consciously in NNEO translation.

Interestingly, for I-NNEO, translators tended to adopt paraphrase after retrieving the meaning from the in-text context. According to the recursive model, the vertical process depends on the in-text context which becomes available later, as processing advances in the chunk or text. Our finding, therefore, reveals that the compensation advances in the vertical translation process particularly when translators retrieve the meaning from the in-text context.

8 Concluding Remarks

Our study seeks to investigate the impact of knowledge of context in Chinese neologism translation by using eye-tracking and key-logging techniques along with a retrospective interview and holistic translation quality assessment. We intend to find out how the absence of knowledge of context can be compensated for Chinese neologism translation. A pilot study with three translators was presented in this chapter.

Our main findings are:

1. Overall, compensation for the absence of knowledge of context induces an increased effort in Chinese neologism translation as revealed by objective measures of effort and subjective assessment of effort. The compensation triggers more effort for the beginner than for the advanced and the professional.
2. In terms of the translation process, the compensation triggers a more vertical process than a horizontal process as revealed by the indicators of early processing effort and those of late processing effort.
3. Different translation strategies are used for compensation in terms of different categories of Chinese neologisms as revealed by the translation output and the retrospective interview. For example, participants tended to adopt literal translation for S-NNEO, ER-NNEO, and A-NNEO and paraphrase for I-NNEO. There is no preferable strategy for I-NNEO.

Our study may have the following implications: (1) Methodologically, the analysis of translations and retrospective interviews should also be taken into consideration except for the analysis of eye movement and keystroke data in order to form the triangulation (Alves 2003), which may give a more complete picture as for the effort involved in neologism translation when knowledge of context is present or not; (2) in translation practice and pedagogy, special attention could be paid to the different categories of Chinese neologisms. For example, teachers can consciously guide students to widely read materials with different categories of neologism expressions, collect the background knowledge of them, and notice the different influence of background knowledge on translators' choice-making.

However, there are some limitations to our study. One is that the translator sample size in our pilot study may be too small which makes it difficult to extrapolate these results to the general population of translators. That is why these results can only be regarded as preliminary, and the conclusion is tentative in nature. In the future, we would like to conduct a bigger study with more participants. Another limitation is the limited number and the genre of neologisms. Since this study is conducted under the eye-tracking and key-logging environment, the length of the passage should be controlled. In addition, the participants may experience fatigue in translating over three passages successively.

Acknowledgments I would like to express my gratitude to my supervisors, Professor Defeng Li and Professor Victoria Lei, from the University of Macau, for their guidance and assistance in my work. Thanks are also given to Professor Michael Carl from Kent State University and Mr. Yuxiang Wei from Dublin City University, for their comments on the chapter.

References

Alves F (ed) (2003) Triangulating translation: perspectives in process oriented research. John Benjamins, Amsterdam

Anderson L (1979) Simultaneous interpretation: contextual and translation aspects. Dissertation. Concordia University, Montreal, QC

Baker M (2006) Contextualization in translator-and interpreter-mediated events. J Pragmat 38:321–337
Carl M (2012a) The CRITT TPR-DB 1.0: a database for empirical human translation process research. In: Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012), pp 9–18
Carl M (2012b) Translog-II: a program for recording user activity data for empirical reading and writing research. In: Proceedings of the eighth international conference on language resources and evaluation (LREC12), pp 4108–4112
Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 13–54
Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process translation. Cogn Behav 2:211–232
Dijkstra T, Wahl A, Buytenhuijs F, Van Halem N, Al-Jibouri Z, De Korte M, Rekké S (2019) Modelling bilingual lexical processing: a research agenda and desiderabilia. Biling Lang Congn 22:703–713
De Groot AMB (1997) The cognitive study of translation and interpretation: three approaches. In: Danks JH, Shreve GM, Fountain SB, McBeath M (eds) Cognitive processes in translation and interpreting. Sage Publications, Thousand Oaks, CA, pp 25–56
Díaz-Galaz S, Padilla P, Bajo MT (2015) The role of advance preparation in simultaneous interpreting: a comparison of professional interpreters and interpreting students. Interpreting 17:1–25
Diriker E (2004) De-/re-contextualizing conference interpreting: interpreters in the ivory tower? John Benjamins Publishing, Amsterdam
Díaz-Galaz S (2011) The effect of previous preparation in simultaneous interpreting: preliminary results. Across Lang Cult 12:173–191
Germann U (2008) Yawat: yet another word alignment tool. In: Proceedings of the ACL-08: HLT demo session, pp 20–23
Gernsbacher MA (1996) The structure-building framework: what it is, what it might also be, and why. In: Britton BK, Graesser AC (eds) Models of understanding text. Routledge, London, pp 289–311
Gile D (1995) Basic concepts and models for translator and interpreter training. John Benjamins, Amsterdam
Gile D (2002) The interpreter's preparation for technical conferences: methodological questions in investigating the topic. Conf Interpreting Transl 4:7–27
Gile D (2005) Empirical research into the role of knowledge in interpreting: methodological aspects. Knowl Syst Transl 2005:149–171
Givón T (1989) Mind, code and context: essays in pragmatics. Lawrence Erlbaum, Hillsdale, NJ
Griffin JS (1995) The role of context, background knowledge, language skill, and translation training and experience in lexical choice in French-English translation. Dissertation. Kent State University, Kent, OH
Gutt E-A (2000) Relevance and translation: cognition and context. St. Jerome, Manchester
Jensen A (2005) Coping with metaphor: a cognitive approach to translating metaphor. HERMES 35:183–209
Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87:329
Kielar BZ (2013) Zarys translatoryki [the outline of translatology]. Wydawnictwo Naukowe IKL, Warsaw
Kim H (2006) The influence of background information in translation: quantity vs. quality or both? Meta 51:328–342
Kintsch W (1988) The role of knowledge in discourse comprehension: a construction-integration model. Psychol Rev 95:163
Lamberger-Felber H, Schneider J (2008) Linguistic interference in simultaneous interpreting with text. Efforts and models in interpreting and translation research. pp 215–236

Mason I (2006) On mutual accessibility of contextual assumptions in dialogue interpreting. J Pragmat 38:359–373
Newmark P (1988) A textbook of translation. Prentice hall, New York
Newmark P (1991) About translation. Multilingual matters, Bristol
Paradis M (1994) Toward a neurolinguistic theory of simultaneous translation: the framework. Int J Psychol 10:319–335
Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychol Bull 124:372
Schaeffer M, Carl M (2013) Shared representations and the translation process: a recursive model. Transl Interpreting Stud 8:169–190
Schaeffer M, Dragsted B, Hvelplund KT, Balling LW, Carl M (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 183–210
Setton R (2006) Context in simultaneous interpretation. J Pragmat 38:374–389
Shreve GM, Schäffner C, Danks JH, Griffin J (1993) Is there a special kind of "reading" for translation? An empirical investigation of reading in the translation process. Target 5:21–41
Sperber D, Wilson D (1986) Relevance: communication and cognition. Blackwell, Oxford
Van Dijk TA (2001a) Text and context revisited. In: proceedings of the first Seoul international conference on discourse and cognitive linguistics: perspectives for the 21st century, Seoul, Korea
Van Dijk TA (2001b) Discourse, ideology and context. Folia Linguistica XXXV:11–40
Weinreich U (1953) Languages in contact. Mouton, The Hague
Wei Y (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Xiang X, Zheng B (2011) Understanding and reformulating metaphors: an empirical study on English-Chinese sight translation. Foreign Lang Teach Res 43:422–436
Xiang X, Zheng B (2015) Background information and interpreting quality of metaphorical expressions: looking into the products of English-Chinese sight interpretation. Foreign Lang their Teach 13:791
Zheng B, Xiang X (2013) Processing metaphorical expressions in sight translation: an empirical–experimental research. Babel 59:160–183
Zheng B, Xiang X (2014) The impact of cultural background knowledge in the processing of metaphorical expressions: an empirical study of English-Chinese sight translation. Transl Interpreting Stud 9:5–24

Part IV
Translation Process Research and Post-cognitivism

Computation and Representation in Cognitive Translation Studies

Michael Carl

Abstract A separation is recently being construed within cognitive translation studies (CTS) between *translation process research* (*TPR*) and *cognitive translatology* (CT), in which TPR is said to implement a *computational* approach, while CT endorses *4EA* approaches. For Muñoz (2017) TPR and CT represent "mutually exclusive views on human cognition," and some research seems to adopt this view (Xiao and Muñoz 2020; Halverson 2019, among others). Muñoz' statement stipulates that either TPR is incompatible with basic assumptions of *4EA* or CT is not computational. We investigate various notions of "computation," some of which rely on a specific meaning of *representation*. We contend that TPR has mainly developed methodologies for empirical research and has *not*, in general, made any particular theoretical or representational commitment. CT, in contrast, supports ontological (i.e., 4EA) views on cognition while being agnostic with respect to methodological issues. We conclude that TPR is compatible with at least some flavors of the CT ontological perspective and that a mechanistic view on computation may offer a more comprehensive prospective on CTS.

1 Introduction

Cognitive translatology (CT) starts out with the assumption that the body and the environment have a constitutive role in cognition. According to Muñoz (2017, 563 ff.), CT assumes that cognition is (1) *embodied*, as it uses the full body (e.g., spatial metaphors); (2) *embedded*, because the brain is nested into both a body and a physical and sociocultural environment; (3) *enacted*, because the environment is selectively created; (4) *extended*, since the brain/mind actively offloads tasks and procedures on the "outside" world; and (5) *affective*, as emotions drive and fine-tune our mental processes and our behavior. As the multitude of these labels

M. Carl (✉)
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_13

indicates, there is a broad spectrum of different theories and models, premises, and assumptions involved which all reject the computational theory of mind (CTM) in some way and may be subsumed under the label *post-cognitivism*.

TPR is concerned primarily with the research question "by what observable and presumed mental processes do translators arrive at their translations?" (Jakobsen 2017, 21). Jakobsen traces the roots of TPR back to Krings' pioneering studies on post-editing, among others. In its quest to explain the findings, Jakobsen posits that "TPR has conceptualized the human mind and brain in information processing and computational metaphors, either implicitly or explicitly." However, for him, TPR "was not motivated by any strong behaviourist convictions" (2017: 28) or for that matter any other theoretical framework or representational commitment. While some TPR texts may be perfectly aligned with the classical cognitivist view, TPR is mainly driven by the possibilities that new technologies provide for the investigation of the translating mind. Consequently, Jakobsen suggests including electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), as well as "physiological reactions, such as respiratory speed, pulse rate, blood pressure, skin conductance, and muscle reactions" and concludes that the "extended perspective is therefore a highly relevant complementation of TPR's traditional focus."

Muñoz' (2017) incompatibility statement stipulates that (1) either 4EA approaches to cognition are not computational or that (2) TPR research is not compatible with the assumption that the body and the environment have a constitutive role in cognition. That is, it is incompatible with a post-cognitivist view on cognition.

Given the obvious focus of TPR on translation aids (e.g., MT post-editing) and the plethora of studies that investigate and in fact show the role of the technology (i.e., the translation environment) on the human translation process (e.g., reduction of effort, etc.), assumption (2) can be ruled out. TPR has—from its beginning—investigated the usage of external resources and used behavioral measures (e.g., keystrokes and gaze) to underpin its hypothesis and conclusions. A review of the chapters in this volume may testify this statement.

We thus start this chapter by addressing claim (1) and investigate what it might mean for a theory (or a cognitive translation approach) to be non-computational. We give several definitions of the term and discuss some assumptions of the "classical cognitivist view," i.e., the CTM. We discuss Steiner's (2014) dichotomy of ontological and methodological approaches and locate CT within an ontological but methodological agnostic framework, while TPR has used a repository of representational and non-representational methodologies in an ontologically agnostic manner. We, thus, find that TPR can be interpreted in a post-cognitivist framework, and we suggest that a mechanistic view of computation may allow us to better distinguish between different approaches within CTS.

The second section introduces four definitions of non-computability and discusses possible views why CT might be considered non-computational. The third section discusses aspects of the CTM which assume a controversial concept of semantic representation, said to be required for computation to take place. Under the

premise that there is no computation without representations, the discussion shifts from the definition of *computation* to that of *representation*. Section 4 provides a taxonomy to classify representationalism and anti-representationalism, and Sect. 5 discusses alternative (e.g., *mechanic*) views on computation that draw on research in biology and neurosciences. We conclude that TPR and CT do not necessarily propose incompatible views on the translation process; they may be distinguished by a dichotomy between ontological and methodological approaches.

2 Non-computational

Is TPR perhaps supposed to be computational because it addresses (a subset of) questions that are computable and CT addresses such questions that are non-computable? What is the borderline, and what does it mean to be computable or non-computable? As we will discuss, there are many definitions of this term, and we start with the most obvious according to which "computability is the ability to solve a problem in an effective manner."[1] This definition contains two terms that need clarification:

- *Problem*: We take it that most people agree that translation can be considered a "problem" at least under some of its very many possible definitions such as *a gap between the existing state and a desired state*, *an intricate unsettled question*, etc. (see also Nitzke 2019: 51ff).
- *Effective:* We also take it that everyone agrees that translations can be produced (e.g., by humans) successfully and impactful.

However, effective here has a narrower meaning. A problem is computable if the solution to it can be described by a procedure which eventually terminates with a (correct) answer. How, then, might one want to reason about the translation process in a non-computational manner, as CT proponents seem to suggest? Showing that a problem is non-computable requires to show that no procedure (or algorithm) exists to solve the problem (in a reasonable amount of time). Here are several reasons why something may *not* be computable[2]:

1. Effectively impossible: for instance, if the computation is NP complete and the best possible algorithm cannot terminate simply because of the time it takes.
2. Logically impossible: an algorithm of any sort cannot be found due to an internal paradox. For instance, it is known that there cannot exist a general algorithm to solve the *halting problem* for all possible program-input pairs, i.e., decide whether the programs eventually terminate.

[1] Wikipedia https://en.wikipedia.org/wiki/Computability. Accessed 26 Apr 2020.

[2] There are more reasons, such as non-computability of uncountably infinite numbers. However, for space and relevance, these reasons are not considered here.

3. Formally impossible: the problem is not sufficiently formalized. There is a narrative and a description of a problem, but the formalization is not advanced enough, or the narrative is too general to allow for an effective stepwise path to a solution.
4. Compression impossible: if the entire universe is the computational device for some phenomenon, say the trajectory of a particle, there might not be a simpler program that can compute the answer faster than the universe itself.

Non-(or anti) computational CT defenders might take one of those positions but would then need to explain how humans do that very task in a psychologically realistic manner. In this section, we go through these options.

Effectively impossible: Even though it is known that translation is NP-complete (Knight 1999), the MT community has developed strategies and heuristics to overcome these intrinsic complexities, one of them being that texts are typically translated segment by segment, instead of an entire text at once. However, within the translation process literature, there is—to the best of the authors' knowledge—no discussion concerning its effective non-computability. To the contrary, as discussed, e.g., by Carl (Chap. 9) and Lacruz et al. (Chap. 11), also humans seem to produce translations in chunks.

Logically impossible: a common way to show that a problem is non-computable (or falls within a certain complexity class) is through a contradiction based on criterion 2: if it can be shown that the easiest solution to solve a problem can also be used to solve the *halting problem*, we know that the solution in question is logically impossible, i.e., non-computable (by a Turing machine). However, this proof of non-computability by contradiction has not been deployed within CT. It is also a matter of debate whether humans perform hyper-computations (i.e., beyond Turing-computable). According to Piccinini (2009), no one has shown up to date how to solve genuinely Turing non-computable problems, such as the *halting problem.*

Formally impossible: Criterion 3 relates to a methodological choice or research paradigm, i.e., whether or not a researcher wants to develop the area of investigation so that crucial aspects of it can be specified into a stepwise approach for a solution. If the formalization or solution is not the goal, a non-computational approach would, perhaps, "aim to enliven rather than report, to render rather than represent, to resonate rather than validate, to rupture and reimagine rather than to faithfully describe, to generate possibilities of encounter rather than construct representative ideal types" (Vanni 2015, 15). It would "find inspiration in the arts, in the poetics of embodied living, in enacting the very unactualized expressive and impressive potentials of social-scientific knowledge, in taking dedicated risks, in exercising passion, and in finding ways to reconfigure thinking, sensing, and presenting by emphasizing the singular powers of action, locution, and thought" (Vanni 2015). Translation studies, including process and ecological translation studies, have plenty of reports that fall into this category. However, to the author's understanding, CT does not fall into this category, as it precisely aims at reporting, describing, representing, validating, etc.

Compression impossible: Non-computability criterion 4 is of an ontological/philosophical nature. It states that a problem cannot be decomposed at all. To alleviate this objection, it could be argued that a big entangled problem (e.g., a translation) could be decomposed and approximated through a set of simpler tractable routines and intermediate solution. However, an approximation may lead to rapid degrading precision, and thus misguiding predictions and conclusions. Muñoz (2017) endorses this non-computability criterion, when, for him on the one hand, CT approaches assume that "many elementary cognitive functions are instrumental in so-called higher cognitive functions" (Muñoz 2017, 564), but, on the other hand, "the divide between lower and higher functions is not that important" (Muñoz 2017), and a modularization is impossible. Similarly, Risku (2014, 339) endorses non-computability criterion 4 when characterizing cognition in translation as consisting of interconnected and self-organizing processes which "includes all operations that work on internal and external representations with the aim of creating translations." She enumerates a large number of cognitive, environmental, and social factors that need to be studied in "authentic, personal, historically embedded environments" (Risku, 2014, 335). As Carl and Schaeffer (2017) point out, if the translation "processes are so heavily dependent on the context in which they develop, it is hard to see how findings in one study can carry over into a second study in a different setting" (Carl and Schaeffer 2017, 59). In a setting with n factors each with m levels, an experiment would have to control for m^n interactions which may very quickly become intractable and thus non-computable. However, not all post-cognitivists take this position.

3 Non-representational

On a different background, also Martín de León (2017, 109–110) claims that within CT "cognition has to be explained in a noncomputational way." She traces computational approaches in translation back to the 1960s and 1970s (i.e., to the Leipzig School and the Paris School (cf. also Muñoz (2017, 562)) and more recently to Gutt's (2000) relevance theory where "[t]ranslation was understood as a rule-guided transformation of symbols from one code into symbols of another code" (Martín de León 2017, 109). For her, as well as for Muñoz, a computational approach entails to defend "the *computational theory of mind*, [by which] cognitive processes can be described and explained as manipulation of formal symbols in a language of thought" (ibid, our italics). This assumption establishes a troublesome criterion 5:

5. Non-representational: computation is a rule-governed manipulation of symbolic representations. The mind contains faithful representations of (aspects of) the environment which are manipulated and formalized in the form of a language of thought (LOT). Theories and models that do not follow this framework are consequently "non-representational," or synonymously "non-computational."

The difficulty with this notion of "non-computational" arises from the fact that it is based on the negation of a controversial concept of representation on which Muñoz's rejection of TPR seems to be based. Muñoz (2017, 561–562) stipulates that "notions that are characteristic of computational translatologies [i.e. TPR] but [which are] rejected by cognitive translatology" allegedly include, among others, claims such as "[t]hought is (mostly) conscious, rational, and logical"; that "[l]inguistic symbols carry stable, self-contained meaning"; that natural languages "can be thought of as entities that are independent from their users"; that "objective meaning [is] subjected to neutral processing [in the human mind]"; etc. These notions of "computation" fit (perhaps) the original idea of a LOT and may be underlying the Leipzig School or the Paris School approaches in the 1970s and 1980s. However, to the best of the authors' knowledge, there is no recent approach within TPR which explicitly supports such notions of a CTM. TPR has mainly been interested in modelling and understanding non-representational phenomena, such as translation effort and translation duration, as many of the chapters in this volume demonstrate.

Despite this—as Muñoz points out—at least "some versions of embodiment allow for internal representations," and hence the question whether we assume or not the existence of mental representation "turns out not to be the touchstone to distinguish subfamilies of cognitive translatologies" (Muñoz 2017, 564). However, Muñoz does not clarify what "representation" actually means for him, how mental representations can emerge without computations, and how representations are in the human mind processed in a non-computational way? And vice-versa, can there be computations on non-representational states? Are all computations rule-based manipulations of representations? Is computation intrinsically representational?

Since the formulation of the physical symbol system hypothesis[3] (PSSH) in the 1970s (Newell and Simon 1972), computation, cognition, and translation theory have undergone a large number of dramatic changes and additions, as also pointed out by Martín de León (2017). Computation—as we know it today—is based on all kinds of representations (e.g., symbolic, propositional, probabilistic, associative, structured or unstructured, connectionist distributed or localist, multimodal, embodied, etc.) which do not (necessarily) correspond to the notion of symbols in the LOT sense. It may thus become more important what exactly these assumed representations are and how they are supposed to be activated and processed. A possible separation may then be made, for instance, between non-mentalistic associations which emerge where an environmental stimulus is directly associated with a behavioral response and mental associations which lead from one associated idea to another associated idea or to a reaction (Pereplyotchik 2016). Associations may be simple or complex (such as connectionist networks) and modelled as mechanical behavioral transducers. Representations can be structured or unstructured, modelled as cognitive maps, and specify correctness conditions,

[3]The PSSH states that a system has the necessary and sufficient means for general intelligent action.

truth values, etc. Steiner (2014) aims at a very general definition of representation that covers many variations, including "parallel and sub-symbolic distributed processing vs. the symbolic conception; action-oriented conceptions of cognition and representation vs. the idea of cognition as a mirror of the environment; mental representations as maps, models or pictures vs. mental representations as propositional sets of symbols . . . " (Steiner 2014:52). Also Martín de León suggests "to drop the representational/nonrepresentational dichotomy and to consider a continuum of degrees and types of representationality" (2017, 119). In support of this suggestion, and a weakening of the non-computational criterion 5, Steiner (2014) argues that many terms (such as Planet, Atom, etc.) have changed their descriptions and meanings over the centuries: "Why should it not also be the case for the concept 'mental representation' as used for denoting a natural, intracranial and subpersonal phenomenon?" (Steiner 2014, 72).

While there are thus a large number of possible notions of "representation," Chemero (2000, 2010) illustrates the limits of the term with the example of the Watt governor, which is based on the distinction between *internal states* and *representations*. The functioning of the Watt governor, he argues, can be understood as a (non-representational) dynamic system and described in terms of *internal states* and precise mathematical terms, without resorting to a language of representations at all. A definition of representation, such as "a set of states . . . produced by one part of the system, for the use by some other part of the system, in adapting the system to some aspect of the environment" would make it possible for *every* system to manipulate representations, including rocks, galaxies, and the Watt governor. He, thus, rejects the *pancomputationalism* (see Sect. 5) which suggests that everything is computation. However, the actual understanding of a dynamical system, he claims, comes from the understanding of the dynamic fluctuations, and it is of little use if a "representational gloss does not predict anything about the system's behavior that could not be predicted by a dynamical explanation alone . . . If one has the complete dynamical story, what is left to be explained?" (Chemero 2010, 77).

Whether or not a system's functioning is grounded in representations—and hence is a computational system according to criterion 5—thus becomes relative to the understanding of the observer and to what extent the system can be described in terms of dynamical system theory. It depends on whether or not a "representational gloss" adds explanatory value to its dynamics. This implies that a representational account becomes increasingly obsolete to the extent we understand, describe, and communicate the intrinsic dynamics of the system. A representational gloss might then just be a means to approximate or communicate the functioning of the dynamic system, but the representationalism (e.g., LOT) and non-representationalism (e.g., a dynamic system) are not necessarily exclusive or incompatible views.

4 Typologies of Representation

A separation can be made between methodological and ontological assumptions about the nature and the status of representations. Ontologies relate to philosophical questions with respect to the reality and the existence of entities, while methodological considerations relate to techniques and tools and the question "Which notion of representation provides better explanatory advantages for scientific findings and modelling?"

Steiner (2014) distinguishes ontological representationalism and ontological anti-representationalism from methodological representationalism and methodological non-representationalism. For Steiner, representations are "contentful (information-carrying) physical structures that have a real ontological and explanatory status" (Steiner 2014, 52), which, however, "does not depend on the existence of linguistic, representational or symbolic human practices" (Steiner 2014: 47).

As shown in Table 1, he suggests all eight combinations of the ontological and methodological perspective. Four of them (in bold) are either agnostic to the ontological or methodological perspective, as they do not imply any methodological or ontological implications or commitments, respectively. **MR** (methodological representationalism) and **MN** (methodological non-representationalism) are agnostic on the issue of the ontological reality of mental representations. **OR** (ontological representationalism) and **OA** (ontological anti-representationalism) are agnostic on the issue of the reality of methodological representations. "Ontological representationalism considers that mental representations exist, whereas ontological anti-representationalism denies their existence" (Steiner 2014, 50), but these positions are rare, as "Ontological representationalism and anti-representationalism almost always ... include methodological commitments" (2016, 52). Four positions are combinations of an ontological and a methodological view, labeled RR, RN, AN, and AR in Table 1.

- **RR**: This representationalism holds that cognitive systems involve the "use, the retrieval or the manufacture of mental representations" and that models of cognitive subsystems "appeal to the concept of 'mental representation'—for descriptive, predictive and explanatory purposes" (p. 52).
- **AN**: The anti-representationalism position claims that cognitive systems and operations as well as scientific models and operations do not include the use, retrieval, or manufacture of mental representations.
- **RN**: This position accepts that mental representations exist but consider that scientific models should better do without the concept of "mental representation."
- **AR**: This position assumes that mental representations do not exist, but still holds that they are our best ways to capture and explain the complexity of cognitive behavior.

Steiner further separates these eight combinations of ontological/methodological positions into three levels, a local, basic, and global level, depending on whether

Table 1 Combination of ontological and methodological positions with respect to representation

	Ontological perspective		
Methodological perspective	Representationalism	Anti-representationalism	Agnostic
Representationalism	RR	AR	**MR**
Non-representationalism	RN	AN	**MN**
Agnostic	**OR**	**OA**	–

the ontological/methodological positions apply only to the cognitive system under consideration (local) and whether it holds for most cognitive subsystems (basic) or to all of them (global). Steiner explains:

> The position is global when it applies to every cognitive system and every cognitive part of it (operations and subsystems such as faculties). It is basic when it applies to most cognitive systems, operations and subsystems. And it is local when it only applies to the particular cognitive system, operations or subsystem under consideration. (Steiner 2014, 51)

Different positions of representationalism can thus be maintained for different cognitive subsystems. Steiner arrives at 24 possible positions that cognitive theories and their subsystem can take. For instance, a basic anti-representationalist position with respect to, e.g., face or voice recognition can still be compatible with a local representationalist position with respect to language production or translation, etc. Only a global position X will not be compatible with any other non-X position. It is conceivable that we will eventually arrive at a huge number of cognitive models that are internally and mutually incompatible, but on the long term, it may be difficult to explain why we need this plethora of different possible conceptions and definitions of "representation."

CT is methodologically agnostic as it embraces a large number of ethnographic, empirical, quantitative, and qualitative paradigms which may or may not rely on a notion of representation (Muñoz 2017). It takes a position of *ontological representationalism* (**OR**), to the extent it assumes there exists a mediation between the translator and the environment through mental representation. For Martín de León (2017, 121), for instance, "embodied, embedded approaches view mental representations as dynamic internal support to meaning construction and translation and interpreting processes." Similarly, Martín and Rojo López (2018) believe that schemas and frames structure the knowledge in our minds, and meaning emerges as an inferential process: "meaning is an online process resulting from the interaction between schematic, ad-hoc knowledge structures and further cognitive or construal operations" (2018: 68). CT in this view is clearly on the ontological-representational side which shares essential properties with the CTM, but which—in contrast to CTM—rejects computationalism. However, it is not clear how representations can emerge and support meaning construction and how those interactions between knowledge structures can be modelled or understood in a non-computational manner. It is also unclear what those "cognitive or construal operations" are other than computational devices that operate on those structures and representations.

In contrast to this, as several chapters in this volume demonstrate, TPR is mostly agnostic toward the ontological reality of mental representations. It may adopt an *ontological anti-representationalism* when assuming a direct contact between the translator and its environment (cf. Raab and Araújo 2019). Carl (Chap. 15), for instance, develops an *ontological anti-representationalist* view of the monitor model (Carl and Dragsted 2012, Schaeffer and Carl 2013) which presumes no need for mental representations for translation. However, TPR more often follows an **MR** approach when using tools such as lexicons and dictionaries, chunks or parsers, semantic and syntactic annotations, etc., as a basis to explain translational behavior, but it has also endorsed non-representational (**MN**) approaches to the extent that the analysis is based on (word) frequencies or probabilities, (production) rhythm, densities, translation entropy, distributed representations, etc.[4]

As can be seen in this discussion, on the one hand, the controversy shifts from the simple dichotomy computational vs. non-computational to representationalism vs. non-(or anti) representationalism and what the nature of those representations might be. The differences in these positions might not be very clear without a precise specification of these terms. On the other hand, there are different accounts of computationalism, as discussed in the next section.

5 Typologies of Computation

Piccinini (2009) suggests a distinction between three types of computation—*causal computation*, *semantic computation*, and *mechanic computation*—depending on the definition of the computational states that are processed. Causal computation takes the view of *pancomputationalisms*, where every possible state in any physical system is a possible computational state, while *semantic computation* restricts that notion to semantic representations, in line with a LOT, and its contradictions. According to the mechanical account, computation is a vehicle-independent *functional explanation* that requires components to be appropriately organized, and the combined activities of a system constitute the capacities of the computational mechanism. With these constructions, he shows that nervous systems (but not all physical systems) perform computations in a generic sense and that computing systems can be coupled with a body, an environment, or both. In this section, we discuss some views on mechanic computation.

Ramstead et al. (2016, 4) develop a mechanistic account of affordances within a dynamic systems approach, in which "computations (digital, analog, neural) can occur without any form of semantic content." Some post-cognitivists describe the

[4]Pereplyotchik (2016, 171) points out that in "classical" cognitive architectures, such as ACT-R, behavior emerges from production rules that fire in response to dynamically changing contents of buffers. While the buffer states are representations, the procedural knowledge that is encoded in the production rules is not. In this view, probabilities, frequencies, rhythm of action, etc. emerge from firing rules but are non-representational in nature.

mind as a dynamic system that can do without a semantic notion of representation (Varela et al. 1991, Chemero 2000, 2010; Hutto and Myin 2013). Carl et al. (2019) develop a dynamic systems-based view on the translation process, in which the human translation process consists of a hierarchy of interacting word and phrase translations systems which organize and integrate as dissipative structures. Carl (Chap. 15) develops a probabilistic notion of translation affordances that explains priming in translation as a maximization of translation abilities which eliminates the need for contentful representation.

Also Damasio (2017) also seems to endorse a mechanistic view of computation. He traces computation back to *homeostasis*, "a seemingly indomitable 'intention' to maintain itself alive and sail forth," which lies the very basis of life and which is based in "chemical pathways" of metabolism (2017; 33ff). Even though bacteria do not have a nerve system, "they have varieties of perception, memory, communication and social governance ... [which] rely on the chemical and electrical networks of the sort nervous systems eventually came to possess." According to Damasio, bacteria perform "computations ... [which] permit them to assess their situation and, accordingly, afford to live independently or gather together if need be" (2017, 54). But if bacteria "compute," then certainly also nerves do, as well as the human brain and body, and neural networks in interaction with their environments. Vida (2020) reports that dendrites can perform XOR computations of two separate inputs, an operation that was previously considered impossible for single neurons. While one processing element in an artificial connectionist network may represent an entire brain area rather than an actual neuron, a single neuron in a biological system is a device, which "may be able to compute truly complex functions. For example, it might, by itself, be able to recognize an object" (2020: np).

Lawrence (2017) seems to promote a mechanistic view in which he locates the difference between machine and human intelligence as well as consciousness in an *embodiment factor* which he defines as "the ratio between an entity's ability to communicate information vs compute information" (2017: 3):

$$\text{embodiment factor} = \frac{\text{compute power}}{\text{communication bandwidth}}$$

Current home computers have a low embodiment factor (~10), as they compute approximately ten times faster than they can communicate (i.e., in the order of gigabits). If an entity can share almost all its "thoughts and imaginings, then that entity is arguably no longer distinct from those which it is sharing with. It could be thought of as merely a sensor" (Lawrence 2017: 3). Humans, in contrast, have very limited communication channels (~60 bits/s), but their brains compute the equivalent of an estimated exaflop (10^{18} floating point operations per second), which yields an embodiment factor of roughly 10^{16}. This extremely high embodiment, Lawrence says, "is arguably the driver of much that is beautiful in our society" and speculates that a concept of "self" and consciousness may have emerged as the consequence of this locked-in nature of intelligence, "because of our inability to directly communicate our mental state" (Lawrence 2017: 7).

According to Villalobos and Dewhurst (2017, 5) "[m]echanistic accounts of computation present a way of understanding computation that remains neutral with respect to representation, but typically do not engage in metaphysical considerations about the relationship between computing systems and the world." There are thus many approaches to define computation and representation, with different implications. Note the potential confusion: Under a mechanical conception, the dynamic systems view of cognition falls into the realm of computation, while in a semantic view of computation, dynamic systems are non-(or anti) computational! Unfortunately, Martín de León and Muñoz (and others) do not point to ways how the translation process ought to be representational but "explained in a noncomputational way."

6 Conclusion

This chapter investigates different definitions of "non-computational" and assesses under which definitions this term might be discriminating between CT and TPR. We find two possible explanations as to why models or theories within CTS might be non-computational: some researchers claim that factors and relations that play in translation are too entangled so that the process of translation cannot be modularized, and it cannot be effectively described and formalized and is, thus, non-computational.

The other reason why an approach within CTS might be non-computational is based on a rejection of the CTM, in particular the claim that the mind performs computations on objects that specify correctness condition and that are faithful representations of an outside world. According to CTM, objects in the environment cause our senses to generate mental representations of the external stimulus (e.g., in the form of a propositions), and the mind infers meaning representations which are located inside the head. Post-cognitivists and 4EA approaches of cognition (and CT) have refused this view of cognition and proposed a large number of alternative theories. While none of the alternatives has become mainstream to date in CTS, CT seems to endorse a concept of representation as a mental reproduction of the environment and the mind as an inference machine over these representations.

At least some CT defenders seem to believe that TPR endorses the classical CTM—which conflates a very specific notion of semantics-based representation with that of rule-based manipulation—and conclude that TPR is therefore incompatible with CT. We show that this is not the case.

1. TPR has mostly been concerned with methodology and technologically heavy empirical research and gathering and analyzing data that can be interpreted in many different theoretical frameworks. As this volume shows, the preferred theoretical framework for TPR has borrowed from connectionist models of bilingualism, but also from dynamic system theory, as well as psychology, linguistics, cognitive sciences, and cognitive or post-cognitive theories.

2. TPR has mainly investigated "non-representational" aspects of cognition, such as translation duration, translation rhythm, and translation effort, and so has no particular affiliation with a CTM, which focusses on representation.
3. TPR has often investigated how human translators cope with their increasingly technological environment. A set of similar questions are raised within CT in terms of embodied, embedded, and enacted cognition.
4. CT endorses an ontological framework which assumes that the body and the environment have a constitutive role in cognition and that cognition consists in inferential steps involving rules (construal operations) and representations.
5. CT is agnostic with respect to methodological issues as it uses ethnographic and meta-cognitive analyses but also behavioral methods, including keyloggers and eyetracker, and is thus compatible with TPR methodology.

Given that TPR findings can be interpreted within a post-cognitivist 4EA framework and that CT makes use of methodologies that are also used in TPR, we conclude that at least certain shades of CT are well compatible with TPR.

Alternative definitions of "computation" have been proposed that do not assume semantics-based representations and rule-based manipulations. A mechanical account of computation can explain the activities and functions of a (cognitive) system as a computational mechanism without the strong assumptions of representation as the CTM does. For Villalobos and Dewhurst (2017, 5), "representation and computation are distinct concepts that should be understood independently of one another." In a mechanistic view of computation, it is possible to reject the representational claim of a CTM without rejecting computationalism all together. We may then be able to talk about computational systems which do not assume (contentful) representation—such as dynamic systems—and we can acknowledge representations in a non-computational context, such as a narration or the report of an experience. Under a mechanical view of computation, post-cognitive translation studies "need not be committed to an anti-computationalist attitude" (Villalobos and Dewhurst 2017, 9).

TPR has produced translation process data and explored different frameworks to explain empirical observations. It has explained data through generalizations and by making use of various theoretical frameworks. We are used to believe that data without a model is just noise; however, Anderson (2017: np) points out that the:

> availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

Alternatively, Hutchinson (2019) suggests that if we:

> forgo the theory and look, really look, at humans as members of social orders interacting with each other and the world, in social settings, and describe what we see in the language available to those members then we will resist the pitfalls of abstraction.

Both are viable paths in post-CTS, but they are not mutually exclusive.

References

Anderson C (2017) The end of theory: the data deluge makes the scientific method obsolete. https://www.wired.com/2008/06/pb-theory/. Accessed 22 Aug 2020

Carl M, Dragsted B (2012) Inside the monitor model: processes of default and challenged translation production. TC3 2(1):127–145

Carl M, Schaeffer MJ (2017) Models of the translation process. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Wiley-Blackwell, Hoboken, NJ, pp 50–70. https://doi.org/10.1002/9781119241485.ch3

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process in: translation. Cogn Behav 2(2):211–232

Chemero A (2000) Anti-representationalism and the dynamical stance. Philos Sci 67(4). https://doi.org/10.1086/392858

Chemero A (2010) Radical embodied cognitive science. MIT Press, Cambridge, MA

Damasio A (2017) The strange order of things: life, feeling, and the making of cultures, Kindle edn. Random House LLC, New York, NY

Gutt E-A (2000) Translation and relevance: cognition and context, 2nd edn. St. Jerome Publishing, Manchester

Halverson SL (2019) 'Default' translation: a construct for cognitive translation and interpreting studies'. Transl Cogn Behav 2:187–210

Hutchinson P (2019) The missing 'E': radical embodied cognitive science, ecological psychology and the place of ethics in our responsiveness to the Lifeworld. In: Backström J, Nykänen H, Toivakainen N, Wallgren T (eds) Moral foundations of philosophy of mind. Springer, Berlin, pp 103–127

Hutto, Myin (2013) Radicalizing enactivism. The MIT Press, Cambridge, MA

Jakobsen AL (2017) Translation process research 21. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Blackwell, Oxford, pp 21–50

Knight K (1999) Decoding complexity in word-replacement translation models. Comput Linguist 25(4):607–615

Lawrence ND (2017) Living together: mind and machine intelligence, In corr. http://arxiv.org/abs/1705.07996. Accessed 22 Aug 2020

Martín de León C (2017) Mental representations. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Wiley-Blackwell, Hoboken, NJ, pp 106–126

Martín M, Rojo López AM (2018) Meaning. The Routledge handbook of translation and culture. Routledge, London, pp 61–78

Muñoz RM (2017) Looking toward the future of cognitive translation studies. In: Schwieter JW, Ferreira A (eds) The handbook of translation and cognition. Blackwell, Oxford, pp 555–573

Newell A, Simon HA (1972) Human problem solving. Prentice-Hall, Upper Saddle River, NJ

Nitzke J (2019) Problem solving activities in post-editing and translation from scratch: a multi-method study. Language Science Press, Berlin. https://langsci-press.org/catalog/book/196. Accessed 16 Feb 2020

Pereplyotchik D (2016) Psychosyntax: the nature of grammar and its place in the mind. Springer, Berlin

Piccinini G (2009) Computationalism in the philosophy of mind. Philos Compass 4(3):515–532. Blackwell Publishing Ltd Part 3

Raab M, Araújo D (2019) Embodied cognition with and without mental representations: the case of embodied choices in sports. Front Psychol. https://doi.org/10.3389/fpsyg.2019.01825

Ramstead MJ, Veissière SP, Kirmayer LJ (2016) Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. Front Psychol 7:1090. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960915/

Risku H (2014) Translation process research as interaction research: from mental to socio-cognitive processes MonTI. Monografías de Traducción e Interpretación:331–353

Schaeffer M, Carl M (2013) Shared representations and the translation process: a recursive model. Transl Interpreting Stud 8(2):169–190. reprint in Describing Cognitive Processes in Translation: Acts and events, Edited by Maureen Ehrensberger-Dow, Birgitta Englund Dimitrova, Séverine Hubscher-Davidson and Ulf Norberg. [Benjamins Current Topics, 77]

Steiner P (2014) Enacting anti-representationalism. The scope and the limits of enactive critiques of representationalism, Avant. Trends Interdiscip Stud 5(2):43–86

Vanni P (2015) Non-representational research methodologies: an introduction. In: Vanni P (ed) Non-representational methodologies: re-envisioning research. Routledge Advances in Research Methods, London

Varela F, Thompson E, Rosch E (1991) The embodied mind: cognitive science and human experience. MIT Press, Cambridge, MA

Vida I (2020) Hidden computational power found in the arms of neurons. QuantaMagazin. Neurosciences. https://www.quantamagazine.org/neural-dendrites-reveal-their-computational-power-20200114/

Villalobos M, Dewhurst J (2017) Why post-cognitivism does not (necessarily) entail anticomputationalism. Adapt Behav 2017:1–12. https://doi.org/10.1177/1059712317710496journals.sagepub.com/home/adb

Xiao K, Muñoz R (2020) Cognitive translation studies. Theoretical models and methodological criticism. Call for papers for a Linguistica Antverpiensia 19. https://lans-tts.uantwerpen.be/index.php/LANS-TTS/announcement/view/12. Accessed 22 Aug 2020

Translation Norms, Translation Behavior, and Continuous Vector Space Models

Michael Carl

Abstract Several models of the bilingual mind have been suggested (e.g., DFM, Multilink), which are inspired by connectionist and artificial neural network models. Those models aim at explaining and predicting translation latencies of human translators based on cross-linguistic similarities of word properties. To foster researching these models, bilingualism studies have developed *translation norms* which enumerate factors (such as concreteness or ambiguity) assumed to be responsible for delays in word recognition and word translation production. While neural machine translation (NMT) systems are currently revolutionizing the translation industry, the compatibility of those identified processes and assumed representations in bilingualism studies have not (often) been compared with measures and representations that can be traced in NMT systems. In this chapter, we map translation norms into word embeddings (i.e., vector representations of words used in NMT systems) and compare model predictions—as gathered from similarity measures of continuous word embeddings—with reported human behavioral latencies. We also investigate to what extent the findings from single-word translation experiments can be carried over to translations in context. We re-assess predictions of DFM and Multilink in the light of the findings.

Keywords Word embedding · Vector space model · Cross-lingual projection · Bilingualism

1 Introduction

Artificial neural networks (NNs) are increasingly replacing traditional natural language processing (NLP) systems including machine translation (MT). In contrast to the traditional NLP system, NNs are usually trained in an end-to-end manner

M. Carl (✉)
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_14

without explicitly encoding linguistic features. The crucial input–output mapping relations between the source language and target language are learned and processed in internal hidden layers, and the weights of connections between the network's nodes detect, connect, and integrate relevant information automatically. While the end-to-end training of NNs achieves in many cases much higher precision and fluency than conventional NLP systems, it is often difficult to understand, visualize, and interpret the learned internal connections and reconstruct what they might represent. However, to assess accountability and reliability of the predictions, and thus increase trust in the systems' performance, researchers have attempted to analyze activation patterns of the NN's internal representations.

Belinkov and Glass (2019) give an overview of recent studies that investigate what kind of linguistic information is captured in NNs for NLP. They report "that neural networks are able to learn a substantial amount of information on various linguistic phenomena" (Belinkov and Glass 2019: 51) including word position, word order, morphological classes, number agreement, syntactic and semantic information, and structure-sensitive phenomena. They report that within multilayer NN architectures, "local features are somehow preserved in the lower layer whereas more global, abstract information tends to be stored in the upper layer" (Belinkov and Glass 2019). In support of such hypothesis, and in analogy with human language processing, Buchweitz and Prat (2013) also report that there is a large overlap in brain activation between two languages in proficient bilinguals: "neuroimaging and behavioral research alike show that there is a shared semantic representation in bilinguals, that is, shared concepts and shared cortical tissue" (Buchweitz and Prat 2013, 430).

Dhar and Bisazza (2020) investigate how syntactic knowledge is represented in a multilingual LSTM-based[1] NMT and what factors trigger the cross-lingual transfer of syntactic knowledge. Their work is inspired by psycholinguistic models (e.g., Hartsuiker et al. 2004) of the bilingual mind and by insights from second language acquisition, where priming studies suggest that lexical and syntactic representations are shared in the mind of bilingual individuals. Trained on huge amounts of bilingual data, a multilingual NN makes it possible to investigate properties of cross-lingual sharing of linguistic features in a simulated context. Dhar and Bisazza (2020) find that "POS [part of speech] categories are shared to a moderate extent, but dependency categories are not at all shared in our multilingual models" (n.p.). They also report "that optimal conditions for lexical-semantic transfer may not be optimal for syntactic transfer" and that Johnson et al. (2017) found semantically equivalent sentences to form well-defined clusters in the high-dimensional space induced by an NMT encoder.

[1]LSTM (long short-term memory) is an NMT architecture in which a deep LSTM encoder projects a source sentence into a fixed dimensional vector from which a deep LSTM decoder generates (i.e., decodes) the target sentence. There are various alternative architectures, such as transformer-based sequence-to-sequence architectures which have recently gained more traction (e.g., Devlin et al. 2018).

The pioneering work in current NMT was conducted by Mikolov et al. (2013), who discovered that continuous word embedding spaces exhibit similar structures across languages. Word embeddings capture—to a certain extent—similarities of the syntactic-semantic properties of words as the models are learned from a huge corpus of text material. Word embeddings underlie the more sophisticated recent NN-based architectures used in recent NLP systems, including LSTM and transformers, such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018). Word embeddings are high-dimensional, continuous vector space models where each word is represented by 300–500 (and more recently up to 800) real-valued numbers. Mikolov et al. trained independent word embedding spaces for several languages and learned a linear translation matrix that would map word embeddings from one embedding space to another embedding space. They used a seed dictionary of the 5000 most frequent translations (i.e., rank <5K) that served as anchor points between the vector spaces. A mapping of a source word embedding into the target language is successful to the extent the two vector spaces are isometric. That is, the word embeddings in the source language space should have similar geometric placements as their translation equivalents in the target language space for a mapping to be successful and precise. Translation difficulties are expected to emerge where this is not the case, i.e., translation disturbances are likely to occur where the relations between source word embeddings and target word embeddings are not symmetrical and the vectors don't have similar geometrical placements in their respective language spaces.

Mikolov et al. (2013) report accuracy of 51% for test words in the frequency ranks of 5K–6K, which can be increased to more than 90% by taking into account several additional parameters, such as the size of the embedding spaces (ST vectors should be 2–4 times larger) and the *word projection accuracy* (see Sect. 2.2). They also report that translation precision decreases for less frequent words so that in the rank of 15K–19K, an accuracy of only 25% is achieved. Surprisingly, the approach works also for typologically unrelated language pairs such as English-to-Czech and even English-to-Vietnamese. However, there was no analysis as to why some parts of the spaces seem to be more isometric than others, other than due to word frequency.

Since their pioneering work, several studies have aimed at improving cross-lingual word embeddings finding more economic, closed-form solutions for producing the translation matrix (Xing et al. 2015) and reducing the size of the seed dictionary (e.g., Smith et al. 2017). Conneau et al. (2018: n.p.) show that a bilingual dictionary between two languages can also be obtained "without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way." However, to the best of the author's knowledge, no investigations exist to date that compare the structure of cross-lingual mapping of word embeddings with predictions that can be gleaned from bilingualism studies.

Within bilingualism studies, several models have been suggested that have been inspired by current multilingual NNs. The distributed feature model (DFM, de Groot 1992) suggests that word meanings are represented in the bilingual human mind as sets of (semantic) features. The DFM suggests that the way in which meanings are

lexicalized into word forms differs across languages, but the inventory of semantic features in both languages is identical. The DFM aims at explaining why bilinguals are slower to translate abstract than concrete words. de Groot suggests that concrete words are strongly connected to few salient features across the two languages, whereas abstract words are weakly connected to some features (De Groot 1992). Accordingly, concrete words have fewer possible translations across languages than abstract words, and translations of concrete words can be activated, retrieved, and produced more easily and faster than abstract words. Tokowicz and Kroll (2007) found that this effect emerges more clearly in less proficient L2 (second language) speakers and when the words had several possible translations.

The BIA model (Bilingual Interactive Activation; Dijkstra and van Heuven 1998), the BIA+ model (Dijkstra and van Heuven 2002), and most recently the Multilink model (Dijkstra et al. 2018) are based on similar assumptions. In their various model incarnations, Dijkstra suggests a two-stage approach of word recognition and production: in the first stage, word forms and meanings are automatically activated in the mind of a multilingual reader in a *nonselective* manner. This accounts for the evidence that bilinguals subconsciously access both languages simultaneously (e.g., Brill and Green 2013). In a second stage, a task/decision component, which is independent of the activation component, makes sure that the right item(s) are selected and produced from the network of activated meanings. While the BIA model assumes that orthographically similar items of both languages are activated in the reader's memory, the BIA+ model suggests that also phonologically similar forms are activated.

Multilink (Dijkstra et al. 2018) predicts that a stimulus word activates orthographic neighbors based on the similarity of phonetic or orthographic properties of the input word. The orthographic activation, in turn, activates associated semantic representations and subsequently their semantic neighbors. Dijkstra et al. explain this model with an example of English HOOD: first English orthographic neighbors are activated, e.g., {FOOD, HOLD, HOOT, ...}, and for Dutch speakers, also their Dutch orthographic neighbors {LOOD, HOND, HOOS, ...}: "orthographic representations will then begin to activate their meaning representations ... and semantically active representations 'spread' their activation to other units ... For instance, 'HOOD' may spread activation to the meanings 'HAT' or 'CAR', and 'FOOD' to 'HUNGRY'" (Dijkstra et al. 2018). This may also include multi-word units, such as NEIGHBORHOOD, HOOD OF MY CAR, FOOD SUPPLY, etc. Activated semantic representations will in turn activate their translations, such as KAP (Dutch for HOOD), also VOEDSEL (Dutch for FOOD), etc. This may then enter a recursive loop: semantic representations may recursively activate linked phonological representations in a language nonselective way, which, in turn, will activate their semantic representations and, successively, their semantic neighbor, etc. The process is likely to end when the task/decision component has selected the appropriate items.

Similar to DFM, Multilink also aims at explaining observed latencies of word comprehension and production times based on cross-lingual properties of words. In Multilink, more items in an orthographic neighbor set (ONS) and/or semantic

neighbor set (SNS) would trigger higher activation of the network which requires more energy and leads to *longer* retrieval times and for selecting the appropriate target item(s) in the set of possible candidates. In DFM, the "translation performance is a function of the number of conceptual nodes shared by a pair of translations: The more shared nodes, the *better* performance will be" (de Groot 1992, 1003, our italics). Both models make thus complementary predictions with respect to translation performance: Multilink predicts that the number of activated words should have a negative effect, while DFM suggests that the availability of a word in close proximity to the stimulus should have a positive effect on translation performance.

Despite the fact that such models of the mental lexicon of bilingual humans have existed since the 1990s and experimental research provides large amounts of evidence for various kinds of asymmetries, as summarized above, little cross-disciplinary research has been done in bilingualism and (machine) translation research communities. Within word embedding spaces, hypotheses from bilingualism research could be tested by comparing the similarity of word embeddings. Word vectors that share many features should be expected to be more similar to each other than word vectors with few or weakly connected features. Following this model, Sahoo and Carl (2019) have operationalized the computation of ONS by means of a Levenshtein distance[2] and SNS with word embeddings. The cardinality of those sets was correlated with behavioral data from monolingual copying, paraphrasing, and summarization tasks. They found an effect of the SNS size on translation latencies but not of the ONS size.

In this chapter, we develop various measures to assess neighbor sets in an English-to-Spanish translation task and investigate their correlation with concreteness scores and translation latencies. We use word embeddings and word2vec,[3] a computational continuous vector space model, for English and Spanish, and we describe a simple mechanism that maps the English word-embedding into the Spanish vector space. We develop three similarity measures that capture different aspects of the interlingual vector-space isometry: *word projection accuracy* (WPA), the *translation projection precision* (TPP)**,** and the *translation semantic similarity* (TSS). We assess those similarity measures against available translation norms (Prior et al. 2007), as well as against data from legacy translation data from the CRITT TPR-DB.[4] We test whether translation projections of concrete words are closer to their translations than abstract words.

In Sect. 2, we give examples of word embeddings for English and Spanish and how we map English word embeddings into the Spanish vector space and explain measures of interlingual vector-space isometry. Section 3 investigates Prior et al.'s (2007) English-to-Spanish translation norms (referred to as P-table) which

[2] *Levenshtein distance* is a distance between two strings. See also Heilmann and Llorca-Bofí (this volume, Chap. 8) and Do Carmo (this volume, Chap. 1).

[3] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

[4] CRITT TPR DB https://sites.google.com/site/centretranslationinnovation/tpr-db

contain concreteness scores and measures of translation ambiguity, among others. We investigate how translation ambiguity is represented in the vector space model. Section 4 applies the vector space model and translation norms to behavioral data from the CRITT TPR-DB. We assess to what extent insights from the translations out of context, as in the P-table, can be carried over to contextualized translation. We conclude in Sect. 5 with a discussion on how bilingual models (DFM and Multilink) are compatible with representations in the vector space model.

2 Word Embeddings

The recent success of NMT systems is, to a large extent, due to (1) an increased computer processing power (memory and speed) and (2) a technique which encodes words into vector spaces (word embeddings). Vector models are continuous space representations that have been used since the 1990s (LDA, LSA) in NLP with some sucess. More recently, NNs have been used to embed words as vectors into a continuous vector space of 100, 500, or even more nodes. These vectors are extracts of the hidden layers from a shallow NN and represent the contextualized usage—and to a certain extent the "meaning"—of the encoded words. The main benefit of vector space models and word embeddings is that they can be learned as an unsupervised task, which does not require pricey annotation of the data. However, they are computationally expensive to train, but luckily some companies (e.g., Google,[5] Facebook[6]) have made some models available that can be freely downloaded from the Internet. The development of word embeddings is a very active field of research, numerous topologies are being developed, and we are likely to see large enhancements and usability scenarios in the near future.

Even though word embeddings are trained on unannotated corpora, they are said to capture some elements of word meanings and thus make it possible to measure the semantic similarity between words, e.g., by using the cosine similarity. The cosine similarity between two vectors is a number between −1 and 1 where a value of 1 represents identity and a value of 0 indicates that the vectors have nothing in common, i.e., they are orthogonal. In our experiments, we use the GloVe[7] word embeddings for English and SBW-vectors-300[8] for Spanish. Both data sets consist of a set of pre-trained word vectors that are freely available from the Internet. The GloVe word-to-vec embeddings were generated from a corpus with six billion (English) words comprising a vocabulary of 400,000 lowercased word forms, each of which is associated with a 300-dimensional vector of real numbers (we use

[5] https://code.google.com/archive/p/word2vec/

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] Global Vectors for Word Representation (GloVe) https://nlp.stanford.edu/projects/glove/

[8] Spanish Billions Word Corpus https://crscardellino.github.io/SBWCE; see also /https://www.kaggle.com/rtatman/pretrained-word-vectors-for-spanish

6B.300d model). The SBW-vectors were also trained on a data set of approximately six billion (Spanish) words and have a (real-cased) vocabulary with over one million different word forms.

List 1: English words color-coded as belonging to three different domains (wortfeld).

['arrests', 'deaths', 'executions', 'kidnappings', 'murders', 'shooting', 'slaying', 'bag', 'briefcase', 'handbag', 'purse' 'wallet', 'doctor', 'hospital', 'midwife', nurse', 'physician']

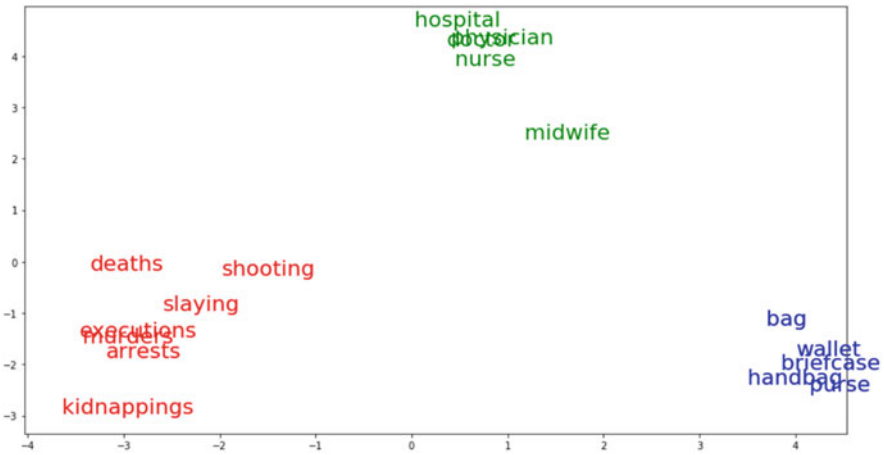


Fig. 1 Proximity of word embeddings of 17 English words from three-word clusters

These vector models can be loaded into a program, and words and their relations in the vector space model can be visualized. Figure 1 plots the proximity of 17 English word embeddings in a two-dimensional space. The words are taken from three semantically related groups of words that are listed in List 1 in different colors. We used Principal Component Analysis (PCA) to map the word embeddings for each of the 17 words from their 300-dimensional vector space into the two dimensions. The word cluster on the bottom left (in red) is related to the *killings*; another cluster in the bottom right (in blue) relates to *purse*; *nurse* is in the center of a third cluster in the top middle of Fig. 1.

The similarities between these words are learned fully automatically from their usage pattern in the monolingual six billion words corpus, and the visualization of those clusters sheds light on their learned (semantic) relatedness. For instance, the *killings* cluster (bottom left) is stretched out in the vector space with more items on the periphery, while the *purse* cluster (bottom right) seems to be denser with words in close proximity. Some clusters have outliers (e.g., "midwife" in the nurse-cluster, top middle), while others are more homogeneous (e.g., the *purse* cluster). Even when projected from a 300-dimensional vector space to only two dimensions, as in Figs. 1, 2, and 3, the three clusters are clearly separated, and some subtle meaning distinctions between words can be pinpointed.

List 2: Spanish translations from English words in List 1.
['arrestos', 'muertes', 'ejecuciones', 'secuestros', 'asesinatos', 'disparos', *'crímenes'*, 'bolso', 'maletín', 'cartera', *'maleta'*, 'billetera', *'doctor'*, 'hospital', 'partera', 'enfermera', 'médico']

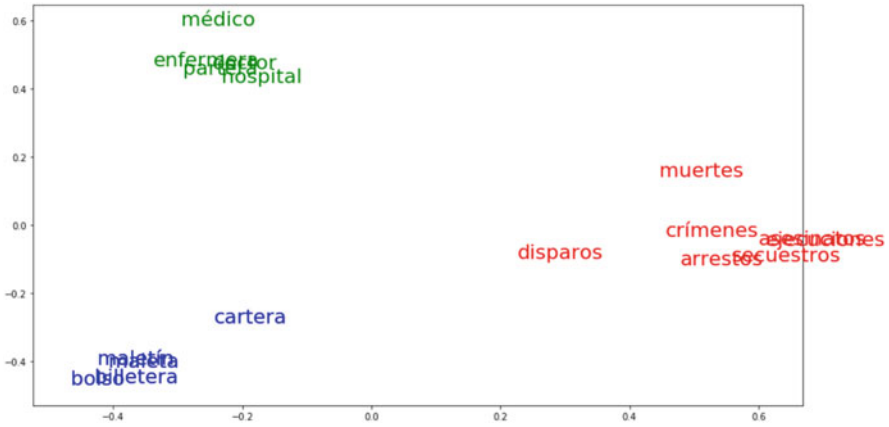


Fig. 2 Proximity of word embeddings of 17 Spanish translations in three-word clusters

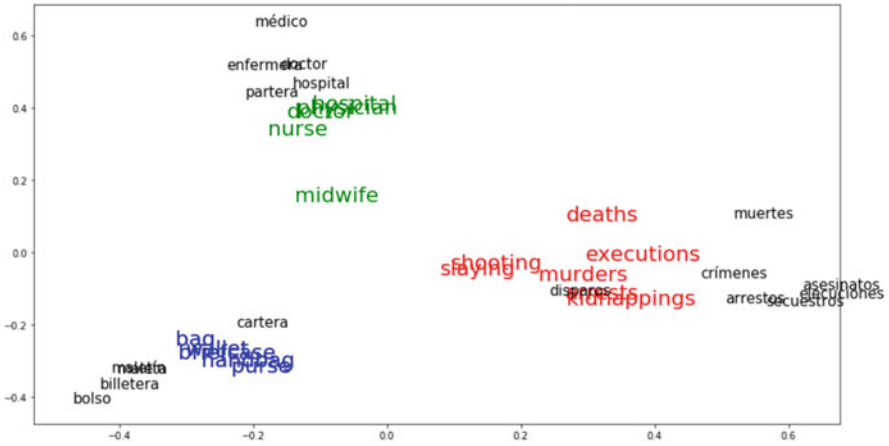


Fig. 3 Mapping of 17 English word embeddings into the Spanish vector space

The continuous vector space model illustrates how a stimulus word might activate semantically related words as predicted in the DFM and Multilink models discussed above. The DFM predicts that words in the close neighborhood of a stimulus may be more easily co-activated than those further away as they share more semantic features. Multilink predicts that a densely populated region with many words may take longer to activate than a region for which the language does not provide many competitors in the close neighborhood. For example, the word *nurse* may also activate *doctor*, *hospital*, and *physician* but to a lesser extent *purse*

and *wallet* which are further apart in the vector space. On the other hand, once a region in the vector space is activated, it may be easier to subsequently recognize and retrieve similar words in the proximity as numerous priming experiments suggest. Word embeddings may give us, thus, a possibility to quantify and verify hypotheses about spreading activation of semantic neighbors to the extent these vector spaces also model a mental reality.

List 2 gives Spanish translations of the English words in List 1, and Fig. 2 plots their Spanish word embedding, with the same color coding. As in Fig. 1, there are three clusters: in red, blue, and green.

However, the clusters are differently distributed in the English and Spanish vector spaces. Translations of the *killing* cluster are on the right side, while (the translations of) the *purse* cluster is on the bottom left, and *nurse* remains approximately at a similar place in Spanish vector space. Notice that the scaling is very different for the English and Spanish spaces. While the English words are distributed approximately across indexes Y:{−3,+4}, X:{−4,+4}, their translations in the Spanish space are distributed over a much smaller space covering from approximately Y:{−0.4, +0.6}, X:{−0.4,+0.6}. It is impossible to say what those indexes mean, as it is also unclear in the first place what each of the 300-dimensions in the original word embeddings represents from which these two-dimensional figures were derived. It is only the relative distance between the words that allow us to see similarities or differences, to cluster words, and to infer their (semantic) relatedness.

Even though the English and Spanish vector space models were trained on completely disjoint data sets, they seem to show some isometry with respect to some of the cluster properties. The English and Spanish clusters are similar to the extent that the *killing* cluster appears more scattered in both spaces, and the words on the periphery of the Spanish clusters (e.g., *disparos, muertes*) tend to be translations of words that are also in the periphery of the English clusters (i.e., *shooting, death*). As discussed above, words that have equidistant relations within and between word clusters in the two language spaces may be more easily directly mapped and might be translated more easily. An (approximate) isometry of relative distances for English and Spanish words enables their easy translation; while the extent to which subspaces are not isometric in terms of relative densities and distances of translation equivalents, it might become more difficult and require more effort to translate words that are members of those clusters.

2.1 Mapping Vector Spaces

Under the assumption that the projection between the two language spaces is approximately isometric, a projection matrix W can be trained that maps English word embedding vectors e into the Spanish vector space with projection landing site s'. The projection matrix W can be considered as (a set of) operators which ideally projects e close to its "real" translation s in the Spanish vector space. W can be trained so that $s' = W \cdot e$ gets close to its "real" Spanish translation s so

that the distance between s and s' is minimized. Following GloVe we can either use Stochastic Gradient Descent or a Closed-Form Solution to learn a translation matrix W from a small number of training samples. For the experiments reported below, we used the *Kronecker product identity*, for which a fast and closed formula exists.[9] To train this matrix, we followed Mikolov et al. (2013) using the 5000 most frequent English words from the BNC[10] and translated them (out of context) with Google Translate into Spanish. We discarded a total of 247 words which were not in either the GloVe or the Spanish SBW word2vec models. We split the remaining dictionary into 4400 words to train the translation projection matrix (W) and tested the matrix with the remaining 353 words.

For testing the accuracy of the translation projection matrix, we projected each English source word into the Spanish vector space. Then we retrieved the five Spanish words in the target vector space that were closest to the English word projection landing site. It was counted as a hit if the "real" (Google) translation was in a set of five closest neighbors. Out of 353 translations in the test set, there were 166 hits, which contained the "true translation" within the five most similar words as predicted by the translation projection matrix. Hence, the accuracy of the projection matrix is 0.47. We tested several alternatives for training W, changing the size and composition of the training and test sets, but could not substantially increase the precision of the translation matrix. We considered to proceed, as conclusions could be drawn even with 47% accuracy and inaccuracies might reveal systematic discrepancies and interlingual non-isometries.

Figure 3 reproduces the Spanish translation space from Fig. 2 and the projection of the 17 English word vectors into the Spanish vector space. It represents the original Spanish words forms from List 2 in smaller black letters and the English vector projections s' into the Spanish space colored with their original English word labels. Even though the Spanish translations and the English projections do not exactly overlap, there is a considerable similarity between the locations of the original Spanish (black) word embeddings and the English projections. The English projections pertaining to the *killing* cluster are swapped to the bottom right and the *purse* clusters to the bottom left, bringing it close to their Spanish translations. Some properties of the English source clusters seem to be preserved in the projections: the *killing* cluster seems more dispersed than the other two clusters, and the words on the cluster periphery resemble those in the English source. For instance, "midwife" is still an outlier in the projected cluster relatively remote from the (real) Spanish translation and the other projected words. The *purse* cluster is much denser in projection than in the original and the Spanish version.

[9] http://stackoverflow.com/questions/27980159/fit-a-linear-transformation-in-python

[10] The British National Corpus (http://www.natcorp.ox.ac.uk/) described below.

2.2 Similarity Measures

Given the English and Spanish vector spaces and a linear mapping of English word embeddings into the Spanish vector space, we define three similarity measures that capture various relations between the English projection landing site (s'), the closest neighbor, and a set of possible translations. These three similarity measures are illustrated in Fig. 4. Figure 4 shows in red the projection landing site of English *nurse* into the Spanish vector space. It shows in green the locations of Spanish semantically related words, *enfermera, enfermero*, *nodriza,* and *assistente*. We define three similarity relations:

1. The *word projection accuracy* (**WPA**) is marked as a black, thin arrow in Fig. 4. **WPA** is the cosine similarity between the English projection landing site and the closest word in the Spanish vector space. For instance, the Spanish word embedding of *enfermera* is the closest word to the English-to-Spanish projection of *nurse*. Its projection similarity is **WPA**(*nurse*) = 0.832. The Spanish word corresponding *projection similarity* (i.e., *enfermera*) is the Spanish word that corresponds *projection similarity* of *nurse*. Thus, the projection translation (**proTra**) of *nurse* is *enfermera*.
2. The *translation projection precision* (**TPP**) is indicated as red dashed lines in Fig. 4. It is the cosine similarity between the projection landing site of an English source word in the Spanish vector space and a given Spanish word (e.g., a translation). The translation similarity of *nurse* and *partera,* i.e., **TPP**(*nurse*, *enfermero*) = 0.635, while **TPP**(*nurse*, *nodriza*) = 0.6266, and **TPP**(*nurse*, *assistente*) = 0.597. If the translation is at the same time the closest position to the projection, the *translation similarity* is identical to the projection similarity, i.e., **TPP**(*nurse*, *enfermera*) = 0.832.
3. The *translation semantic similarity* (**TSS**) is operationalized as the average cosine similarity between every pair of words in a set of alternative translations

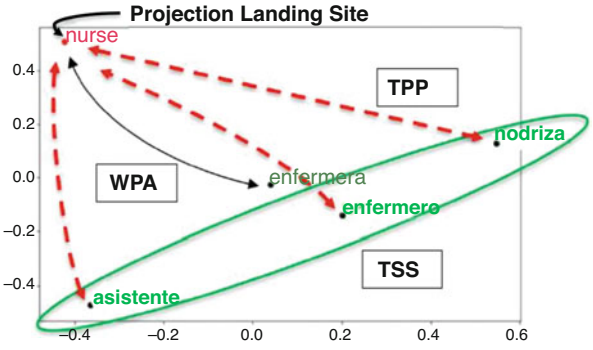


Fig. 4 Similarity measures between English projection and various Spanish neighbors

for the same source word.[11] For instance, the green oval circle in Fig. 4 contains three Spanish translations of *nurse*, and the mutual similarity of these three Spanish words **TSS**([*enfermero*, *nodriza*, *assistente*]) = 0.6242. A similar measure (*translation semantic variability* (TSV)) was first suggested by Bracken et al. (2017) as a continuous measure to quantify relatedness of form- and meaning-ambiguous translations. Translations of polynyms (unrelated in meaning) would receive a low score, whereas near synonyms would receive a high score. The **TSS** score of a set with only one translation will be 1.

The **WPA** indicates how remote the English word projection lands in the target space. We take it that **WPA** is an indicator of the isometry of the two language vector spaces. A word with a low projection accuracy might be uncommon, or it might be differently conceptualized in the target language, which leads to a non-isometric projection. **WPA** accounts for and measures an observation also made in Mikolov et al. (2013, np):

> If we apply the Translation Matrix to a word vector in English and obtain a vector in the Spanish word space that is not close to [any] vector of any Spanish word, we can assume that the translation is likely to be inaccurate.

The **TPP** indicates how close a given translation is to the English projection point. As a consequence of Mikolov's assumption, it follows that translations with a high **TPP** values should represent more reliable translations than those with lower **TPP** values. Finally, the **TSS** indicates how related (i.e., similar) a set of alternative translations is. Conneau et al. (2018) note that "some vectors, dubbed hubs, are with high probability nearest neighbors of many other points." **TSS** quantifies this *hubness* of a set of words as their average cosine similarity.

3 Translation Norms in the Vector Space

To assess the importance of translation ambiguity as a factor influencing translation performance, Prior et al. (2007) produce "Translation norms for English and Spanish." Eighty bilinguals were asked to name, among other things, Spanish translations for 670 single English verbs and nouns. The words were presented out of contexts, and half of the words received more than a single translation across participants; some words have up to eight or nine different translations. The number of different translations is indicated by a feature (**Ntra**), and each alternative translation has also a probability (**ProbT**).

[11] Leewnberg et al. (2016) suggest a "relative cosine similarity" which gives a high score to pair of words that have a high cosine similarity as compared to their top ten most similar words. If all words in the top ten most similar words have almost an equal cosine similarity, they will get a lower score. We did not see a (big) difference as compared to just using the average cosine similarity between all n^2 alternative translation.

The Prior data (henceforth P-table) contains a frequency score, but we also added frequency values to the words of the P-table that were extracted from the British National Corpus[12] (BNC). The BNC is a large, balanced corpus of English texts comprising more than 6 M sentences and more than 110 M words (tokens) and >1 M types (i.e., different form-POS combinations). We processed the BNC in several steps: tokenizing, tagging, lemmatization, and creation of a word-frequency dictionary, which resulted in more than 778.000 word forms with their frequency counts. In the analysis, we used the raw count, as well as a log transformation. We found that frequency information of the P-table very strongly correlates ($r = 0.95$) with the BNC frequency (BNCfreq) and the log-transformed version (lgFreq).

Most of the English source words and their Spanish translations have scores for imageability (**Simg** and **Timg**), concreteness (**Sconc** and **Tconc**), familiarity (**Sfam** and **Tfam**), and age of acquisition (**AoA**) for the source words and their translations, respectively. The data is publicly available free of charge and was assessed in this part of the study.[13]Prior et al. (2007) found that word frequency is negatively correlated with imageability (**Simg**) and the number of translations (**Ntra**) which can be confirmed with findings in Table 1.

The concreteness scores of the source word (**Sconc**) and their translation (**Tconc**) correlate also strong correlation ($r = 0.719$) when only considering the 602 (out of 670) English ST words with a total of 1062 Spanish translations for which source and target concreteness ratings are available in the P-table. Note that concreteness scores of source and target words correlate negatively but weakly ($\rho = -0.22$ and $\rho = -0.26$) with word frequency (**lgFreq**), which indicates that, surprisingly, more concrete words seem to be less frequent. It is also interesting to note that imaginability (**Simg**) strongly correlates with both source and target concreteness scores ($r > 0.7$). The number of translation alternatives (**Ntra**) correlates with all other values negatively and weakly, and in some cases not significant. Concreteness of translation (**Tconc**) and familiarity (**Tfam**) both decrease weakly as the number of translation alternatives increases ($r = -0.134$ and $r = -0.167$) indicating that less concrete and less familiar translations tend to have more translation variation.

3.1 Alternative Translations and the Vector Space

The number of possible different translations for an ST word has been shown to affect performance in single-word recognition and production tasks (e.g., Tokowicz and Kroll 2007) as well as translation production of coherent texts (e.g., Schaeffer et al. 2016; Carl and Schaeffer 2017a). Tokowicz et al. (2002) showed that words that have more than a single translation are judged to be less semantically similar to each of their possible translations than are words that have only a single translation.

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

[13]The data can be downloaded from https://link.springer.com/article/10.3758/BF03193001

Table 1 Pearson (r) and Spearman (ρ) correlation for various features in the P-table

	Ntra	Pfreq	lgFreq	Sconc	Tconc	Tfam	Simg	
Ntra	–	−0.085**	*−0.025*	−0.063*	−0.134***	−0.167***	−0.113***	Pearson correlation r
Pfreq	*−0.058*	–	0.663***	−0.163***	−0.145***	0.159***	−0.231***	
lgFreq	*−0.012*	0.922***	–	−0.250***	−0.275***	0.228***	−0.331***	
Sconc	*−0.056*	−0.169***	−0.226***	–	0.718***	*0.032*	0.883***	
Tconc	−0.139***	−0.218***	−0.264***	0.719***	–	0.064*	0.705***	
Tfam	−0.188***	0.241***	0.264***	*0.056*	0.09**	–	*0.03*	
Simg	−0.123***	−0.223***	−0.294***	0.881***	0.717***	*0.046*	–	
	Spearman correlation ρ							

The asterisks indicate the significance 345 levels of the correlations, where one The The asterisk (*) refers to a significant effect (p-value <0.05), two asterisks (**) for a highly significant effect (p-value <0.01), and three asterisks (***) for a very highly significant effect (p-value <0.001).

In this section, we investigate the relation between translation ambiguity and the three vector space measures, **WPA**, **TPP**, and **TSS** in the P-table.

Altogether, there are 1408 English-to-Spanish translations in the P-table for the 670 ST words which amount to an average of 2.1 Spanish translations per ST word. However, only half of the 670 words have more than a single translation. A distribution of alternative translations is shown in Fig. 5. Prior et al. (2007, 2013) point out that the reason for ambiguity can be very different. The English word *soap*, for instance, has two meanings: the material used for washing and in the meaning of *soap opera*, which translates into Spanish *jabón* and *telenovela*, respectively. The English verb *to fire* translates into *despedir,* while the noun *fire* is rendered as *fuego*. The English word *cook* can be a verb, which translates into *cocinar*, or the person (i.e., the noun) which translates into Spanish *cocinero*. The verb *know* covers knowing facts and knowing people, which are two distinct verbs in Spanish, *saber* and *conocer*, while *hair* may be translated synonymously into *cabello* or *pelo.*

These translations show a different degree of similarity in meaning vs. form (e.g., *jabón*, *telenovela* vs. *cabello*, *pelo*). In order to capture these differences in a single measure, Bracken et al. (2017) suggest replacing the dichotomy *form-ambiguous* (i.e., synonyms) and meaning-ambiguous (polynym) translations with the concept of translation semantic variability "which quantifies the degree of semantic relatedness between the translations of translation-ambiguous words" (Bracken et al. 2017: 784). Similarly, the **TSS** computes a "hubness score," i.e., the average cosine similarity between a set of alternative translations in the Spanish vector space. A high **TSS** score would indicate close (synonym-like) relation, as in the case of *cabello* and *pelo* (**TSS** ([*cabello*, *pelo*]) = 0.919) or *saber* and *conocer* (**TSS** ([*saber*, *conocer*]) = 0.779), which are form-variant translations of *hair* and *know*, respectively. Polysemous words such as *soap* with different translations may have lower scores, **TSS** ([*jabón*, *telenovela*]) = 0.595. Note that **TSS** is likely to be smaller as the number of alternative translations increases, such as in examples *act* and *edge* in Table 2.

Names of months seem to cluster closely together (higher *hubness*) in the Spanish vector space, with, e.g., **TSS** (["abril", "junio"]) = 0.987. The close proximity

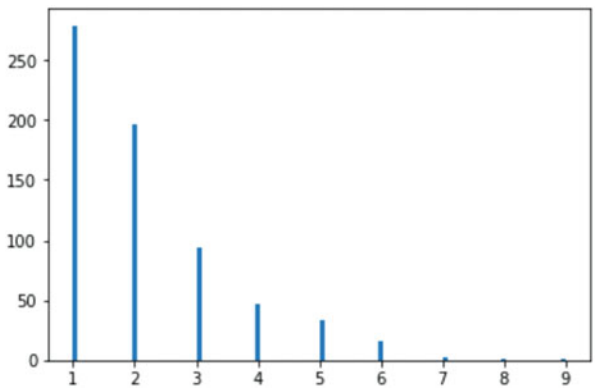


Fig. 5 Distribution of translation alternatives

Table 2 Alternative translations from the P-table with various features

Word	Translation	ProbT	TPP	WPA	TSS	lgFreq	HTraP
Nurse	Enfermera	1.0	0.832	0.832	1.0	7.92	0.0
Father	Padre	0.950	0.892	0.892	0.640	9.88	0.286
Father	Papa	0.050	0.328	0.892	0.640	9.88	0.286
Potato	Papa	0.941	0.301	0.696	0.682	6.60	0.323
Potato	Patata	0.059	0.625	0.696	0.682	6.60	0.323
Hair	Cabello	0.400	0.740	0.786	0.919	9.47	0.971
Hair	Pelo	0.600	0.750	0.786	0.919	9.47	0.971
Know	Saber	0.650	0.787	0.841	0.779	11.68	0.934
Know	Conocer	0.350	0.543	0.841	0.779	11.68	0.934
Soap	jabón	0.947	0.511	0.592	0.595	7.06	0.299
Soap	Telenovela	0.053	0.425	0.592	0.595	7.06	0.299
Act	Actuar	0.632	0.585	0.723	0.576	9.31	1.168
Act	Acto	0.316	0.597	0.723	0.576	9.31	1.168
Act	Pretender	0.053	0.511	0.723	0.576	9.31	1.168
Edge	Borde	0.385	0.664	0.761	0.425	8.84	2.039
Edge	Esquina	0.308	0.565	0.761	0.425	8.84	2.039
Edge	Filo	0.154	0.440	0.761	0.425	8.84	2.039
Edge	Orilla	0.077	0.549	0.761	0.425	8.84	2.039
Edge	Limito	0.077	0.263	0.761	0.425	8.84	2.039

of several semantically related words may lead to a preferred wrong translation e.g., English. *june* ➔ *abril* (not shown in the table). However, the correct Spanish translation (i.e., *junio* in this case) is among the five closest translations.

Table 2 shows a number of alternative translations from the P-table with their respective translation probability values (**ProbT**), the three vector space measures, word frequency, and word translation entropy (**HTraP**). As discussed in detail in various chapters in this volume (e.g., Carl this volume, Chap. 5; Ogawa et al. this volume, Chap. 6), the word translation entropy (**HTra**) is computed based on the translation probabilities of alternative translations according to[14]:

$$\mathrm{HTra}(A) = \sum\nolimits_{\mathrm{ProbT} \in A} \mathrm{ProbT} \times \log\left(1/\mathrm{ProbT}\right)$$

Higher **HTra** (and also **HTraP**) values indicate more equal distribution among multiple possible translations, while low **HTra** values have few translations. For instance, *Enfermera* is the only translation for *nurse* in the P-table and has, thus, the translation probability **ProbT** = 1 and **HTraP** = 0. The English word *father* has

[14]We use the formula to compute word translation entropy for the translation alternatives in the P-table and later in Sect. 5 for the BML12 study. To avoid confusion, we call the values computed from the P-tables **HTraP** and from the BML12 tables **HTra**.

two Spanish translations in the P-table *padre* and *papa*, which have very different distributions. Spanish *padre* occurs 95% of the time, while *papa* makes up only 5% of the translation choices for *father*. The **HtraP** value is *0.286,* which is, accordingly, relatively low. In contrast, *hair* has also two different translations in the P-table, but their distribution is more equal, with 60% of the translations being *pelo* and 40% *cabello.* This more equal choice leads to a quite higher **HTraP** value of *0.971*. As can be seen in Table 2, with more alternative translations that are available for a source token, the **HTraP** value tends to also increase. The **HTraP** value indicates the complexity of a lexical translation choice, where the selection of an item from a list of equally distributed translation alternatives is more effortful than if one translation option is more entrenched. As Campbell (2000, 30) puts it, "the more complex choices a translator has to consider, the more effortful is the translation of a particular item."

In contrast to **HTraP,** the **TSS** measure does not take into account the probability of the translations. Both measures assess translation choices differently. While **TSS** measures the density of the vector space in which the translation alternatives are located, **HTraP** quantifies the distribution of the translation choices. They both measure potential difficulties to discriminate between alternatives. As we will discuss in Sect. 4.2, we expect a strong correlation between these two measures.

Given the isometric properties of the vector spaces, the DFM would suggest that **TPP** correlates with **ProbT:** an entrenched translation (high **ProbT**) is likely to share many features with the ST word (high **TPP**), while a marginally related translation (low **TPP**) is likely to have also a low **ProbT**. For instance, Spanish *padre* is the closest translation of the English-to-Spanish projection of *father,* which is also the most likely translation in the P-table (**ProbT** $= 0.95$, **TPP** $= 0.892$), while the alternative translation *papa* has a translation probability of **ProbT** $= 0.05$ and a translation precision of **TPP** $= 0.328$.

However, there are exceptions to this rule. For instance, English *potato* has two translations: Spanish *papa* (**ProbT** $= 0.941$, **TPP** $= 0.301$) is much more likely than *patata* (**ProbT** $= 0.059$, **TPP** $= 0.625$), yet their translation precision values do not reflect their translation probabilities. This might be another indicator for non-isometric vector spaces. A low **TPP** for *papa* might be due to the fact that *papa* has also the meaning *father* and *pope,* but there is only one point in the vector space to encode each word, irrespective of its polysemous status. Spanish *papa* might thus occupy a place somewhere in the middle between the Spanish equivalents of *pope*, *father,* and *potato,* which makes it remote to all of them.[15]

[15]More recent word embeddings, such as BERT, can take different contexts into account. It would then, however, be difficult to map lists of single word translations, since there is no context.

3.2 Translation Norms and the Vector Space

The correlation matrix in Table 3 shows Pearson and Spearman's correlations of the features in the P-table discussed above. From the original 670 ST words, we excluded 83 English source words that had no **Sconc** value and 286 translations which had no **Tconc** value. Since some of the words have no value in both, we took out a total of 346 words, which left us with 1062 translations of 602 English source words. Table 3 shows some of the values that are of more interest to us here; some observations are listed below.

- The number of different translations (**Ntra**) correlates negatively and in most cases significantly with all other vector space measures. In particular, as the number of alternative translations increases:
  - **TPP** decreases ($\rho < -0.5$), indicating that to the extent an ST word has more alternative translations, each of the translations is also semantically more remote from the *source word*.
  - **TSS** decreases ($\rho < -0.8$), corroborating the hypothesis that the more translations a word allows for, the more they are also semantically more distant from each other.
- A moderate correlation ($r > 0.5$) between **WPA** and **TPP** suggests that as the source and target vector spaces become less isometric (as measured by **WPA**) and also the translation precision decreases.
- The moderate correlation between **WPA** and word frequency (**lgFreq**) ($r > 0.5$) suggests that more frequent words tend to have better mapping into the target space and that more frequent words share more isometric vector spaces across the two languages.
- The concreteness scores of the source (**Sconc**) and the target (**Tconc**) correlate—surprisingly—negatively though weakly with **WPA** ($r = -0.16$ and $r = -0.11$). A multivariate regression[16] shows, however, that this negative correlation is a spillover effect due to the negative correlation between word frequency (**lgFreq**) and number of translations (**Ntra**) and the two concreteness values. The multivariate analysis suggests that none of **Sconc** or **Tconc** has a significant effect on **WPA**; however, **lgFreq** and **Ntra** have. The negative correlation in Table 3 is thus a word and translation frequency effect, and not related to the concreteness value.
- A similar analysis can be made with the two concreteness scores (**Sconc** and **Tconc**) and their effect on **TSS**. Table 3 indicates that **Tconc** is weakly but significantly correlated with **TSS** ($r = 0.16$), while **Sconc** is not. Running the same multivariate analysis with **TSS** as dependent variable shows that **Sconc,**

[16] We used the ols function in python's statsmodel library to compute (`WPS ~ Sconc + Tconc + lgFreq + Ntra`) and obtained **Sconc**: $p = 0.69$, **Tconc**: $p = 0.99$ and **lgFreq**: $p < 0.001$, **Ntra**: $p < 0.001$.

Table 3 Correlation matrix for vector space measures in the P-table

	TPP	WPA	TSS	Ntra	ProbT	lgFreq	HTraP	Sconc	Tconc	
TPP	–	0.55***	0.6***	−0.46***	0.59***	0.18***	−0.46***	−0.06	0.09**	Pearson r
WPA	0.59***	–	0.33***	−0.28***	0.19***	0.5***	−0.28***	−0.16***	−0.11***	
TSS	0.62***	0.36***	–	−0.81***	0.67***	0.01	−0.81***	*0.02*	0.16***	
Ntra	−0.53***	−0.3***	−0.87***	–	−0.65***	−0.03	0.94***	−0.06*	−0.13***	
ProbT	0.62***	0.19***	0.66***	−0.71***	–	−0.02	−0.67***	0.06	0.17***	
lgFreq	0.16***	0.44***	−0.01	−0.01	−0.05	–	−0.03	−0.25***	−0.28***	
HTraP	−0.5***	−0.3***	−0.83***	0.95***	−0.67***	−0.02	–	−0.05	−0.11***	
Sconc	−0.02	−0.12***	0.04	−0.06	0.07*	−0.23***	−0.05	–	0.72***	
Tconc	0.11***	−0.07*	0.17***	−0.14***	0.19***	−0.26***	−0.11***	0.72***	–	
	Spearman correlation ρ									

Tconc, and **Ntra** have a significant effect ($p < 0.001$) on **TSS**, while **lgFreq** has no effect ($p = 0.501$). It thus seems that concrete words tend to cluster closer together in the Spanish target vector space than less concrete words.

- The moderate negative correlation ($r = -0.46$) between **TPP** and **HTraP** indicates that less entrenched translation choices (i.e., high **HTraP**) are more likely to occur when translations are remote from the source word projection landing site. The moderate correlation ($r = 0.55$) between **TPP** and **WPA** suggests that less isometric translation mappings (lower **WPA**) also lead to more remote translations (lower **TPP**).

4 Mapping Translation Norms and Behavioral Data

This experiment investigates how the results from P-table described in Sect. 3 for single word translations out of context scale to translations in context. The CRITT TPR-DB contains a large number of translations with recorded keystrokes and production times. We extract an English-to-Spanish subset from the CRITT TPR-DB that matches entries in the P-tables and adapt the vector space measures to contextualized translations. We investigate whether translations in context have similar properties as single word translations discussed in the previous section. The only related study we know of is Prior et al. (2010). However, they use translated texts and thus do not have access to behavioral data, such as translation production times.

4.1 Behavioral Data

For our experiment, we used the BML12 study, which is a subset of the *multiLing* dataset from CRITT TPR. The multiLing corpus consists of six different English source texts with a total of 40 segments and 847 words. These texts were translated, post-edited, and edited by 31 Spanish translation students into Spanish, into their L1, who produced a total of 184 texts with 25,939 target words. The data was recorded in 2012, using Translog-II and a Tobii eye tracker, TX120. The logging data was uploaded to the Translation Process Research (TPR) database and manually aligned and is publicly available.[17] For this study, we use the ST tables, which provide, among other things, information about the translation product such as lemmatization, and PoS tags of the source words, different alternative translation renderings, measures of variation and entropy of the produced translations, etc., as

[17]The database can be downloaded from https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies. See Carl this volume, Chap. 5, for a description.

well as process information, including production times for each ST word, produced keystrokes, number of insertions and deletions, etc.

The 847 English words in the six *multiLing* texts are made up of 408 different types (lemmas). From these 408 different lemmas, 77 match entries in the P-table. Some lemmas have different forms in the *multiLing* corpus, so that we ended up with 139 word forms that are extracted from the BML12 corpus. Thus, around 16% of the ST words in the BML12 study can be enriched with information from the P-table. As every English source text is translated by up to 31 different translators, this leaves us 4311 translated words (observations). However, around 2/3 of these translations were produced in a post-editing or blind post-editing task, and due to an artifact that relates to delayed copy and pasting in the key-logging tool, we only kept words that have a production duration of more than 20 ms. This filter further reduces the number of translated words from the original 25,939 observations to 1472 translated words, with an average of almost 11 alternative translations for each of the 139 different English source words. For these 1472 observations, we copy **TSS** (as **TSS_P)** and **Sconc** information from the P-tables over into the BML12 ST tables so that it contains the product and process information form the TPR-DB and some of the features from the translation norms.

List 3: Alternative translations for "acting" in context.

['actuando', 'actuado', 'actuaba', 'había actuado', 'estado actuando', 'había estado actuando', 'portaba', 'mantenido un comportamiento'].

4.2 Adaptation of Vector Space Measures

In contrast to the P-tables, many of the translations in the BML12 table are groups of more than one word. For instance, English *act* has—according to the P-table—three Spanish translations, *pretender*, *acto*, and *actuar* (see Table 2). However, in our reduced BML12 dataset, the translations depend on the context in which the words occur. For instance, there are 12 Spanish translations from 12 translators for the sentence *Norris had been* ***acting*** *strangely* in our BML12 subset. The word *acting* in this context has been translated in eight different ways, as shown in List 3.

The translation *actuando* was produced five times, and each of the other seven translations occurred only once (which makes a total of 12 translations). Ten translators actually used the verb *actuar* in their translations, however, in differently inflected versions. Two out of the 12 translators chose to render the translation in a different form as those given in the P-table, i.e., *mantenido un comportamiento* and *portaba*, while the P-table lists two translations *acto* and *pretender* that were not used by the translators.

While similarity scores between different translations in the P-table are just the cosine similarity between their word embeddings, for translation groups in the BML12 study, we might need to compute the similarity between phrases with several words. To do this, we average over the most similar correspondences in the two phrases. For instance, to compute the translation precision **TPP** (*acting,*

mantenido un comportamiento), first *acting* is projected into the Spanish vector space. Then the similarity between the landing site vector and each of the word embeddings of *mantenido, un,* and *comportamiento* is computed. The maximum of the three values is taken to be the similarity for **TPP.**

For the computation of the **TSS** in the BML12 context, we need to compute the average similarity of a set of groups of words. For each token in group 1, we compute the most similar token in group 2 and average over their similarity scores of tokens in group 1. We define the **TSS** score for a set of n alternative translations to be the average of all the $n \times n$ similarities in the set. This ensures that **TSS** $= 1$ for a set with only one translation (i.e., $n = 1$), since the cosine similarity between two identical vectors is 1. For instance, to compute **TSS** for eight different translations in List 3, we compare each of the 8×8 translations, some of which are phrases. The similarity between *había actuado* and *estado actuando* is the average of two similarities (había × *estado actuando* and *actuado* × *estado actuando*) which results in a **TSS** score of 0.51. With this method, the TSS score of List 3 is 0.5461.

4.3 Correlation of Contextualized Translation

Table 4 shows the correlation matrix of the 1472 translations from the BML12 study with various vector similarity scores as well as the **Sconc** concreteness score from the P-table.

- The negative and rather strong correlation between **TSS** and **HTra** ($r = -0.68$) indicates that with increasing word translation entropy, the semantic similarity between the different translation alternatives decreases. This finding matches those of single-word translations in the P-table (see Table 3), where we observe a very strong negative correlation between **TSS_P** and **HtraP** ($r < -0.8$). It also corroborates previous findings (e.g., Tokowicz et al. 2002) who report that ST words with more different translations are semantically more remote from their various translations than ST words with one or fewer translations.
- The moderate correlation ($r = 0.32$) between the BML12 **TSS** score and the P-table **TSS_P** score indicates that results from single-word experiments may carry over to translations in context to a limited extent. There are very likely much fewer meaning variants in the set of contextualized alternative translations. For instance, an occurrence of *soap* in a text would very likely leave no doubt whether the washing material or a melodrama is meant, while this is different with a list of decontextualized words. Accordingly, there is only a weak correlation ($r = 0.23$) between the number of translations in the P-table (**Ntra**) and the equivalent feature in the BML12 study (**AltT** not shown in the table); also the word translation entropy (**HTra** and **HTraP**) values correlate only very weakly ($r = 0.184$, also not shown in the table).
- The moderate correlation ($r = 0.43$) between **TPP** and **WPA** suggests that (1) the projection landing site is not always systematically close to the word embeddings

of the actual translation and (2) translation alternatives are semantically less related if they are not in the proximity of the source word projection landing site. This second conclusion is also supported by the negative correlation between **TPP** and **HTra** ($r = -0.41$), which indicates that a wider variation of less entrenched translation choices is likely to occur if the translations are further apart from the English projection landing site.

- A weak but significant correlation ($r = 0.17$) between translation production duration (**lgDur**) and word translation entropy (**HTra**) has been reported in previous research (e.g., Schaeffer et al. 2016, Carl and Schaeffer 2017b, Carl this volume, Chap. 5). It suggests that the choice of a translation rendering from a larger number of equally likely possibilities is linked to higher effort and, thus, to longer production durations. It is also to be expected that **lgDur** correlates negatively with the word frequency (**lgFreq**) ($r = -0.32$) since more frequent words tend to be shorter and typing patterns more automatized, hence quicker to type and easier to produce than longer (or more infrequent) words.
- As previously noted, concreteness (**Sconc**) correlates negatively with frequency. Surprisingly, concreteness scores are negatively correlated to duration (**lgDur**) suggesting that concrete words take *longer* to produce, thus contradicting the assumptions underlying the DFM. They are also weakly correlated with **TPP** and **WPA** indicating that more concrete words tend to have translations further away from the English word embedding projection.

It is particularly the latter points that made us look deeper into the contributions of various variables on production duration.

4.4 Correlation Contextualized Translations

As Table 4 suggests and in contradiction to the predictions of DFM, translation production duration (**lgDur)** correlates positively with **Sconc** (i.e., more concrete words need *longer* to translate) and with **WPA,** and only insignificantly with **TSS**. To understand the contributions of these variables, we ran a hierarchical regression analysis starting with six independent variables (**WPA**, **TPP**, **TSS**, **HTra**, **lgFreq**, and **Sconc**) to assess their effect on dependent variable **lgDur**. It turned out that none of **Sconc** and **TSS** were significant. As in Table 3, it appears that the correlation between **Sconc** and **lgDur** is a side effect of the fact that concrete words are less frequent ($r = -0.65$).[18] Similarly, the effect of **TSS** and **HTra** is strongly correlated ($r = -0.68$) so that the effect of **TSS** on **lgDur** is neutralized by **HTra.** The effect on **lgDur** can be explained to some extent by the four remaining variables, which resulted in the following model in R:

```
lm(lgDur ~ TPP + WPA + HTra + lgFreq, data = df)
```

[18] Also de Groot (1992) indicates that only high-frequency words show the concreteness effect.

Table 4 Correlation of vector space measures behavioral and norm features in the BML12 table

	TPP	WPA	TSS_P	TSS	lgFreq	lgDur	HTra	Sconc	
TPP	–	0.43***	0.20***	0.51***	−0.11***	*0.01*	−0.41***	0.23***	Pearson *r*
WPA	0.47***	–	0.52***	0.24***	−0.11***	0.24***	*−0.04*	0.22***	
TSS_P	0.26***	0.49***	–	0.32***	0.10***	0.13***	−0.22***	0.10***	
TSS	0.57***	0.27***	0.31***	–	*0.03*	−0.06*	−0.68***	0.16***	
lgFreq	−0.13***	−0.19***	*−0.02*	*−0.03*	–	−0.32***	−0.23***	−0.65***	
lgDur	*0.04*	0.26***	0.17***	*−0.03*	−0.35***	–	0.17***	0.23***	
HTra	−0.40***	*0.01*	−0.19***	−0.67***	−0.23***	0.20***	–	*0.02*	
Sconc	0.25***	0.26***	0.29***	0.18***	−0.67***	0.26***	*0.04*	–	
	Spearman correlation ρ								

Table 5 Effects on production duration

	Estimate	Std. error	*t*-Value	Pr(>|*t*|)
(intercept)	1.34115	0.48925	2.741	0.0062**
TPP	−0.72280	0.22992	−3.144	0.0017**
WPA	6.20959	0.68840	9.020	<2*e*−16***
HTra	0.08183	0.03126	2.617	0.0090**
lgFreq	−0.40232	0.03560	−11.301	<2*e*−16***

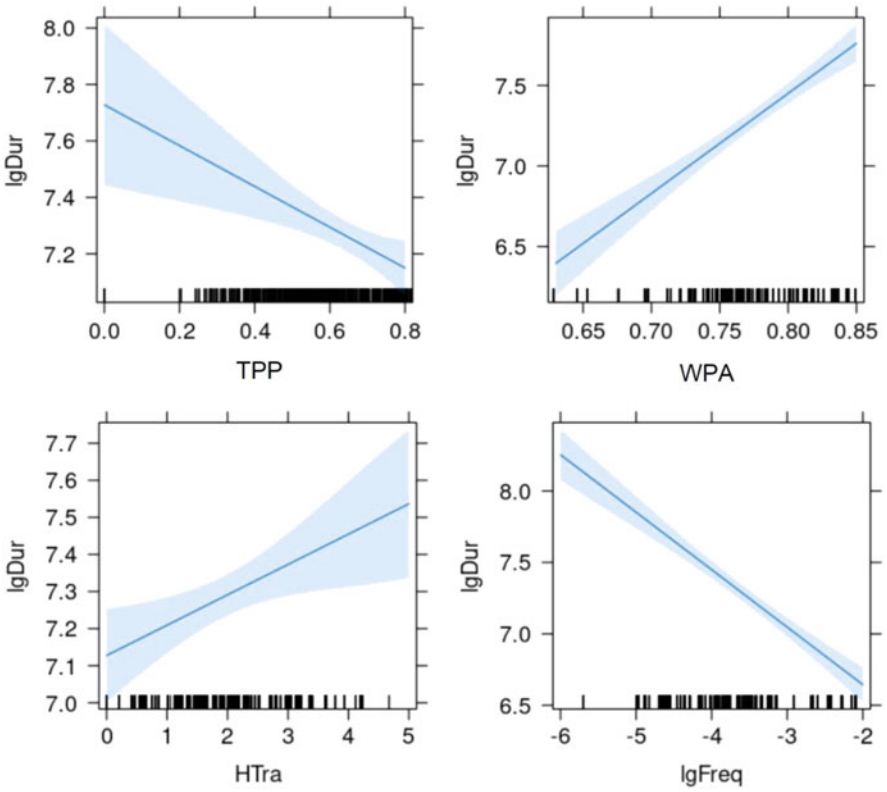


Fig. 6 Effects of **TPP**, **WPA**, **HTra**, and **lgFreq** on **lgDur**

The overall fit for this model was $r^2 = 0.168$ ($F = 69.11$, df $= 1368$, p-value: <2.2e−16). The summary in Table 5 indicates that all four independent variables have a significant effect on the dependent variable **lgDur**, **WPA** and **HTra** have a positive effect, and **TPP** and **lgFreq** have a negative effect on **lgDur** (see Fig. 6). While the effect of word frequency (**lgFreq)** and word translation entropy (**Htra**) on **gDur** has already been reported elsewhere (e.g., Schaeffer et al. 2016), the effect of the other two variables, **TPP** and **WPA,** has not yet been observed. Interestingly, **WPA** and **TPP** have the opposite effect, respectively, positive and

negative, suggesting that larger distances between the projected source word and the actually realized translation imply more effortful translation.

For instance, given their **WPA** and **TPP** values, we can hypothesize that translating English *edge* into Spanish *orilla* or *limito* may be more effortful than translating *nurse* as *enfermera*. The current model may explain this in terms of isometry in the English and Spanish vector spaces. For the translation *nurse-enfermera*, the English projection of *nurse* is close to the word embedding of the produced translation *enfermera* (**WPA** = **TPP** = 0.832), while for *edge*, the projection accuracy is **WPA** = 0.761, while the translation precisions for *orilla* and *limito* are **TPP** = 0.549 and **TPP** = 0.263, respectively. A purely frequency based model would not allow for such conclusions, as the frequency of *nurse* is lower (and translations presumably less entrenched) than of *edge* (see Table 2).

In this view, a source word activates a region with certain syntactic-semantic properties in a target vector space which may or may not be populated by expressions in the target language vocabulary. A translation is particularly time-consuming to produce if there are target language expressions in the proximity of the activation center, but the translator chooses—for some reasons—to produce an alternative that is further away from the center of the stimulated area. The variation of possible translation choices (**HTra**) adds to translation latencies to the extent that there are several equally likely translation possibilities for one source word.

5 Discussion

The chapter introduces word embeddings as a new method to investigate, model, and explain findings from TPR and to assess empirical data and theoretical considerations within a connectionist framework. Word embeddings (e.g., word2vec) have been shown to capture aspects of word meaning with a huge potential to advance all areas of NLP, language research, NMT, and artificial intelligence in general. Word embeddings are trained on huge text corpora, and the learned vectors represent syntactic-semantic properties of words as points in a high-dimensional continuous vector space. The vector space model suggests that meaning variations are continuous, rather than discrete, and there are some points in the vector space for which languages provide words. It has been discovered (Mikolov et al. 2013) that those vector spaces are to some extent isometric across different languages so that a projection matrix can be trained to map and thus translate words across the languages.

We report two experiments using those word embeddings and projection matrix to assess findings of "translation norms" and from a TPR-DB study. For our experiments, we used two pre-trained monolingual English and Spanish word embeddings, and we trained an English-to-Spanish projection matrix on a small set of English-to-Spanish translations. The projection matrix projects English word embeddings into the Spanish target vector space, where the English projection can

be closer to or further apart from possible Spanish translations, and thus activates more or less strongly possible Spanish translation options.

The notion of "spreading activation"—as alluded to in several connectionist models of the bilingual mind—can be modeled as the (cosine) similarity between two (or more) vectors: word embeddings in close neighborhood are more likely to be co-activated than word embeddings further away, as activation is more easily spread to close neighbors. The spreading assumption predicts that co-activation facilitates the retrieval and production of words. The further an item is away from a center of the activation, the more unlikely, time-consuming, resource-intensive, or effortful will be the retrieval or production. Translation is smooth and unproblematic to the extent the source and the target language vector spaces are isometric, and source expressions and their translation(s) are close neighbors. Translation becomes effortful if target items are retrieved from positions that are distant from the center of activation in the vector space. With a vector space model, we can thus test or corroborate predictions about priming, linguistic co-activation, and assumed translation effort (i.e., expended energy) by measuring the (cosine) similarity between word embeddings.

Several models of the bilingual mind have been suggested—including the DFM (de Groot 1992) and Multilink (Dijkstra et al. 2018)—which presume (localist) connectionist models in which word activations are spread through connections and activate related concepts and meanings. Kroll et al. (2010, 6), for instance, assume that "conceptual features are sampled and linked to word forms," but it is unclear what exactly those features might be. In contrast to the localist interpretation of cognitivist models, a continuous vector model represents "meaning" without recourse to any explicit linguistic features at all. Distributed representations provide us with the possibility to represent words as points in a continuous n-dimensional hyperspace, where the notion of "feature" (in a linguistic sense) may only emerge post hoc, e.g., when analyzing and comparing groups of words and how they cluster in the vector space. However, linguistic features are not—a priori—representations on which the operations in the network are based.

In this chapter, we mainly investigated three aspects:

1. Are concrete words easier to translate than less concrete words?

   The impact of concreteness ratings on language processing and translation has been of great interest (Tokowicz and Kroll 2007; Brysbaert et al. 2014; Dijkstra et al. 2018). The DFM (de Groot 1992) postulates an effect of "concreteness" on translation performance due to larger number of overlapping features between the source and the translation. While we observed a weak correlation between concreteness ratings and **WPA** scores ($r = 0.22$), we found a significant correlation ($r = 0.23$) between translation duration and concreteness, suggesting that more concrete words take *longer* to translate. However, concrete words have also lower frequency in our corpora ($r = -0.65$), and it was argued—among others in de Groot (1992, 1002)—that we may "expect shorter translation times and less translation errors for high-frequency words than for low-frequency words." A multivariate analysis taking into account frequency effects confirms

this expectation, suggesting that the apparent increased translation duration of concrete words is actually due to their lower frequencies.

2. Can Multilink and DFM predictions be substantiated with the three vector space measures?

- Multilink (Dijkstra et al. 2018) predicts that more activated translations imply longer processing time. Several of our findings confirm this hypothesis:
  - **HTra** has a significant negative effect on **lgDur** indicating that higher translation variation has a positive effect on production duration. More translation variation is thus more difficult to process. This effect is also observed in other studies, e.g., Lacruz et al. (this volume), Chap. 11) and Carl (this volume), Chaps. 5 and 9).
  - **HTra** correlates negatively and strongly with **TSS** ($r = -0.67$), indicating that more translation variation is observed when alternative translations are less similar.
  - **HTra** correlates negatively and moderately with **TPP** ($r = -0.40$), indicating that more translation variation is observed when nondominant translations—i.e., less "feature overlap" or lower similarity between ST and TT words—are selected or the translation projection is rather imprecise.

  Nondominant translation: The moderate correlation between **WPA** and **TPP** ($r = 0.4$) indicates that often nondominant translation solutions are produced in our translation data. In addition, the relatively strong correlation between **TPP** and **TSS** ($r = 0.51$) suggests that the similarity between alternative translations tends to increase when more dominant translations are produced.

  Interestingly—as discussed in Fig. 4—**TPP** has a significant *negtive* effect on **lgDur,** which suggests that more isometric mappings are easier to process. However, **WPA** has a significant *positive* effect on **lgDur**, which—in combination with the negative **TPP** effect—suggests that translation production is more time-consuming when producing nondominant translations. Eddington and Tokowicz (2013) did not find a dominance effect in their single word translation recognition task. However, dominance seems to be an important factor for our contextualized translation in which meaning translation-ambiguous words may not play as important a role as in single word translation.
- The DFM (de Groot 1992) predicts that larger semantic overlap between the source and the target facilitates translation production. Several of our findings confirm this hypothesis:
  - **TPP** has a significant negative effect on **lgDur**, indicating that translations with more isometric projections—i.e., better mapping of "semantic" features between ST and TT words—are also easier to translate.

– **TPP** and **TSS** are strongly correlated ($r = 0.51$), which—as discussed above—indicates that translations of more isometric projections have less translational variation.

Episodic memory: De Groot (1992, 1017) suggests that "the conceptual elements in these memory structures" may encode more than just lexical knowledge; it may also contain contextual and episodic knowledge. The "conceptual memory structures" can be seen as a "context availability measures" which represent also traces of episodic memory and may, for instance, impact familiarity ratings.

This view coincides with the way how word2vec vectors are built: word embeddings are trained within an n-gram context. The similarity between vector representations—as can be measured by their cosine similarity—is automatically learned, merely through collocational examples. Similar to episodic memory, these collocational examples are "personal" in the sense that they depend entirely on the provided training samples. The view is also supported by recent enactive-ecological approaces to cognition which argue that "episodic memory functions by strengthening the connection among nodes in a network, not by storing content" (Carvalho and Rolla 2020, 9).

3. To what extent can findings from single word translation carry over to translations in context?

According to Degani and Tokowicz (2010), words with multiple translations in another language trigger a *Fan Effect* in which multiple concepts and words are activated. This explains that retrieval processes are slowed down and translation production takes longer. This effect has been observed for translations in context and out of context. Our results show that there is a weak correlation ($r = 0.23$) between the number of alternative translations in context and out of context. There is also a significant but weak correlation ($r = 0.32$) between average translation similarity scores for alternative translations in context (**TSS**) and out of context (**TSS_P**). These results indicate that findings from translation norms produced out of context may carry over to translations in context with caution.

Single word translations may be meaning translation-ambiguous and/or synonym translation-ambiguous, but a context may eliminate any meaning ambiguities. Eddington and Tokowicz (2013, 453) note:

> When a bilingual translates ambiguous words out of context, they may be more influenced by the level of ambiguity than when processing in a richer semantic context. For example, a bilingual translating words embedded in a discourse contextmay be less influenced by ambiguity

The correlations between the vector space measures are in general weaker in the contextual translations but point to the same direction and seem to be similarly significant as for translations out of context. Heilmann and Llorca-Bofí (this volume), Chap. 8) report a similar finding for words that are translated more literally out of context than within a context.

6 Conclusion

Kroll et al. (2010, 7) wonder "whether the bilingual's two languages draw on semantic representations that are fundamentally shared." The revised hierarchical model (RHM) (Kroll and Stewart 1994) assumes independent lexical stores for each of the languages a speaker knows and a common "conceptual" store for all languages. Word embeddings—as discussed in this chapter—model two independent and continuous vector spaces, one for the source and one for the target language, and a mapping between those spaces that is trained with a small set of frequent translations. This amounts to the rather unlikely situation in which a fluent speaker of two languages learns the translation relations between those two languages at a very late stage, as the monolingual patterns in both languages are already independently established.

It may be worth considering how more realistic scenarios could be implemented and to investigate whether they allow us to model more fine-grained distinctions. Much of the research in bilingualism has investigated and modeled processes of bilingual development and language learning. The RHM (Kroll and Stewart 1994), for instance, is a model of word production which accounts for "observed asymmetries in translation performance by late bilinguals who acquired the second language (L2) after early childhood and for whom the first language (L1) remains the dominant language" (Kroll et al. 2010, 373) The RHM predicts that speakers for whom the L2 is relatively weak will exploit the L1 lexical translation equivalent for the purpose of accessing meaning linked to their more fluent L1. But as L2 proficiency increases, this effect vanishes, and speakers will access L2 meaning directly. More proficient bilinguals do not use the translation equivalent as a mediator to retrieve the meaning of the L2 word. Other branches of bilingualism research investigate the recovery from lesion, brain injury, or aphasia taking into account various parameters, such as language history with regard to the age of L1 and L2 acquisition, proficiency, the dominance of the treatment language, etc. The possibility of amazingly powerful word embeddings to model, confirm, or refine possible mechanisms that underlie such phenomena has yet to be investigated.

We believe that analyzing translation processes with the support of word embedding models has a great potential to establish and test new hypotheses to tightly integrate aspects of bilingualism and TPR and to further the research. More recent models—such as multilingual BERT (e.g., Devlin et al. 2018; Pires et al. 2019)—compile multiple languages into one single language model which allows to assess and compare predictions under the assumption that syntactic-semantic representations are fundamentally shared across languages. It might be worthwhile to investigate whether and how these models might be helpful to shed more light on possible relations of the bilingual mind.

References

Belinkov Y, Glass J (2019) Analysis methods in neural language processing: a survey. Transactions of the association for. Comput Linguist 7:49–72

Bracken J, Degani T, Eddington C, Tokowicz N (2017) Translation semantic variability: how semantic relatedness affects learning of translation-ambiguous words. Bilingualism: language and. Cognition 20(4):783–794. https://doi.org/10.1017/S1366728916000274

Brill, Green (2013) Bilingual stroop in English speakers with Russian as a second language: exploring the model of the bilingual mind. FIVE 2(2):Spring 2013

Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods 46:904–911

Buchweitz A, Prat C (2013) The bilingual BraIn flexibility and control in the human cortex. Phys Life Rev 10:428–443. https://doi.org/10.1016/j.plrev.2013.07.020

Campbell S (2000) Choice network analysis in translation research. In: Olohan M (ed) Intercultural faultlines: research models in translations studies. St. Jerome, Manchester, pp 29–42

Carl M (this volume) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Schaeffer M (2017a) Why translation is difficult : a corpus-based study of non-literality in post-editing and from-scratch translation. Hermes 56:43–57

Carl M, Schaeffer M (2017b) Sketch of a noisy channel model for the translation process. Empirical modelling of translation and interpreting. In: Hansen-Schirra S, Czulo O, Hofmann S (eds) . Language Science Press, Berlin, pp 71–116. (Translation and Multilingual Natural Language Processing; No. 7)

Carvalho EM, Rolla G (2020) An enactive-ecological approach to information and uncertainty. Front Psychol. https://doi.org/10.3389/fpsyg.2020.00588

Conneau A, Lample G, Ranzato M, Denoyer L, Jegou H (2018) Word translation without parallel data. In: International conference on learning representations (ICLR)

De Groot AMB (1992) Determinants of word translation. J Exp Psychol Learn Memory Cogn 18(5):1001–1018. https://doi.org/10.1037/0278-7393.18.5.1001

Degani T, Tokowicz N (2010) Ambiguous words are harder to learn. Biling Lang Congn 13:299–314

Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Dhar P, Bisazza A (2020) Understanding cross-lingual syntactic transfer in multilingual recurrent neural networks. Comput Lang. https://arxiv.org/abs/2003.14056. Accessed 20 Aug 2020

Dijkstra T, van Heuven W (1998) The BIA-model and bilingual word recognition. In: Grainger J, Jacobs A (eds) Localist connectionist approaches to human cognition. Lawrence Erlbaum, Mahwah, pp 189–225

Dijkstra T, van Heuven W (2002) The architecture of the bilingual word recognition system: from identification to decision. Biling Lang Congn 5:175–197

Dijkstra T, Wahl A, Buytenhuijs F, Van Halem N, Al-Jibouri Z, De Korte M, Rekké S (2018) Multilink: a computational model for bilingual word recognition and word translation bilingualism: language and cognition: (1 of 23). Cambridge University Press, Cambridge. https://doi.org/10.1017/S1366728918000287

Do Carmo F (this volume) Editing actions: a missing link between translation process research and machine translation research. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Eddington CM, Tokowicz N (2013) Examining English–German translation ambiguity using primed translation recognition. Biling Lang Cogn 16(2):442–457

Hartsuiker RJ, Pickering MJ, Veltkamp E (2004) Is syntax separate or shared between languages? Psychol Sci 15(6):409–414. https://doi.org/10.1111/j.0956-7976.2004.00693.x

Heilmann A, Llorca-Bofí C (this volume) Analysing the effects of lexical cognates on translation properties: a multi-variate product and process based approach. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Johnson M, Schuster M, Le Q, Krikun M, Wu Y, Chen Z, Thorat N, Vigas F, Wattenberg M, Corrado G, Hughes M, Dean J (2017) Googles multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguist 5:339–351

Kroll JF, Stewart E (1994) Category interference in translation and picture naming: evidence for asymmetric connection between bilingual memory representations. J Memory Lang 33(2):149–174. https://doi.org/10.1006/jmla.1994.1008

Kroll F, van Hell J, Tokowicz N, Green D (2010) The revised hierarchical model: a critical review and assessment. Biling (Camb Engl) 13(3):373–381. https://doi.org/10.1017/S136672891000009X

Lacruz I, Ogawa H, Yoshida R, Yamada M, Martinez DR (this volume) Using a product metric to identify differential cognitive effort in translation from Japanese to English and Spanish. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Leewnberg A, Vela M, Dehdari J, van Genabith J (2016) A minimally supervised approach for synonym extraction with word Embeddings. Prague Bull Math Linguist 105:111–142

Mikolov T, Le Q, Sutskever I (2013) Exploiting similarities among languages for machine translation. arXiv arXiv:1309.4168

Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and source-text background information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual BERT? https://arxiv.org/abs/1906.01502

Prior A, MacWhinney B, Kroll JF (2007) Translation norms for English and Spanish: the role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. Behav Res Methods 39:1029–1038. https://doi.org/10.3758/BF03193001

Prior A, Wintner S, MacWhinney B, Lavie A (2010) Translation ambiguity in and out of context. Appl Psycholinguist 32(01):93–111. https://doi.org/10.1017/S0142716410000305

Prior A, Kroll JF, MacWhinney B (2013) Translation ambiguity but not word class predicts translation performance. Biling Lang Cogn 16(Spec Issue 02):458–474. https://doi.org/10.1017/S1366728912000272

Sahoo D, Carl M (2019) Lexical representation and retrieval on monolingual interpretative text production. In: Dublin machine translation summit XVII: second MEMENTO workshop on modelling parameters of cognitive effort in translation production, 20 august, 2019. MT Summit XVII, Dublin, pp 14–16

Schaeffer M, Dragsted B, Hvelplund K, Balling L, Carl M (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer. ISBN 978-3-319-20357-7, Berlin, pp 183–210

Smith S, Turban D, Hamblin S, Hammerla N (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: International conference on learning representations

Tokowicz N, Kroll J (2007) Number of meanings and concreteness: consequences of ambiguity within and across languages. Lang Cognitive Process 22. https://doi.org/10.1080/01690960601057068

Tokowicz N, Kroll JF, de Groot AMB, van Hell JG (2002) Number-of-translation norms for Dutch-English translation pairs: a new tool for examining language production. Behav Res Meth Instrum Comput 34:435–451

Xing C, Wang D, Liu C, Lin Y (2015) Normalized word embedding and orthogonal transform for bilingual word translation. Proceedings of NAACL

A Radical Embodied Perspective on the Translation Process

Michael Carl

Abstract This chapter develops a post-cognitivist perspective of the translation process based on the dynamic systems approach of Chemero's radical embodied theory of cognition. We introduce the notion of translation affordances, extend it with a probabilistic layer, and show how probabilistic translation affordances are optimized with respect to environmental (e.g., textual) features and the subjects' (e.g., translators') abilities. The model is compatible with research in bilingualism studies, as well as the monitor model and with Schaeffer and Carl's recursive model of shared representations. Probabilistic translation affordances explain translation abilities as effects of horizontal (priming) processes and the optimization of textual material during visual search as vertical monitoring processes. The proposed dynamic system account views translation affordances as basic units of translational cognition and sheds new light on the conception of translation units as cycles of perception-action in translation production.

Keywords Translation affordances · Translation units · Horizontal and vertical processing · Translation priming · Monitoring · Mental representation

1 Introduction

In this chapter, we develop a radical embodied perspective on the translation process that is compatible with a dynamic systems view on cognition. Dynamic systems have been proposed as one radical alternative to the *computational theory of mind* (CTM) by stating that perception and action can be direct without the need for internal representations. Chemero's (2000, 2010) radical embodied perspective on cognition relies on the notion of *affordance* (Gibson 1979), which conceptualizes perception as possibilities for action. It posits that information exists as a product

M. Carl (✉)
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_15

of the relation between environmental features and the abilities of a perceiver (or actor) opening up the possibility for direct action without the need for any kind of internal representation or "higher-level" cognitive processing.

We apply Chemero's radical embodied theory to the translation context. We posit that priming processes enable direct translational action, while textual search amounts to maximizing the readiness of environmental features. Both processes (priming and search) interact with each other; they both optimize translation affordances and, thus, maximize translational action. To account for these processes, we extend the notion of affordances with a probabilistic dimension. This probabilistic extension becomes instrumental when modeling the optimization of translation affordances through subliminal priming processes and through deliberate gazing patterns which both facilitate translational action. To illustrate these processes, assume a translator is about to translate a text. One likely scenario is that the translator will direct her attention to the first word(s) in the source text and start typing as soon as translation solution comes to her mind. Two processes contribute to the optimization of the translation process: the maximization of the translation ability through subliminal priming processes and the deliberate arrangement and scanning of textual material in the environment that is to be translated.

We thus observe the construction of a translation unit (as defined by, e.g., Alves and Vale (2009); see also Carl this volume-b, Chap. 9) which consists of a translation act (i.e., ST reading) and a translation event (TT typing) and that structures the processed textual material according to the translator's translation ability. In this view, a translation unit is a physical instantiation of a translation affordance—the dynamic linkage of the translator's mind, body (i.e., gaze and hands), and the translation environment—that traces the mutual and interactive optimization of environmental *features* and translation *abilities*.[1] Translation units are dynamically updated and integrated as the translation proceeds. Priming mechanisms trigger implicit learning which facilitates the translation of successive passages as an effect of executing the translation task. At the same time, the emerging translation is part of—and thus changes—the translator's environment which is validated and cross-checked as the translation evolves.

The CTM, in contrast suggests that translators build up mental representations of the textual content that allows them to derive meaning hypothesis which are then translated into the target language (for a discussion, see Carl this volume-c, Chap. 13). However, in the radical embodied framework, translation can be described as a "reorganization of the organism-environment system ... [where] cognitive processes are different aspects of the organization and dynamics of the organism-environment system ... there is no need for a part of the system (the organism) to

[1] According to Ramstead et al. (2016, 16), "an ability is simply the capability of an organism to coordinate its action-perception loops to skillfully engage an affordance in a way that is optimal under the free-energy principle." The free-energy principle, in turn, explains how the organism (e.g., a translator) reduces entropy (e.g., in the translation process) restricting herself to a limited number of possibilities. For a discussion on translation entropy, see, e.g., Carl this volume-a, Chap. 5; Wei this volume, Chap. 7; Ogawa et al. this volume, Chap. 11 or Carl et al. (2019).

represent the other part, or to represent parts of itself (the body), or representing the interactions between both parts" (Raab and Araújo 2019, 3), as would have been assumed in CTM.

This chapter lays out a dynamic systems view on the translation process. It introduces an affordance theory within a radical embodied view on the translation process. Section 2 develops basic assumptions of the dynamic system view on radical embodied cognition and locates the dynamic systems approach of radical embodied cognition within *ontological anti-representationalism/methodological non-representationalism* (cf. Carl this volume-c, Chap. 13). In Sect. 3, we extend the notion of affordances with a probabilistic layer. We show how affordances can be optimized by maximizing the subject's ability and the configuration of environmental features, and we use Bayes' rule to split complex affordances into multiple more simple affordances. Section 4 interprets results of bilingualism research and the translation of isolated words in terms of probabilistic affordances that optimize the translator's ability through priming processes. Section 5 extends the notion of translation affordances to translation in context. It discusses instances of default and challenged translation and the recursive configuration of textual features and abilities that determine probabilistic translation affordances (see Carl this volume-b, Chap. 9). Section 6 argues that networks of affordances can model dynamic processes of the translation process without resorting to a notion of representation.

2 Affordances and Anti-representational Cognition

Chemero (2000, 2010) elaborates a radical embodied cognitive theory that is based on the concept of *dynamic systems.* According to Chemero, dynamic systems do not "represent" their environments but rather react to them in a more or less direct manner. Chemero (2000, 2010) makes a distinction between *internal states* and *representations*. He illustrates the difference with the example of the Watt governor. The functioning of the Watt governor, he argues, can be understood as a (non-representational) dynamic system and described in terms of precise mathematical terms without resorting to a language of representations at all. The actual understanding of a dynamical system, he claims, comes from the understanding of the dynamic fluctuations, and it is of little use if a "representational gloss does not predict anything about the system's behavior that could not be predicted by a dynamical explanation alone … If one has the complete dynamical story, what is left to be explained?" (Chemero 2010, 77).

For Hutto and Myin (2013), representations have content which specifies "correctness conditions" so that "anything that deserves to be called content has special properties – e.g. truth, reference, implication – that make it logically distinct from, and not reducible to mere covariance relations" (2013, 67). Within a dynamical system, Chemero (2010, 60ff.) distinguished between different forms of relationships between a (external) stimulus and a corresponding internal target

state (the representation) based on the extent to which the stimulus and the target state are in constant contact with each other or whether and how long they can be decoupled from each other. Different durations of decoupling imply different forms of representational content that may be used by some other part of the system:

1. *Effective tracking* implies that an input stimulus needs to be constantly causally connected with the internal target state. This amounts to a state of "direct perception" which produces necessary information to guide the behavior for the next state of the system as output.
2. *Non-effective tracking* allows for temporary decoupling of the stimulus and the target "for a few milliseconds" (Chemero 2010, 60). It requires "the capacity to use inner states to guide behavior in the absence of the environmental feature represented" (Chemero 2010: 62), for instance, moving an arm when grasping a glass of water with temporary loss of sight. Chemero suspects that such *non-effective tracking* processes might be "ubiquitous in the nervous system ... [during] any degree of behavioral control by expectations from sensory feedback" (62).
3. *Registration* implies a strong decouplable and potentially absence of the stimulus. It "requires abstraction in that the subject must ignore many of the details of the object" (Chemero 2010: 57).

According to Chemero, there is an agreement among scholars that internal states which are triggered by effective tracking can be modeled as associations and that associations are different from representation proper: associations are covariance relations that do not specify correctness conditions. It is controversial, however, whether the existence of a target state with a potentially absent stimulus must count as proper representation and whether strong decouplability is a necessary condition for representation. On the one hand, Saphiro (2019, 192) maintains that a representation must be there despite strong coupling: "Contact without representation is useless." On the other hand, Kiverstein and Rietveld (2018, 11) suggest modelling decoupled cognition as "nested states of action," which eliminates the need for representation all together.

Chemero supports the anti-representationalist view which allows for representation-free action, but also does not eliminate the possibility for representation all together. He suggests that the coupling of stimulus-target are dynamic systems, which can be modeled with coupled oscillators. According to Chemero, dynamic systems do not make use of explicit representations, instead "appropriately connected, intelligently situated activity emerges, apparently without the building or maintenance of representations of the environment" (2000, 626). Chemero (2010) argues that "the best way to understand cognition is with the tools of dynamical systems theory, by ... providing non-representational explanations of cognitive phenomena that are both convincing and sufficiently rich in their implications to guide further research." Chemero suggests three types of non-representational internal states:

(a) *Relaxation oscillators* (i.e., electrical and neural systems) are capable of synchronizing quickly a stimulus with a target. However, they "cannot keep hold of a represented target in the absence of the stimulus" (Chemero 2010, 49) even though a target must not be constantly present. Relaxation oscillators are thus suited to model effective and non-effective tracking to some extent.
(b) *Physical systems* may be used to form a class of coupled oscillators which have momentum and internal dynamic due to their intrinsic mass distributions. They synchronize less quickly with a stimulus than relaxation oscillators, but their mass can keep a rhythmic pattern even in the absence of the stimulus. They are suited for modeling, for instance, motor control tasks.
(c) *Hybrid oscillators* join desirable properties from relaxation and from physical oscillators which allow them to synchronize quickly with the input signal and to "keep the beat even in the absence of the input signal" (Chemero 2010, 50). "Complex adaptive oscillators are required to have representations that are strongly decoupable, to be able to represent absent features of the environment" (Chemero 2010: 58).

In order for direct perception to be possible without being mediated through an explicit internal representation (i.e., registration), the concept of affordances is introduced. Affordances are relations between features of the environment (*feature*) and abilities of an agent (*ability*); they are "opportunities for behavior, which ... are the main things that animals perceive" (Baggs and Chemero 2018, 3). According to this view, the "world is inherently meaningless, but the environment is not; the environment contains affordances" (Chemero 2000, 4). Chemero's notation of affordances is as follows:

$$\text{affords}_X \text{ (feature, ability)}$$

Chemero (2010) gives an example for a "gap-crossing" affordance, the success of which depends on the perceived width of a gap and the stepping ability of an agent. The stepping ability, in turn, depends on several parameters, such as the size of the steps (but surprisingly not length of the leg), the age of a person who is crossing the gap, etc. and could be formalized as follows:

$$\text{affords}_{\text{gap}-\text{crossing}} \text{ (gap-width, stepping-ability)}$$

As this example shows, affordances "arise along with the abilities of animals to perceive and take advantage of them" (Chemero 2010, 146). Affordances allow instances of perception-action to be direct (i.e., non-mediated through an internal representation) where the relation between the perceiver and the affordance can be conceptualized as:

$$\text{perceives} \left[\text{animal, affords}_X \text{ (feature, ability)}\right]$$

However, the animal (e.g., human) is usually only aware of the affordance itself, but not of its constituting *feature* and *ability* parameters, which remain subliminal.[2] In addition, abilities—and the affordances that emerge with them—are probabilistic in nature; "individuals with abilities are supposed to behave in particular ways, and they may fail to do so" (Chemero 2010, 145). Abilities may also fail to become manifest at all, and affordances may not be realized in all cases; they may depend on environmental, cultural, or situational context.[3]

3 Affordances and Probabilities

It has been controversial whether, as for Gibson (1979), any affordance either exists or does not exist. Franchak and Adolph (2014, 2) posit that "affordances are not categorical" if the "performance is variable across repeated trials." This suggest that affordances could be modeled in terms of probabilities. Franchak and Adolph discuss some criteria to define a critical point that would allow to bin performance into two categories (e.g., possible/impossible), but they conclude that "affordances are better considered as continuous, probabilistic functions that represent an individual's likelihood of successful performance." In line with these assumptions, we assign an affordance with a probability P, which determines the "likelihood of successful performance." However, the "success" of the performance may be determined (or measured) in different ways, e.g., by the probability[4] of the affordance to be recognized; its estimated duration; the anticipated energy expended; the probability of associated dangers; financial, economic, and environmental risks; ethical implications; etc.:

$$\text{affords}_X\ (\text{feature, ability}) :: P\ (\text{affords}_X) = P\ (\text{feature, ability})$$

Similar to Ramstead et al. (2016), we develop a Bayesian account of affordances that may be optimized based on the selection of environmental features or the agent's capabilities.

On the one hand, a complex *feature* may have different environmental configurations $\{f_1, f_2, \ldots, f_m\}$ that may play a role in the success of affordance performance. For instance, to optimize the success of a gap-crossing affordance, it might be possible to cross the gap in a diagonal or orthogonal fashion. There might also be different starting points and landing sites on that gap which appear differently secure or slippery. Some landing sites might be more elevated than others, making certain ways of gap-crossing more effortful than others, etc. The possible combinations

[2] The more restricted term "perceived affordance" requires an agent to be aware of the affordance.

[3] See also Hutchinson (2019: 121ff.) discussion on the *p-affordance.*

[4] For simplicity and readability reasons, we omit the index "x" in probability notation, which could also be "P_x".

of those features constitute different environmental configurations, the selection of which may result in different success probabilities for the gap-crossing affordance given the set of abilities of the gap-crosser. We could perhaps say that all these different configurations account for different affordances from which a gap-crosser picks one given his/her abilities. As we will see, it might be easier to model the potentially exponential number of environmental configurations in a probabilistic manner.

On the other hand, an ability may depend on various performance capabilities $\{a_1, a_2, \ldots, a_n\}$ that contribute to the expected affordance success. For instance, choosing the left or the right as the takeoff leg, take a running start, swinging arms or moving the body in a certain way, etc. may all have an effect on the gap-crossing success, where the coordination and performance of those abilities (the gap-crossing event) depend on the particular environmental configuration that was selected for gap-crossing.

Chemero thinks of the interaction between features in the environment and abilities of an agent as coupled oscillators. Environmental configurations are not *represented* in the agent's mind, but they resonate with the agent's capabilities and offer opportunities for action. According to Norman (1988), an affordance is a relationship between the properties of an object and the capabilities of the agent that determine how the object could possibly be used. Environmental features and abilities "causally interact in real time and are causally dependent on each other" (Chemero 2010, 152). This implies that affordances can be optimized in two ways, by maximizing the performance capabilities ($\{a_1, .. \ a_n\}$) or by maximizing the readiness of environmental configurations ($\{f_1, .. f_m\}$).

For instance, a *gap-crosser*$_1$ may prefer a shorter gap with a slippery landing site that resonates better with her capabilities, while for another *gap-crosser*$_2$, the diagonal route that allows for a running start may be more suitable given a different set of capabilities. The two gap-crosser optimize the affordance differently by maximizing the conditional probability of environmental features f_i given their individual abilities:

$$P\,(\text{feature, ability}) = \max\,(f_i \in \text{feature})\ \{\ P\,(f_i \mid \text{ability}) \times P\,(\text{ability})\ \}\,.$$

Given the different abilities of *gap-crosser*$_1$, she may select and thus optimize the environmental configuration in a different manner then *gap-crosser*$_2$. However, an affordance can also be optimized by maximizing the conditional probability of a capacity a_i for a given environmental setting. Thus, for a certain gap topology, it might be best to jump with both feet instead of jumping with one foot or to run-start. The gap-crosser will thus select (i.e., maximize) the set of her capabilities for the given environmental configuration:

$$P\,(\text{feature, ability}) = \max\,(a_i \in \text{ability})\ \{\ P\,(a_i \mid \text{feature}) \times P\,(\text{feature})\ \}$$

According to Gibson and Pick (2002), affordances can be learned through a process of differentiation by splitting an existing, more general affordance into

multiple, more specific affordances and actions. A complex affordance can thus also break down into a sequence of more simple affordances. For instance, a complex gap-crossing affordance might break down into several interconnected affordances, which involve a running-start affordance, a take-off affordance, a landing-site affordance, etc. The success of the compound affordance (afford_x) would then depend on each of the more specific affordances (afford_i), which can be modeled as the product of their associated probabilities:

$$\text{affords}_x\ (\text{feature, ability}) :: \prod_{\text{affords}_i \in \text{affords}_x} P\ (\text{affords}_i)$$

Each of those new affordances has sets of properties for their respective features and abilities that can be maximized according to the opportunities for action they represent.

Given that affordances and ability "interact in real time" and are perceptually re-assessed at each moment, the different factorization may result in different affordance probabilities over time. The perception of an affordance with probability *P* may have an impact on the agent's ability to perceive and react and to maximize probabilities of successive similar affordances. For instance, once a gap-crossing has been performed, the affordance for a next gap-crossing might be differently assessed, the environmental features may be perceived differently and reconfigured, and the abilities may have changed. Gallagher (2017, 18) sees two directions of fit in this process:

> [t]he first involves updating predictions or updating priors on the basis of ongoing perceptual experience – the world-to-brain direction. The second involves acting on the world to directly shape or re-sample it in such a way as to directly test our prior expectations . . . for example, active ballistic saccades do not merely passively orient towards features but actively sample the bits of the world that fit my expectations or resolve uncertainty

According to Baggs and Chemero (2018, 9–10), "the world specifies structure in energy arrays (patterns in sound, in light, etc.), which in turn specifies what an animal perceives." The different factorization models show, on the one hand, how "the animal [can] explore energy arrays such that what it perceives specifies appropriate information," i.e., which environmental features fit the given ability. On the other hand, the optimization of the animal's ability models how the perceived information "specifies structure in the world that is adaptive for the animal's purposes," i.e., enhance the abilities that correspond to the given environmental feature(s). The optimization of affordances through different factorization possibilities captures this mutual relation between features of the "umwelt" and abilities of the acting agent.

The basic notion of affordances stipulates that perception is a direct guide to action, without a need for a mental model that duplicates the sensory information. However, the affordance theory of direct perception does not rule out the possibility of indirect perception as a complementary process, which is mediated by mental models and "non-action constructs." Chemero suggests simulating affordances as coupled oscillators. Oscillators are probabilistic devices that produce, for instance, sine or square wave output with different wavelengths. A coupling of oscillators can

be modeled as a sequence of probabilistic processes (i.e., Markov chains), in which the state of one oscillator depends on the state of the previous oscillator. A recursive interaction of relaxation and hybrid oscillators could thus also be modeled as a probabilistic process. In the next section, we assess how such a view on affordances might be suited to explain the translation process.

4 Affordances and Translation Priming

At the very basis of the translation is the recognition of words, which, according to most theories and models, requires access to a "mental lexicon." Several models describe how the mental lexicon is organized and how words are accessed and retrieved during the translation process (De Groot 1992, Kroll and Stewart 1994, Dijkstra et al. 2018; see also Carl this volume-d, Chap. 14). There is a general agreement that initial word recognition and translation is a subliminal, automatized process, which depends on several parameters. This process can be modeled as a translation affordance:

$$\text{affords}_{\text{translating}}\ (\text{expression, translation\_ability})$$

The probability $P_{\text{dur}}(\text{expression, translation\_ability})$ for the execution duration of the translation affordance (i.e., the perception-action loop) depends on several factors such as word frequency, the kinds of translation ambiguities, word length, experience of the translator, etc. Translation is a partially automatic process, which is slowed down if, for instance a word has more than one translation alternative (see also Carl this volume-d, Chap. 14).

The degree of automatization of this process can be tested through priming studies. Priming is a technique whereby the exposure to one stimulus (the prime) has an impact on the recognition or response time to a subsequent stimulus (the target) without conscious guidance or intention. In terms of the proposed affordance model, priming studies investigate processes of *non-effective tracking*, as the prime and the target stimuli are usually separated by a short lapse of time. According to Hartsuiker and Bernolet (2015), priming is a form of implicit learning by which complex knowledge and skills can be acquired without the awareness of what is being learned, i.e., in the absence of consciously accessible knowledge.

Within a dynamic systems perspective, priming processes can be explained as an activation of areas within a network of possible internal states (e.g., oscillators) that facilitate the recognition and processing of similar successive stimuli. It suggests that this learning is based on mere covariation of a prime and a target and does not rely on representations that involve the evaluation of correctness conditions, truth, reference, or implication.

Priming effects have been shown to exist—among many other phenomena—for phonetic (e.g., cognates), semantic (Dimitropoulou et al. 2011, Schoonbaert et al. 2011), and syntactic structures (Bangalore et al. 2016, Maier, et al. 2017).

These studies show that a related prime results in faster response times for the successive target than an unrelated prime. A number of priming studies showed that priming takes place also in translation (Tokowicz and Kroll 2007, Laxén and Lavaur 2010, Boada et al. 2013, Eddington and Tokowicz 2013, Prior et al. 2013). These studies have shown that a related prime can facilitate translation recognition speed compared to an unrelated prime.

In a translation recognition task, Eddington and Tokowicz (2013) presented unambiguous translations, synonym translation-ambiguous source words, and meaning translation-ambiguous source language words. A synonym translation-ambiguous word in English is, for example, "shy" which can be translated into German in different forms as *schüchtern* or *scheu*. In contrast, a meaning translation-ambiguous word is a homograph with several meanings: "odd" can refer to an odd number or something strange. Depending on which meaning is used, the translations into German are different (*ungerade* or *merkwürdig*) (Eddington and Tokowicz 2013: 442). Bilingual participants were presented with English-German word pairs that were preceded by a related or unrelated prime and were asked to decide if the word pairs were translations. They found that translation ambiguity slows down translation recognition regardless of the source of ambiguity (synonym translation-ambiguous or meaning translation-ambiguous). Participants were slower and less accurate to respond to words that had more than one translation compared to unambiguous words.

To explain these observations, Eddington and Tokowicz develop the *Revised Hierarchical Model of Translation* which links the two languages on two levels: (1) direct lexical links between the two languages which allow for fast processes and (2) conceptually mediated links through a space of distributed meaning representations. The model allows for synonym translation-ambiguous words to have a different direct link to each synonym in the TL, while meaning translation-ambiguous words can have different links to distinct conceptual representations, which are shared between the two languages. Eddington and Tokowicz suggest that:

> Translation unambiguous word pairs … have the strongest associative strength … whereas translation ambiguous words would have weaker associations between a source word and each translation, resulting in longer, more difficult processing … For translation-ambiguous words, more than one alternative translation is available for selection, which may lead to active competition between the possible translations. Selecting one translation over another would require the inhibition of the unselected translation alternatives, leading to slower and less accurate responses. (2013: 453)

The model assumes that words and possible translations are activated due to a *fan effect* (Anderson 1974), which is modeled within the Multilink model (Dijkstra et al. 2018) as follows. When reading a text, orthographic neighbors of the input words are automatically activated in a language nonselective manner. That is, first orthographically (and phonetically) similar words in the source and the target language(s) will be activated. In a successive step, semantic representations are activated: "orthographic representations will then begin to activate their meaning representations ... and semantically active representations 'spread' their activation to other units" Dijkstra et al. (2018: 2). Finally, a task-dependent decision process

selects (an) appropriate candidate(s) from this network of activated words and makes sure that the correct translation is produced in the correct (target) language. Word recognition and translation time are a function of the size of these automatically activated networks, which depends—among other things—on the ambiguity, frequency, and length of the activated words.

In this view, Multilink models the internal dynamics of translation affordances in which the activation and coupling of interconnected networks lead to the recognition or translation of a single word that is being presented. Eddington and Tokowicz suggest that what the task-dependent decision process is in Multilink might function as an "inhibition of the unselected translation alternatives." Priming, according to Eddington and Tokowicz, may either narrow the number of activated word associations or the selection of the task-dependent decision process. In both cases, priming speeds up the translation process by strengthening direct pathways which eliminate the need for the evaluation of correctness conditions or inferential reasoning over meaning representations.

Priming effects can thus be understood as a subliminal optimization of affordances to maximize future recognition and response abilities of a related target stimulus. Exposure to a prime (e.g., a source text word) conditions a subject to changing or adapting the translation ability which leads to more efficient (e.g., quicker) access and processing of a successive similar stimulus. In a dynamic systems account of the mind, the priming effect can be explained as a local activation of nodes in the neighborhood network of the prime (cf. Carl this volume-d, Chap. 14) so that successive items with similar properties can be more easily activated, accessed, and produced, without a need for intermediate representation.

5 Affordances in the Translation Context

While these priming studies report translation production of single words, coherent translation production has also been described as a process of (*effective*) tracking without an apparent need to produce internal representations. Carl and Dragsted (2012) show that stretches of fluent translation production are similar to text copying into another language in which a target text emerges at a maximum possible typing speed of the translator: while the eyes take in new information of the source text, the fingers type out a target text, apparently restricted only by the speed of the finger movements and motor control. Carl (2013, 125) notes that "[i]n an unchallenged translation situation source text fixations trigger target text production, with only little look-ahead and a linear word-for-word translation production." Carl et al. (2011) find such typical translation patterns in professional translators which they label *head starter*. Head starters start translating right away, looking only few words into the ST context. With an eye-key span of three or so words, it is impossible for the translator to build a meaning representation of the sentence, or even the phrase that is being translated. Rather those experienced translators seem to be confident typing out translations word-by-word whenever the context allows for.

Carl (2013) compares two simulations of this unchallenged translation process, a rule-based ACT-R implementation (Anderson 2007) and a statistical model. The statistical model is based on two probabilistic processes, a probabilistic reading process of a source word *s*: $P(R_s)$ and a probabilistic process of writing the translation *t* given a word *s* was read: $P(W_t \mid R_s)$. The model takes into account the average typing speed of frequent character combinations since more frequent combinations such as "er" or "ations" are typed quicker than less frequent character sequences. It also factors in average gaze durations on words, as shorter and more frequent words receive less and shorter fixations than longer or less frequent words. This leads to a chained probabilistic process model $P(W_t \mid R_s) \times P(R_s)$, which represents a sequence of coupled oscillators. Carl (2013) shows that the statistical simulation captures better fine-grained fluctuations in reading and writing patterns than the rule-based ACT-R implementation. This corroborates Chemero's assumption, who argues that dynamical systems are better suited for modelling details of cognitive processes, while rule-based systems may also reproduce global means of the observations but miss finer-grained variations.

Behavioral patterns change dramatically as the translation process becomes more entangled, less compositional, and less monotone. Models of single-word translation are not designed and not suited to take into account recursive processes and contextual translation integration, monitoring of emerging target texts, or revision behavior. Much effort is spent in single-word translation studies to eliminate contextual interference by showing distractors and lists of unrelated words. Eddington and Tokowicz (2013, 453) mention that "a bilingual translating words embedded in a discourse context may be less influenced by ambiguity," as the context will disambiguate meaning translation-ambiguous source language words. The resulting models of bilingualism are thus designed to give a de-contextualized snapshot of the mind when dealing with one controlled input. They do not account for complexities encountered in the translation of texts where recursive gazing patterns on the source text (ST) and the target text (TT) and revision processes take place. Simple priming models measure the maximization of the ability given the environmental feature (i.e., the prime), but they do not account for the maximization of environmental feature configurations, given the translator's ability, nor do they explain how translation affordances and abilities recursively interact in real time.

An instance of how environmental (i.e., textual) features may be arranged and configured in the translation process is discussed in Wei (this volume, Chap. 7). Wei investigates gaze patterns triggered during the translation of a highly ambiguous metaphor. He explains various examples of how translators pick up, with high precision, those textual clues that help understand and disambiguate the metaphorical expression under scrutiny, which then enables them to pursue with the translation production. When drawing attention to specific words, the translator mentally rearranges environmental features into *translation units*, searching for a collection of textual items for which a translational equivalent can be established. According to Alves and Vale (2009), a translation unit consists of an *act* of (ST) reading and the *event* of (TT) production that relate to each other (cf. Carl this volume-b, Chap. 9). Under this view, a translation unit is an instantiation of a

translation affordance,[5] which is jointly determined by the properties of the object (i.e., the text) and the ability of the agent (Norman 1988). It can thus be argued that the notion of *translation unit* is a fluent construct; it is a "reorganization of the organism-environment system" (Raab and Araújo 2019, 3) which constantly changes not only across different texts and translators but also from moment to moment. Any (nontrivial) text allows for an exponentially large number of different segmentations and (mental) configurations of its textual elements each of which potentially activate numerous translations in the translator's mind and thus allows for a large number of different translation units, some of which are more entrenched than others.

Within the affordance model, a translator enters into a recursive action-perception loop, thereby optimizing translation affordances in which specific translation abilities and the disposition of textual features are mutually maximized. In the search for a translation solution, the translator collects textual features that respond to the current translation problem. This targeted contextual search triggers priming mechanisms, which activate translation options that complement and integrate with already activated translation alternatives.[6] In Wei's analysis, when searching the metaphor context, translational entropy decreases in the translator's mind, depending on properties of the words attended to. As sufficient disambiguating information is gathered and the translation entropy lowered below a certain limit, it is possible for the translator to formulate a (first draft) translation, and a translation unit is completed. The translation solution becomes, in turn, a textual element in the translation environment which is subject to visual search and scrutiny. The translator, thus, enters into a recursive loop in which the produced translations themselves become environmental features that are susceptible to visual search, secondary priming processes, and extended translation units.

Schaeffer and Carl (2013) propose a recurse model of the translation process, which includes recursive horizontal and vertical translation processes that operate on the ST and TT. Schaeffer and Carl assume that horizontal processes are based on priming mechanisms that automatically activate translation options and shared representations early during source text reading and which serve as a basis for generating the target language, as discussed in Sect. 4. Vertical processes act as a monitor for target text production which "control the acceptability of the target text" and "assess whether source and target texts are compatible in terms of propositional content and shared conceptual representations" (Schaeffer and Carl 2013, 38). As vertical processes "depend on context which becomes available later, as processing

[5]It is controversial as to whether affordances are permanent or change over time. While for Gibson affordances do not change relative to an agent's internal states, Chemero's *affordance 2.0* is a dynamic relation between the abilities of the individual and features of the environment. Ramstead et al. (2016) introduce "landscapes" and "fields" of competing affordances which "changes through cycles of perception and action."

[6]Carl et al. (2019) model this integration as a hierarchy of interacting word and phrase translations systems which organize and integrate as dissipative structures. This view is compatible with the "free-energy principle" as suggested in Ramstead et al. (2016).

advances in the chunk or text" (Schaeffer and Carl 2013: 37), they are naturally not addressed in bilingual models of single-word translation. However, vertical processes are crucial in human translation production and arguably the main source for cognitive effort (see Lacruz et al. this volume, Chap. 11).

While horizontal processes are based on priming mechanisms which reduce or eliminate the need for contentful representations, vertical processes may involve representations that allow to control for and that specify conditions to assess the correctness of the produced translations. Within the suggested theory of translation affordances, horizontal processes thus facilitate the translation production trough implicit priming mechanisms. Vertical translation processes trigger gaze patterns to take in or contextualize new information or to check initial draft translations. Vertical processes thus optimize configurations of textual features that enable evaluation and monitoring processes. Both processes are statistical, and their mutual optimization is essential for successful translational action.[7] Schaeffer and Carl (2013) assume a:

> recursive cycle which integrates horizontal and vertical source and target language processes: the monitor needs to compare whether the source is the same as the target, but it is equally important to make sure that the target is the same as the source.

It implies that priming processes take place not only from source to target but also from target to source as well as within each of the source and target languages which may provide explanations for shining-through effects and also for normalization and other universals in translation (Halverson 2003, Hansen-Schirra et al. 2017).

Within TPR, translation effort and thus the interaction between horizontal and vertical translation processes have mainly been studied on a rather abstract level through analyzing accumulated typing duration and gazing patterns. A point in case is pause analysis. Pause analysis has been one of the main topics of TPR (Schilperoord 2001, O'Brien 2006, Alves and Vale 2009, Timarová et al. 2011, Carl and Dragsted 2012, Lacruz et al. 2012, Kumpulainen 2015). The analysis of the rhythm and temporal structure in translation production is taken as an indicator for cross-lingual priming effects and cognitive effort. Lacruz et al. (2012) suggest that inter keystroke delays of up to 300 ms could be due to problems of motor control. She, therefore, suggests that pauses relevant for assessing cognitive effort (i.e., vertical processes) should be longer than 300 ms. However, rather than investigating the dynamic interaction of horizontal and vertical processes, the preferred method has been to average over segments of texts, typically a sentence. Pause analysis examines the lag of time between successive keystrokes during text production and

[7]Instead of optimizing environmental features within affordances, one could also say that each affordance comes with (a set of) fixed but different environmental features, and the optimization consists in the selection of the most appropriate affordance for a given ability. While both models could explain the same behavior, the former model may better account for large number of environmental configurations, while the latter version may tremendously inflate the "field" of affordances if environmental features and abilities are continuous. It would then be difficult to understand how affordances can exist independently of specific individual organisms.

can be combined with gazing data (or other behavioral measures). It is assumed that stretches of relatively regular, fluent, and uninterrupted typing activity are indicators of easy translation that mostly rely on horizontal priming processes (i.e., default translations), while longer pauses indicate translation problems and higher cognitive effort, more strongly involving vertical processes (see also Carl this volume-b, Chap. 9).

6 Conclusion

Increasingly, TPR authors make use of bilingualism models (including several authors in this volume) to explain translation processes, but only rarely address the dynamic interaction between the horizontal and vertical processes. If a word or passage is ambiguous or unclear (cf. Wei this volume, Chap. 7), if a cultural item has no immediate translation (cf. Lacruz et al. this volume, Chap. 11), or if the meaning of a neologism can only be inferred through the context (Chen this volume, Chap. 12), we may observe an extended typing pause in which the translator searches for disambiguating clues in the source or the target or external resources. As Kussmaul (1995) illustrates, longer pauses (e.g., >5 s) are often characterized by displacement activities. While the mind waits for alternative translation options to pop up, the gaze may wander around and shift attention from the text into the room or out of the window. This *illumination phase*, "which is often achieved by some parallel activity . . . such as going to the kitchen, eating a bar of chocolate" (Kussmaul 1995, 188), ends with a "Eureka" moment, a short but important moment in the sequence of translation steps in which a (preliminary) solution is found.

As the translation becomes more entangled, it requires more elaborate inferences, thought, or memory processes to take place. Without proper representations, it has been claimed, it is difficult to explain how such "higher" cognition can be achieved. The underlying mental processes that are decoupled from the environment have been traditionally conceptualized and explained as "representations." In order to model such instances of "representation hungry" cognition, Kiverstein and Rietveld (2018, 11) suggest to model thought and memory as a "complex form of coordinating nested states of action." Instead of "representing" past events, a subject "re-enacts" his past experiences in which "the subject pretends to perform the same activities they would perform were they coupled to affordances" (Kiverstein and Rietveld 2018: 12). According to Kiverstein and Rietveld (2018), this reenactment is based on "systematic patterns of covariation that hold between the model and whatever it is modeling in the world." Kussmaul's observation of parallel activity could then be explained in terms of *hybrid oscillators and as an effect of* covariance, which does not require a (truthful) representation of an outside world at all.

References

Alves F, Vale D (2009) Probing the unit of translation in time: aspects of the design and development of a web application for storing, annotating, and querying translation process data. In: Göpferich S, Jääskeläinen R (eds) Process research into translation competence. Spec Issue Across Lang Cult 10(2):251–273

Anderson J (1974) Retrieval of propositional information from long-term memory. Cogn Psychol 6(4):451–474

Anderson J (2007) How can the human mind occur in the physical universe? Oxford series on cognitive models and architectures. University Press, Oxford

Baggs E, Chemero A (2018) Radical embodiment in two directions. Synthese. https://doi.org/10.1007/s11229-018-02020-9

Bangalore S, Behrens B, Carl M, Ghankot M, Heilmann A, Nitzke J, Schaeffer M, Sturm A (2016) Syntactic variance and priming effects in translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research. Springer, Berlin, pp 211–238. ISBN 978-3-319-20357-7

Boada R, Sánchez-Casas R, Gavilán JM, Garciá-Aleba JE, Tokowicz N (2013) Effect of multiple translations and cognate status on translation recognition performance of balanced bilinguals. Biling Lang Cogn 16(1):183–197

Carl M (2013) A computational cognitive model of human translation processes. In: Bandyopadhyay S, Naskar SK, Ekbar A (eds) Emerging applications of natural language processing: concepts and new research. IGI Publishing, Hershey, PA, pp 110–128

Carl M (this volume-a) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M (this volume-b) Micro units and the first translational response universal. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M (this volume-c) Computation and representation in cognitive translation studies. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M (this volume-d) Translation norms, translation behavior, and continuous vector space models. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Dragsted B (2012) Inside the monitor model: processes of default and challenged translation production. TC3 2:127–145

Carl M, Dragsted B, Lykke Jakobsen A (2011) A taxonomy of human translation styles. http://translationdirectory.com/articles/article2321.php

Carl M, Tonge A, Lacruz I (2019) A systems theory perspective on the translation process. Transl Cogn Behav 2(2):211–232

Chemero A (2000) Anti-representationalism and the dynamical stance. Philos Sci. https://doi.org/10.1086/392858

Chemero A (2010) Radical embodied cognitive science. MIT Press, Cambridge, MA

Chen J (this volume) Translating Chinese neologisms without knowledge of context: an exploratory analysis of an eye-tracking and key-logging experiment. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

De Groot AMB (1992) Determinants of word translation. J Exp Psychol Learn Mem Cogn 18(5):1001–1018

Dijkstra T, Wahl A, Buytenhuijs F, Van Halem N, Al-Jibouri Z, De Korte M, Rekké S (2018) Multilink: a computational model for bilingual word recognition and word translation. Biling Lang Cogn 22(4):657–679

Dimitropoulou M, Duñabeitia JA, Carreiras M (2011) Two words, one meaning: evidence of automatic co-activation of translation equivalents. Front Psychol. https://doi.org/10.3389/fpsyg.2011.00188

Eddington CM, Tokowicz N (2013) Examining English–German translation ambiguity using primed translation recognition. Biling Lang Cogn 16(2):442–457

Franchak J, Adolph K (2014) Affordances as probabilistic functions: implications for development, perception, and decisions for action. Ecol Psychol 26(1–2):109–124. https://doi.org/10.1080/10407413.2014.874923
Gallagher S (2017) Enactivist interventions rethinking the mind, Oxford
Gibson JJ (1979) The ecological approach to visual perception. The Psychology Press, New York, NY
Gibson E, Pick AD (2002) An ecological approach to perceptual learning and development. Oxford University Press, Oxford
Halverson S (2003) The cognitive basis of translation universals. Targets 15(2):197–241
Hansen-Schirra S, Nitzke J, Oster K (2017) Predicting cognate translation. In: Empirical modelling of translation and interpreting. Language Science Press, Berlin, pp 3–22. https://zenodo.org/record/1090944#.XysAGChKg2w
Hartsuiker RJ, Bernolet S (2015) The development of shared syntax in second language learning. Biling Lang Cogn 1:1–16. https://doi.org/10.1017/S1366728915000164
Hutchinson P (2019) The missing 'E': radical embodied cognitive science, ecological psychology and the place of ethics in our responsiveness to the lifeworld. In: Backström J, Nykänen H, Toivakainen N, Wallgren T (eds) Moral foundations of philosophy of mind. Springer, Cham
Hutto, Myin (2013) Radicalizing enactivism. The MIT Press, Cambridge, MA
Kiverstein JD, Rietveld E (2018) Reconceiving representation-hungry cognition: an ecological-enactive proposal. Adapt Behav 26(4):147–163
Kroll JF, Stewart EJ (1994) Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. J Mem Lang 33:149–174
Kumpulainen M (2015) On the operationalisation of 'pauses' in translation process research. Transl Interpreting 7(1):47–58
Kussmaul P (1995) Training the translator. Benjamins, Amsterdam
Lacruz I, Shreve GM, Angelone E (2012) Average pause ratio as an indicator of cognitive effort in post-editing: a case study. In: O'Brien S, Simard M, Specia L (eds) Proceedings from AMTA, 10th Conference of the association for machine translation of the Americas, workshop on post-editing technology and practice (WPTP 2012). AMTA, San Diego, CA
Lacruz I, Ogawa H, Yoshida R, Yamada M, Martinez DR (this volume) Using a product metric to identify differential cognitive effort in translation from Japanese to English and Spanish. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham
Laxén J, Lavaur J-M (2010) The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. Biling Lang Cogn 13(02):157–183
Maier RM, Pickering MJ, Hartsuiker RJ (2017) Does translation involve structural priming? Q J Exp Psychol. https://doi.org/10.1080/17470218.2016.1194439
Norman D (1988) The psychology of everyday things. Basic Books, New York, NY. http://web.stanford.edu/~rldavis/educ236/readings/doet/text/ch01.html
O'Brien S (2006) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Lang Cult 7(1):1–21
Prior A, Kroll JF, Macwhinney B (2013) Translation ambiguity but not word class predicts translation performance. Biling Lang Cogn 16(02):458–474
Raab M, Araújo D (2019) Embodied cognition with and without mental representations: the case of embodied choices in sports. Front Psychol. https://doi.org/10.3389/fpsyg.2019.01825
Ramstead MJ, Veissière SP, Kirmayer LJ (2016) Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. Front Psychol 7:1090. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960915/
Saphiro L (2019) Embodied cognition. Routledge, London
Schaeffer M, Carl M (2013) Shared representations and the translation process: a recursive model. Transl Interpreting Stud 8(2):169–190. reprint in Describing Cognitive Processes in Translation: Acts and events, Edited by Maureen Ehrensberger-Dow, Birgitta Englund Dimitrova, Séverine Hubscher-Davidson and Ulf Norberg. [Benjamins Current Topics, 77]

Schilperoord J (2001) On the cognitive status of pauses in discourse production. In: Rijlaarsdam G, Olive T, Levy M (eds) Studies in writing, vol. 10: contemporary tools and techniques for studying writing. Kluwer Academic Publishers, Dordrecht, pp 61–87

Schoonbaert S, Holcomb PJ, Grainger J, Hartsuiker RJ (2011) Testing asymmetries in noncognate translation priming: evidence from RTs and ERPs. Psychophysiology 48(1):74–81. https://doi.org/10.1111/j.1469-8986.2010.01048.x

Timarová S, Dragsted B, Hansen IG (2011) Time lag in translation and interpreting. In: Alvstad C, Hild A, Tiselius E (eds) Methods and strategies of process research: integrative approaches in translation studies. John Benjamins, Amsterdam, pp 121–146

Tokowicz N, Kroll JF (2007) Number of meanings and concreteness: consequences of ambiguity within and across languages. Lang Cogn Process 22(5):727–779

Wei Y (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Index

M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8