

Empirical studies in translation and discourse

Edited by

Mario Bisiada

Translation and Multilingual Natural
Language Processing ??



Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Johannes Gutenberg-Universität Mainz)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.
2. Hansen-Schirra, Silvia & Sambor Grucza (eds.). Eyetracking and Applied Linguistics.
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3 I.
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II.
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). Language technologies for a multilingual Europe: TC3 III.
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). New perspectives on cohesion and coherence: Implications for translation.
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds.). Empirical modelling of translation and interpreting.
8. Svoboda, Tomáš, Łucja Biel & Krzysztof Łoboda (eds.). Quality aspects in institutional translation.
9. Fox, Wendy. Can integrated titles improve the viewing experience? Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data.
10. Moran, Steven & Michael Cysouw. The Unicode cookbook for linguists: Managing writing systems using orthography profiles.
11. Fantinuoli, Claudio (ed.). Interpreting and technology.
12. Nitzke, Jean. Problem solving activities in post-editing and translation from scratch: A multi-method study.
13. Vandevoorde, Lore. Semantic differences in translation.

ISSN: 2364-8899

Empirical studies in translation and discourse

Edited by

Mario Bisiada


Bisiada, Mario (ed.). 2021. *Empirical studies in translation and discourse* (Translation and Multilingual Natural Language Processing ??). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/000>

© 2021, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

ISBN: no digital ISBN

no print ISBNs!

ISSN: 2364-8899

no DOI

ID not assigned!

Cover and concept of design: Ulrike Harbort

Fonts: Libertinus, Arimo, DejaVu Sans Mono

Typesetting software: X_YLaTeX

Language Science Press

xHain

Grünberger Str. 16

10243 Berlin, Germany

langsci-press.org

Storage and cataloguing done by FU Berlin

Contents

Preface	iii
1 Post-editing: a genealogical perspective on translation practice Maeve Olohan	1
2 Testing the Gravitational Pull Hypothesis on modal verbs expressing obligation and necessity in Catalan through the COVALT corpus Josep Marco	21
3 Stylistic normalisation, convergence and cross-linguistic interference in translation: The case of the Czech transgressive Olga Nádvorníková	45
4 The internationalized text and its localized variations: A parallel analysis of blurbs localized from English into Arabic and French Madiha Kassawat	79
5 Movement or debate? How #MeToo is framed differently in English, Spanish and German Twitter discourse Mario Bisiada	95
6 Investigating patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German Felix Hoberg	119
7 What can Euclidean distance do for translation evaluations? Éric André Poirier	141
8 Between audiovisual translation and localization: The case of <i>Detroit: Become Human</i> Laura Mejías-Climent	175
9 Analysing the Dimension of Mode in Translation Ekaterina Lapshinova-Koltunski	195

Contents

Index	213
--------------	------------

Preface

Empirical translation studies has moved beyond being a mere methodological approach within translation studies to becoming an established sub-field of translation studies. This is not only shown by the amount of volumes dedicated to it (Hansen 2002; Carl et al. 2015; Ji 2016; de Sutter et al. 2017; Ji & Oakes 2019). The idea of empirical translation studies has a strong tradition, and perhaps to some extent its origin, in corpus-based translation studies (de Sutter et al. 2012; Ji et al. 2016), but empirical approaches have also been adopted in other paradigms of translation studies (Laviosa et al. 2016).

In their programmatic article *Towards methodologically more rigorous corpus-based translation studies*, de Sutter et al. (2012) offer a set of minimal requirements for research papers in the field of corpus-based translation studies. According to this, papers should:

1. “provide a meticulous overview of the corpus materials used and of the exact procedures for selecting, annotating and sifting the data”
2. “comment on any specific problems encountered during data selection and annotation, including explicit and motivated statements as to the solutions being adopted”
3. “include elaborate testing for statistical significance as a complement of, not in opposition to, thorough qualitative analysis”

With the slight modification that the data does not necessarily have to be corpus data, I think we can adopt at least the first two requirements for a wider view of empirical translation studies, as those are steps that allow other researchers to reproduce the study in question. As for the issue of statistical significance, it is true that the uptake has been slow in our discipline, though studies generally tend to include significance testing. At the same time, the notion of “statistical significance” is continuously questioned in science in general (see, e.g., McShane et al. 2019) and in corpus linguistics in particular (Koplenig 2019).

A more recent suggestion for an improved research agenda in empirical translation studies comes from de Sutter & Lefer (2019), who demand passing on from

Preface

the analysis of well-studied phenomena towards exploring new linguistic indicators, such as “linguistic features that have been said to typify other forms of constrained communication, such as non-native language varieties, editing and student writing” (de Sutter & Lefer 2019: 19). Starting from the basic assumption that “translation products and processes are multifaceted and multidimensional” (de Sutter & Lefer 2019: 18), so that their study should draw on multi-methodological designs. They argue that “understanding translation inevitably entails an interdisciplinary approach to translation, building on theoretical frameworks and findings from neighbouring disciplines, including, but not restricted to, variational corpus linguistics, bilingualism studies and (cognitive) sociolinguistics” (de Sutter & Lefer 2019: 18).

It is in this spirit, then, that the present volume seeks to contribute some studies to the subfield of Empirical Translation Studies and thus aid in extending its reach within the field of translation studies and thus in making our discipline more rigorous and fostering a reproducible research culture. The Translation in Transition conference series, across its editions in Copenhagen (2013), Garmersheim (2015) and Ghent (2017), has been a major meeting point for scholars working with these aims in mind, and the conference in Barcelona (2019) has continued this tradition of expanding the sub-field of empirical translation studies to other paradigms within translation studies. This book is a collection of selected papers presented at that fourth Translation in Transition conference, held at the Universitat Pompeu Fabra in Barcelona on 19–20 September 2019.

While maintaining the interdisciplinary focus and the strong standing of corpus-aided research in empirical translation studies, the conference has also seen input from fields such as audiovisual translation studies, cross-linguistic discourse studies. The papers in this volume are ordered roughly by the type of language they analyse, on a cline from primarily written to primarily spoken language, passing on its way via a hybrid type of conversational or “informally written” (McCulloch 2019) language. A connecting thread is the analysis of features and style of language in multilingual discourse environments. This can be seen in Josep Marco’s, Éric Poirier’s and Ekaterina Lapshinova-Koltunski’s chapters, which suggest ways of analysing the shifts of informationally and linguistically salient elements in the target texts as well as in Olga Nádvorníková’s, Laura Mejías-Climent’s and Madiha Kassawat’s chapters analysing stylistic features of the target language in entertainment products, as well as in Bisiada’s chapter, which compares the semantic features of particular expressions across languages to represent feminist movements.

Maeve Olohan develops a practice-theoretical conceptualisation of post-editing as one of several activities that make up the changing practice of translating,

alongside other activities such as editing translation memory fuzzy matches. This contrasts with a view of post-editing as a practice in its own right that competes with or complements the practice of translating. She examines how post-editing is reconfiguring translation practice, through changes in constituent elements of the practice, including the tools and materials deployed, the competences and knowing that transpire in practice, and various understandings of the practice. By exploring this reconfiguration, we may extend our genealogical understanding of translation, as a practice that changes over time.

Josep Marco examines Halverson's (2010) Gravitational Pull Hypothesis which draws on cognitive linguistics and bilingual theory to address the problem that translated texts have been shown in some cases to over-, in others to under-represent of typical target language elements, which creates the dilemma that both claims cannot be generally true or predicated of the same set of data. Halverson (2010) argues that patterns of prototypicality in the target language, conceptual structures or the representation of the source language item, and patterns of connectivity are possible cognitive causes of this issue. Marco draws on the COVALT corpus consisting of English-Catalan and French-Catalan parallel corpora and a Catalan comparable corpus to test her hypothesis on the Catalan verb *caldre*.

Olga Nádvoříková argues in her chapter that the Czech transgressive is a non-finite verb form belonging to the cross-linguistic converb category. In contrast with other converbs, the Czech transgressive has a strong stylistic mark and is very rare in contemporary language. Using a parallel multilingual corpus and a comparable corpus of translated and non-translated Czech, the chapter investigates the differences in the frequency of the transgressive in translated and non-translated fiction and non-fiction. The data shows the effect of convergence in both fiction and non-fiction and the effect of stylistic normalisation in fiction only. The results of the potential effect of cross-linguistic interference are much less conclusive, indicating that a thorough contrastive analysis of different language pairs is required first.

Madiha Kassawat argues in her chapter that, in an increasingly globalised world, accessibility to digital content has become indispensable for people around the world, which in turn makes translation indispensable. As the majority of products is promoted for and sold on the internet, their web pages are often localised based on the market. The required speed in this type of work, its tools and process influence the quality of the localised texts, which necessitates their analysis and an exploration of the different interpretations of the same source text in several languages. Her study compares the product descriptions provided in English and localised into Arabic and several French versions.

Preface

Mario Bisiada studies tweets on #MeToo in English, Spanish and German from 2019, revealing how MeToo is most commonly referred to as a “movement” in English and Spanish but as a “debate” in German, a difference that echoes German-language press habits. Based on an analysis of semantic prosody, the chapter demonstrates that words indicating longevity such as *era* and *times* collocate with MeToo in English and Spanish, but not in German. This points to a framing of MeToo as influential and long-term in English and Spanish and as exaggerated and short-term in German. Reflecting this difference, MeToo is talked about in more negative terms in German tweets compared to English and Spanish, as shown by a qualitative analysis of evaluative author stance.

Felix Hoberg investigates the patterns of saccadic eye movement when using Microsoft’s Skype Translator between Catalan and German through a case study of 21 German-speaking participants as part of an overall evaluation of the Skype Translator on a dialogue-oriented level. Despite not having any proficiency in Catalan, these participants had to text-chat with Catalan native speakers via Skype, while the Skype Translator was activated. The sessions were observed by an eye tracking system. The collected data thus represents a naturalistic starting point to evaluate how users structure computer-mediated communication situations when real-time machine translation is involved while having to rely on that output.

Éric Poirier describes an empirical method to screen informational translation shifts in parallel segment pairs extracted from bilingual or multilingual translation corpora, based on character length and lexical word count. The method applies to most known languages and in one or the other of the two translation directions (direct or inverse). The chapter argues that heteromorphic segment pairs, as opposed to isomorphic ones, are more likely to contain informational translation shifts. The objective and reproducible method described in his chapter allows for semi-automatic identification of problematic translations and uncovering of textual and linguistic facts revealing translation processes, contingencies, and determinism.

Laura Mejías-Climent’s chapter is an analysis of the dubbing of the video game *Detroit: Become Human*. She wants to shed some light on the convergences of audiovisual translation studies and localisation from the specific perspective of dubbing, in a product that, in turn, poses some questions to the genre it belongs to. This chapter aims to highlight some of the differences and convergences between AVT and localisation analysing the dubbing synchronies applied in a video game belonging to a genre closer to traditional movies, compared to other adventure games, due to the strong presence of cinematic scenes and the lower level of interaction.

References

Ekaterina Lapshinova-Koltunski analyses English-to-German translations and interpretations, focussing on the variation in English-to-German translation that involves the dimension of mode, i.e. variation between spoken and written language production. She argues that the resulting variation is reflected in the linguistic features of translations and interpretations, e.g. preferences for modality meanings, proportion of nominal or verbal phrases and others. These features offer the opportunity of analysing and modelling the dimensions involved. The methodological focus of her chapter is on quantitative distributions of these linguistic features reflected in the lexico-grammar of texts.

As is evident, the studies are united by a strong empirical aspect, based for instance on corpus analyses or eye-tracking approaches, and a few also have a theoretical focus on features of translated language or the effect of post-editing on translation practice at the workplace. The studies come from subfields as diverse as audiovisual translation, machine translation, cultural mediation and contrastive linguistics and include a range of languages such as Spanish, German, Arabic, Czech, Catalan, French as well as English.

Mario Bisiada

Barcelona, November 2020

References

- Carl, Michael, Srinivas Bangalore & Moritz Schaeffer (eds.). 2015. *New directions in empirical translation process research: Exploring the CRITT TPR-DB*. Heidelberg: Springer.
- de Sutter, Gert, Patrick Goethals, Torsten Leuschner & Sonia Vandepitte. 2012. Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures* 13(2). 137–143.
- de Sutter, Gert & Marie-Aude Lefer. 2019. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives: Studies in Translation Theory and Practice* 28(1). 1–23.
- de Sutter, Gert, Marie-Aude Lefer & Isabelle Delaere. 2017. *Empirical translation studies: New methodological and theoretical traditions*. Berlin: de Gruyter.
- Halverson, Sandra. 2010. Cognitive translation studies: Developments in theory and method. In Gregory M. Shreve & Erik Angelone (eds.), *Translation and cognition*, 349–70. Amsterdam: John Benjamins.

- Hansen, Gyde. 2002. *Empirical translation studies: Process and product*. Copenhagen: Samfundslitteratur.
- Ji, Meng. 2016. *Empirical translation studies: Interdisciplinary methodologies explored*. London: Equinox.
- Ji, Meng, Lidun Hareide, Defeng Li & Michael Oakes (eds.). 2016. *Corpus methodologies explained: An empirical approach to translation studies*. London: Routledge.
- Ji, Meng & Michael Oakes (eds.). 2019. *Advances in empirical translation studies: Developing translation resources and technologies*. Cambridge: Cambridge University Press.
- Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.
- Laviosa, Sara, Adriana Pagano, Hannu Kemppanen & Meng Ji. 2016. *Textual and contextual analysis in empirical translation studies*. Heidelberg: Springer.
- McCulloch, Gretchen. 2019. *Because internet: Understanding how language is changing*. London: Random House.
- McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett. 2019. Abandon statistical significance. *The American Statistician* 73. 235–245.

Chapter 1

Post-editing: a genealogical perspective on translation practice

Maeve Olohan

University of Manchester

This paper develops a practice-theoretical conceptualization of post-editing, as an activity that increasingly forms part of translation practice. This contrasts with a prevailing conceptualization of post-editing as a practice in its own right, competing with or complementing translation practice. Adopting a genealogical perspective, I trace this particular evolution of the translation practice through some of the interdependent changes in the materials constituting the practice, the competences or know-how that transpire in the practice, and the meanings of the practice, in particular as constructed through the discourse of language service providers and the international standards that normatively regulate the practice. The paper concludes with some implications of this practice-theoretical approach for future research on post-editing.

1 Introduction

Machine translation (MT) is increasingly deployed by language service providers (LSPs) and translators. This means that some professional translators work more often with machine translation outputs, editing them to make them fit for purpose, an activity known as post-editing (henceforth PEMT but also referred to elsewhere as MTPE).

In this paper I develop a practice-theoretical conceptualization of PEMT as one of several activities that make up the changing practice of translating, alongside other activities such as editing fuzzy matches retrieved from translation memory (TM). This contrasts with a view of PEMT as a practice in its own right that competes with or complements the practice of translating. I examine how PEMT



Maeve Olohan

is reconfiguring translation practice, through changes in constituent elements of the practice, including the tools and materials deployed, the competences and knowing that transpire in practice, and various understandings of the practice. By exploring this reconfiguration, we extend our genealogical understanding of translation, as a practice that changes over time. The account of translation practice put forward in this paper has been abstracted from my own workplace observations in several LSPs and formal and informal interviews with translators, translation project managers and LSP managers. Practitioner performances and articulations from particular sited practices are not further elaborated on here; rather, those understandings are theorized and contextualized through relevant practice-theoretical and translation studies scholarship to conceptualize a particular trajectory of the practice of translation.

The paper first outlines the contours of a practice-theoretical understanding of translation. It then examines the reconfigured practice in more detail, focusing on how PEMT-related changes in the practice are linked to changes in material elements, know-how and meanings participating in it. To conclude, I reflect on how empirical research can shed further light on PEMT activity in translation, by employing methods typical of ethnographic research but also complementing them by other methods that may be productive in studying ongoing changes in translation practice.

2 Understanding translation as a practice

Practice theory refers to a range of theoretical approaches to the study of social practices, as elaborated by [Schatzki \(1996; 2002\)](#), [Reckwitz \(2002\)](#), [Shove et al. \(2012\)](#), [Nicolini \(2012\)](#) and [Warde \(2016\)](#), among others. Although there is some divergence in thinking and terminology, these contributions share some specific priorities and concerns. Crucially they place practices at the centre of their conceptualizations and analyses, conceiving the social world as a plenum of practices in which practices hang together in constellations or complexes ([Schatzki 2016](#)). This focus on practice entails key ontological and methodological shifts ([Postill 2010](#); [Reckwitz 2002](#)), moving away from research that seeks to explain social phenomena through individual actions, informed by rational choice theory. Similarly, a focus on practices also entails a shift away from systems-oriented thinking that looks for explanations in social systems and norms. [Reckwitz \(2002: 250\)](#) cautions against trivializing practice theory because much of its terminology appears to resemble our everyday descriptions of human behaviour. Terms such as *activity*, *material*, *competence*, *knowing*, *meaning*, *practical understanding*, *gen-*

eral understanding, *rules* and, of course, *practice* are used in those more technical senses below.

Definitions of practice vary. One of the most influential is Schatzki's (2001: 2) depiction of practices as "embodied, materially mediated arrays of human activity centrally organized around shared practical understanding". This definition emphasizes the human-centred nature of the activities but also the situated, embodied performances by practitioners. The activities are bodily "doings and sayings" that unfold in time and space (Schatzki 1996: 89). The activities are mediated by materials, i.e., non-human entities, whether physical, biological, chemical or artefactual. Moreover, the activities are organized around shared practical understanding, i.e., knowing how to identify the doings and sayings that make up a practice and knowing how to carry them out. In addition to practical understanding, three other kinds of practice-organizing elements are proposed (Schatzki 2002), namely general understandings, rules and teleo-affective structures.

General understandings refer to the general sense or ethos of a practice, otherwise described as the "senses of worth, value, nature or place of things, which infuse and are expressed in people's doings and sayings" (Schatzki 2012: 16). For example, in (2021: 76–77) I draw on corpus data to show that LSPs, in their promotional discourse, seek to convey a general sense of translating as being in the service of globalized trade.

For Schatzki, rules are the explicitly normative formulations that are found in regulatory or legislative frameworks. Rules of this kind figure less in the organization of translating practice than in other professional practices, e.g., medicine or chartered accounting, although other mechanisms conveying what is considered an acceptable performance of translating include tests, accreditations, prizes, client feedback, etc.

Finally, the teleo-affective structure brings together the teleological and the affective dimensions of practice organization. The first refers to the ends, projects and tasks that hierarchically order the activity (Schatzki 2002: 80). These ends, projects and tasks function normatively, in the sense that they are the ends, projects and tasks that participants ought to realise as they produce what can be considered an acceptable performance of the practice. Practitioners complete tasks, which are part of projects, which serve ends. Freelance translating in different settings may pursue a number of ends, ranging, for example, from earning a living to disseminating a particular cultural product or to supporting a humanitarian cause. The affective dimension refers to the emotions and moods that are permitted, encouraged or considered acceptable or obligatory for participants to exhibit when carrying out a practice. Translating practice is generally not

Maeve Olohan

strongly ordered by affectivity, so there are often no obvious affect-related expectations; a translating practice may be deemed to be acceptable whether the practitioner is excited, bored or despairing in its performance, for example.

For analytical purposes, it is often helpful to consider different constituent elements of practice in turn, as will be done in the next sections, but it is important to emphasize the interconnectedness of those elements, without which the practice would not exist. Reckwitz's (2002: 249) definition of a practice helpfully highlights the interconnections between elements that produce a "routinized form of behaviour" consisting of "forms of bodily activities, forms of mental activities, 'things' and their use, a background knowledge in the form of understanding, know-how, states of emotion and motivational knowledge". A key consequence of this thinking is that a practice cannot be reduced to any single element (Reckwitz 2002: 249). This has implications for empirical investigations and research methods, as noted in the concluding section of this paper.

Practice theory has shed light on practices in many domains, from everyday practices of eating (Warde 2016) and consuming energy (Shove et al. 2015) to professional practices performed in workplaces such as hospitals (Nicolini 2011), schools (Kemmis et al. 2012) and engineering and construction sites (Buch 2015), to give just a few examples. Many of these practices have been theorized and empirically investigated by researchers working in sociology, organization studies and consumption studies but practice theory has made its way into many other academic disciplines too, for example, political science (Jonas & Littig 2017) and media studies (Bräuchler & Postill 2010).

Practice scholars pursue a range of research questions but are often interested in investigating the nature of specific practices, how practices emerge and evolve, as well as how they endure or fade (Shove et al. 2012; Schatzki 2019). An understanding of how practices interconnect with and are dependent on other practices is also highly relevant (Hui et al. 2017; Spaargaren et al. 2016). Underlying these investigations is a distinction between practice as performance and practice as entity. Individual performances of a practice occur in specific times and spaces; they are "continual improvisations" along "more or less precise or fuzzy parameters" (Warde 2016: 46). The practice entity is the encapsulation or abstraction of what makes the performances recognisable as acceptable performances of the practice. A practice requires repeated performance to endure, and changes in performances may eventually lead to changes in the practice entity. Formulated in another way that is helpful for the purposes of this paper, practices are "open-ended, spatial-temporal sets of organized doings and sayings" and they can be extended through additional doings and sayings (Schatzki 2019: 28).

The value of theorizing translation as a practice and researching it empirically rests in the holistic perspective, compared with approaches to studying translation that may focus on translation product, cognitive process or practitioner. A practice-theoretical perspective encourages us to consider all of the various elements that make up the practice of translating, including the human body, material entities, know-how and meanings of the practice. In focusing on often overlooked material and embodied elements of practices, as well as the know-how that is enacted in practices, practice theory provides a productive framework for a dynamic and materially aware understanding of translation practice and for an examination of the emergence of new configurations of the practice (Olohan 2021). This paper examines aspects of how translation practice is being reconfigured through the integration of the PEMT activity, by considering the interdependence of the diverse elements that constitute the practice.

3 The proliferation of MT and post-editing

Since 2016 the dominant MT model has been neural machine translation (NMT), which displaced the predecessor model of statistical machine translation (SMT). NMT relies on machine learning performed via neural networks, and NMT developers, starting with Google and Microsoft, were quick to claim fairly substantial increases in translation quality for the new approach when compared with SMT (Wu et al. 2016). Unless otherwise stated, MT refers here to NMT.

The concepts of pre-editing and post-editing emerged some decades ago in the era of older, rule-based MT technologies and were applied more often in the context of research systems than in commercial applications. Post-editing (PE), as currently understood, is defined in international standard ISO 18587:2017 as editing and correcting machine translation output (British Standards Institution 2017). Distinctions between different PE modes will be made below but it is first useful to consider PEMT's prevalence in today's language services sector. There are no sector-wide measures of how widely PEMT is being performed in language services but some indicators attest to the ever-increasing deployment of MT by LSPs for their clients. For example, the *Slator 2019 Language Industry Market Report* (Faes 2019: 16) notes that MT is "well on its way to becoming the single most important productivity enhancement technology for human translators". This report also acknowledges the use of MT by enterprises who thereby forego the intermediary services of LSPs; they refer to the market for "stand-alone" or "pure play" MT, described as MT without any human translation services, i.e., raw MT output that is used without post-editing by linguists (Faes 2019: 16).

Maeve Olohan

Raw MT is considered useful for content that would otherwise be too ephemeral or too voluminous to be commissioned for human translation, while human involvement is usually preferred for the production of high visibility target-language content on which commercial reputations rest. User-generated content (e.g., customer reviews) and customer support are content types for which raw MT is considered acceptable in some situations. These preferences are confirmed by the European Commission's survey on likely uptake of MT in small and medium enterprises (*Directorate-General for Communications Networks, Content and Technology 2020*), where most respondents considered MT to be useful for understanding websites or social media, gathering information about or corresponding with companies or partners, and purchasing and selling products or services, including offering after-sales service. By contrast, respondents expressed a clear preference for human translation for activities relating to negotiating and signing contracts, resolving conflicts in commercial transactions, dealing with public administration in other countries and conducting marketing and promotional activities.

These survey responses of business representatives highlight the interdependence of the practice of risk management and the practice of translation. A concrete example of a company's concern to avoid reputational damage from potentially low quality raw MT is offered by *Schmidtke & Groves (2019)* in their account of the deliberations of Microsoft as it sought to introduce raw MT into software localization, having previously published some raw MT in technical and end-user support documentation for Microsoft Office. This is an example of one of three levels of NMT-related risk identified by *Canfora & Ottmann (2020)*, namely the damage that can be incurred by clients and end users from errors in the MT output. This is seen as an issue for NMT in particular because errors are not readily predictable and the output can resemble a convincing piece of target language discourse. Errors of accuracy can therefore be overlooked by post-editors or revisers. A second level at which risks have to be managed concerns the attribution of liability and accountability when NMT tools are used and damage is incurred. There is no legal clarity on this matter as yet. While traditional legal notions of misconduct or negligence apply to human behaviour and not AI systems, they could potentially be applied to those who produce, own or use the AI system, thus also possibly extending to post-editors (*Canfora & Ottmann 2020*: 63). The third level at which risks are incurred and must be managed are those related to data security, a particular problem when NMT is used via free, online, generic MT services (*2020*: 64).

The use of raw MT has not reduced demand for translations of a specified (high) quality for which human involvement, through PEMT activity, is gener-

ally required and expected. A European survey of 298 LSPs and 905 individual translators conducted in 2018 reported that more than half of companies and individuals were using MT in some form (ELIA et al. 2018). In a worldwide survey of 7,363 translators and interpreters at the end of 2019, almost all respondents (97%) provide translation services, 72% offer editing or proofreading, and PEMT is the next largest service offered, by 35% of respondents (Pielmeier & O'Meara 2020). However, 55% of respondents report that they use MT, including on projects when the client does not request it, which means they are also post-editing as part of their own translation services. 23% of the MT users find that they deliver better quality when they use MT, and 52% say that MT speeds up their work (Pielmeier & O'Meara 2020: 45). The aforementioned Slator report asserts that there is an increasing demand for “professional linguists who can interact with machine translation output”, given that LSPs’ corporate clients are looking for bespoke MT solutions tailored to their content, workflows and preferences (Faes 2019: 22).

Approaches to PEMT activity are discussed in language services and academic research. The ISO 18587:2017 standard and numerous MT technology providers differentiate between full and light PE. Full PE is the “process of post-editing to obtain a product comparable to a product obtained by human translation”, while light PE is a “process of post-editing to obtain a merely comprehensible text without any attempt to produce a product comparable to a product obtained by human translation” (British Standards Institution 2017: 2). This distinction, using similar or different terminology, is also made by MT promoters and developers, e.g., TAUS (2015), KantanMT (2019) and SDL (2020). However, it may be misleading to suggest that there are two (or more) PE modes that are easily defined and recognized, and in demand in commercial practice, or that translators can easily switch between them. Light PE appears to be much less relevant in practice and, indeed, the ISO standard restricts its detailed prescriptions to full PE. Similarly, formulations that refer to the product of human translation as the aspirational goal of PEMT are common and perhaps understood as a shorthand but may be unhelpful, since they reflect unrealistic notions of all human translation being of invariably appropriate quality.

Finally, another indicator for the increasing importance of PEMT is the general growth of research on the phenomenon. Much of the earlier research investigated PEMT as performed by students or novices or as a stand-alone activity, typically also in experimental settings. A relatively high proportion of studies also focus on MT research systems and are more concerned with performance or assessment of the technologies rather than PEMT *per se* or as it occurs with commercial systems. However, there is growing interest in studying PEMT in the

Maeve Olohan

professional workplace (e.g., Góis & Martins 2019; Vardaro et al. 2019; Macken et al. 2020) and in assessing the acceptability of PEMT for end users in typical usage settings (Girletti et al. 2019). Accounts of professional deployments of PEMT are also becoming more prevalent in the literature (Zaretskaya 2019a,b; Kosmaczewska & Train 2019; Premoli et al. 2019; Nunziatini 2019).

Having outlined a practice-theoretical framework and having established that PEMT is increasingly deployed in language services, we now examine in more detail the PEMT-related evolution of translating practice. To do this, we trace a selection of the changes in the constituent elements of the practice, namely its materials, competences and meanings, following Shove et al. (2012). Materials include artefacts such as software and hardware, other tools, devices and infrastructures, as well as the human body. Competence refers to practical understanding or know-how. Meanings bring together general understanding, teleo-affectivity and other elements that normatively organize the practice.

4 Changing materials

Material entities of many kinds participate in practices, including humans, organisms, phenomena of nature and artifacts (Schatzki 2019: 39). Shove (2017) distinguishes different roles that may be played by material entities in practices, namely as infrastructures, devices and resources. It is beyond the scope of this paper to consider all relevant material aspects of PEMT and translation practice, so we will consider one example of each of these three categories in turn, to illustrate how changes in materials shape changes in the practice.

The first role to be considered for material entities is that of infrastructure. These are understood as things in the background that are necessary for the practice to be performed but are not directly engaged with it (Shove 2017). For translation with or without MT, the infrastructure that is usually necessary for the practice to be performed includes buildings, lighting, heating, electrical power, the Internet and information and communications technologies, among other elements. An infrastructural addition that is specific to the PEMT activity is the NMT engine. As noted above, NMT relies on neural networks, and an NMT engine has been trained and tested on language data, usually in large quantities and for a specific language pair. In addition, it is often customized or fine-tuned by adding further smaller datasets comprising texts from a specific subject domain, in order to improve the quality of outputs when deployed for that domain.

The rapid advances in NMT and other machine-learning technologies over the past five years are themselves partly attributed to the material changes in com-

puter systems that came with the realization that neural networks can run relatively efficiently on graphical processing units (GPUs). GPUs are the computer processors designed for rendering graphics and games, and it was discovered that they outperform conventional processors (central processing units, or CPUs) for implementing and training neural networks. The capabilities of GPUs and subsequent enhancements, as well as the availability of large datasets for training MT engines, thus enabled significant developments in machine learning, including NMT, which, in turn, are changing the trajectory of the translation practice by reshaping some of the activities that constitute the practice.

The building of NMT engines and their adaptation to domains is technically complex and beyond the capabilities of most LSPs and individual translators (see [Gupta et al. 2019](#) and [Silva 2019](#) for descriptions of some of the processes involved). Thus, the viability of MT deployment for an LSP can be considered in terms of the computational infrastructures required. Some LSPs, like SDL and Tilde, develop MT systems for their own use in their language services businesses and also for sale to other LSPs or translators. However, most LSPs are dependent on buying an NMT service from a specialist provider, either as an off-the-shelf product or as a customized engine that the provider will build, test and perhaps maintain and host on their behalf. As noted by [Faes \(2019: 33\)](#), increasing commercial deployment of NMT is being driven by some of the global, big tech companies: Microsoft, Google, Facebook, Amazon, IBM, SAP, Salesforce, Alibaba, Baidu, iFlytek and Sogou. These companies have invested very heavily in developing NMT, initially to help them to deliver their core businesses, but some then take advantage of the opportunity to sell the MT technology to smaller companies, either as a stand-alone service or as part of a wider suite of technological applications. MT technologies and services are also being sold to LSPs and linguists by another group of technology companies for which MT is their core business; these include DeepL, KantanMT, Omnisien Technologies, Systran and PROMT, among others. Thus, LSPs are often relieved of the material requirements to purchase and run specific hardware or software or to ensure data security and confidentiality on their own premises. However, the potential success of customized MT engines is dependent on LSPs being able to provide large corpora of source texts and translations for the language pair and subject domain so that the system can be appropriately trained, and they still need to be able to give clients the necessary assurances regarding data security for engines hosted by a third party.

The second role to consider for materials is as devices, i.e., things that are in the foreground of practices and participate directly in them ([Shove 2017](#)). Devices

Maeve Olohan

that are undergoing material changes as PEMT is integrated into translation practice include the translator's desktop environment. The most typical deployment of MT is through an application programming interface (API) that connects the NMT service with computer-assisted translation (CAT) tools. Thus, the CAT environment combines resources from MT, translation memory (TM), and terminology management tools, and translators using MT work in their usual editor and follow workflows that are familiar from their non-PEMT practices. Typically, the TM software first retrieves, from its database, full matches (i.e. 100%) and fuzzy matches (typically 75% to 99%) for segments of the source text that formally resemble source text segments already stored in the TM. Then, for those segments of text for which there are no full or fuzzy TM matches, an MT suggestion is generated and inserted into the editor, so that the translator is confronted with suggestions for all segments of text and generally proceeds to post-edit the MT suggestions and edit TM matches to produce a translation of the requisite quality (see [Zaretskaya 2019a](#), [Premoli et al. 2019](#) and [Nunziatini 2019](#) for descriptions of this process as implemented in different LSP settings).

The resources are handled in this way because an assumption is made that a fuzzy TM match is more useful to the translator than an MT suggestion, so the TM takes precedence and the MT is only provided where the TM can offer no assistance. However, as a study at TransPerfect shows ([Zaretskaya 2019b](#)), when NMT engines are customized for the domain and the quality of the MT suggestions is high, it is desirable to give the MT suggestions priority over fuzzy TM matches. In those cases, as demonstrated for short segments of text (typically 4 to 6 words) in the TransPerfect research, the TM fuzzy matches required more editing than the MT suggestions (as measured by the post-edit distance, PED).

Although the translation practice still happens in the familiar interface, it is changed materially by the change in quantity and type of data presented to the translator, and the material organisation of that data. ISO 18587:2017, the international standard for post-editing, makes an explicit, material distinction between translation and post-editing by describing PEMT as involving three texts: the source text, the MT output and the final target text, while translation only involves two ([British Standards Institution 2017](#): 5). In the working environment just outlined, the translator deals not only with MT output but also with TM matches, with some visual differentiation through colour coding and the addition of metadata.

NMT systems operate on a sentence level and translation suggestions are proposed segment by segment, as is also the case with TM (where segments are typographically delimited and often equate to a sentence, heading, bullet point, etc.). However, as argued in [Olohan \(2021: 51–54\)](#), since many texts follow a narrative

structure, segment-based organisation of TM databases is at odds with the texts' narrative logic. Moreover, the algorithmic nature of NMT is at odds with both database and narrative logic. One manifestation of the MT's algorithmic logic is its relative lack of transparency compared to TM suggestions. Translators prefer to have some information on provenance on and the nature of TM matches (Teixeira 2014; Cadwell et al. 2018). However, the inner workings of neural networks are inscrutable so it is virtually impossible for translators (or system developers) to predict MT outcomes and it is difficult to explain MT errors. These clashes in narrative, database and algorithmic logic underlying the material configuration of data may be at the heart some of the frustrations experienced by translators working with TM and MT (e.g. Moorkens & O'Brien 2017; LeBlanc 2014; Cadwell et al. 2016).

Suggestions for potential improvements that are not yet generally implemented in commercial MT applications include MT quality estimators that are meaningful in the context of the post-editing process, e.g., identification of segments that require revision, or estimates of post-editing efficiency, rather than abstract quality metrics (Stahlberg 2019). Other desired changes are delivered, to some extent, by interactive and adaptive MT systems, where the MT suggestion is changed on the basis of what the user types, and the system also learns from the corrections made (Daems & Macken 2019; Karimova et al. 2018). Pielmeier & O'Meara (2020: 43) report that, of their 2,059 respondents to questions about MT use, 71% agree with the statement "I prefer to work with adaptive MT like Lilt rather than raw MT output". Lilt promotes its interactive, adaptive MT for use with "high-value content" in particular. It changes the material working environment of the translator further, in that fuzzy TM matches are no longer helpful, so the translator is working with MT suggestions for all segments.

The final role to consider for materials is as resources, i.e., things that are used up or consumed in the practice (Shove 2017). Translation practices consume resources, with or without the deployment of MT (see also Cronin 2017). However, as might be concluded from the description of computer processors above, the building and training of NMT engines is considerably more resource-intensive than the compilation and use of TMs. Indeed, NMT engines not only consume more processing resources but also require longer training times than the previous SMT systems. In resource terms, the technology developers appear to be moving in two different directions. On the one hand, there are attempts to enable machine learning applications like NMT to use CPUs more efficiently so that they may be run on conventional PCs and mobile devices, to reduce both the need for specialized hardware and the training times (Devlin 2017). SDL's latest NMT product, for instance, the Enterprise Translation Server, is offered in

Maeve Olohan

both GPU and CPU modes. The benefit of running NMT with CPU is presented as lower infrastructure costs, although it entails compromises on speed or quality. On the other hand, MT research is also pulling in the opposite direction, towards massively multilingual NMT systems that require billions of words as data and very substantial computing power (Aharoni et al. 2019).

5 Changing competences

Translation scholars have long been interested in competences, understood and articulated in a variety of different ways, ranging from Pym's (2003) minimalist definition of translation competence to the complex, multi-dimensional models and competency frameworks proposed by the PACTE (Hurtado Albir 2017), TransComp (Göpferich 2013) and EMT (EMT Board 2017) projects, among others. In Olohan (2017) I argue that a focus on knowing-in-practice (i.e., knowing as it transpires in and through practice) is desirable because it pays due attention to the situated, embodied, relational, and materially mediated aspects of knowing, alongside the embrained knowing that is more traditionally accorded primacy in discussions of competence, training and education. In Schatzki's terms, this is practical understanding, as introduced above, i.e. knowing how to perform the doings and sayings that constitute the practice and also recognizing when these are performed. Practical understanding is alternatively described as a "a battery of bodily abilities that results from, and also makes possible, participation in practices" (Schatzki 2001: 9) Shove et al.'s understanding of competence similarly encompasses "skills, know-how and technique" (2012: 15).

Multi-dimensional translation-related competency frameworks have generally been developed with professional practice in mind, and through consultation with practitioners and other relevant stakeholders. Such frameworks typically seek to formalize the practice by formulating an understanding of what it means to be competent that can serve as a competency standard. However, there is some variation in the practice that is being addressed. A framework such as the EMT's is strongly focused on learning outcomes and arguably formalizes what it means to be competent in the learning practice rather than competent in the translation practice. Others, such as the PACTE framework, focus on capacities that professional translators should demonstrate, formulated as list of tasks or activities that practitioners should be able to complete. In both kinds of cases, outcomes are foregrounded, with relatively less consideration of the performances from which those outcomes ensue. A practice-theoretical approach, by contrast, is interested in the situated, social, embodied and materially mediated nature of the

knowing that makes participation in the practice both possible and appropriate. It also recognizes that there are different ways of carrying on a practice.

Where scholars have considered post-editing as a separate practice from translation, they nonetheless develop PEMT competency frameworks that are strikingly similar to those for translation (e.g., Nitzke et al. 2019). These similarities have also been codified in the international standard for the post-editing of machine translation output, ISO 18587:2017, where a substantial focus is on competences. A comparison of the post-editing standard and the standard for translation services, ISO 17100:2015, reveals an almost identical description of competences, classified as translation competence; linguistic and textual competence in the source language and the target language; competence in research, information acquisition, and processing; cultural competence; technical competence, and domain competence (British Standards Institution 2015; 2017). Likewise, the qualifications required by the standards are very similar. Translators are required to have a formal degree in translation or full-time professional experience in translating, or a combination of professional experience and a degree in another field. For post-editors, the formal degree simply needs to include significant translation training (so it can be a more general degree in language studies). The professional experience required can be in translating or post-editing. These prescriptions of substantially similar know-how for PEMT and translation provide further support for this paper's argument that PEMT constitutes an additional activity that may take place as part of the translation practice, rather than a separate, recognizable practice in its own right. The overlaps extend to the standards' expectations on the role of formal training in abstracting and codifying that know-how.

Despite these competence-related convergences, Slator's *Neural Machine Translation Report* (Slator 2019) noted a growing demand for qualified post-editors and a growth in companies developing training courses to fill this demand. LSPs who expect their linguists to perform PEMT also frequently acknowledge the need for training in this activity. Transperfect, for example, provides training and a certification programme in PEMT for some thousands of freelance linguists (Zaretskaya 2019a: 137), and training was also required at TranslateMedia when post-editors switched from editing SMT to NMT (Kosmaczewska & Train 2019: 170).

The training that is delivered tends to address additional requirements that are given in ISO 18587:2017 in a section entitled 'Professionalism'. Here it is stipulated that post-editors should have general knowledge of MT technology, basic understanding of common MT errors and a general knowledge of CAT tools (British Standards Institution 2017: 8). This know-how is deemed important, not only for the execution of changes to the MT output but also because LSPs

Maeve Olohan

collect data on how MT is used and they usually require translators to report on frequently encountered errors. These reports are fed back to technology developers, to contribute to improving the MT engines.

The standard also requires post-editors to have “the knowledge and ability to establish whether editing MT output makes sense, in terms of time and effort estimation”, and the “ability to follow instructions received” and “to focus on specific issues and make specific corrections as given” ([British Standards Institution 2017](#): 8). These aspects address key know-how from an LSP’s perspective, often linked to post-editing speeds and productivity. A decision-making process is mapped by [Nitzke et al. \(2019\)](#) who propose a decision tree to help users to decide whether or not to use MT and how to approach PE. Factors to be considered in making the decision include possible risks and benefits, resources needed and available, data sensitivity and security issues, quality of MT output produced and the client’s or end user’s quality requirements. Once MT has been deployed, the two-second rule ([Graciet 2018](#)) encapsulates the rapid decision making required of translators about whether an MT suggestion is usable with editing, or whether the translator needs to produce a translation from scratch.

[Blue & Shove \(2016\)](#) posit that practices constitute the knowledge that they need to continue to exist, and that there are various mechanisms by which this happens. The translation practice takes some of its know-how from closely related practices, such as the practice of learning a language or the practice of writing literature or other genres. It cultivates other aspects of its know-how, for example for the PEMT activity, through know-how that is already embedded in material forms. With MT integrated materially into the familiar CAT interface, translators know how to interact with MT suggestions through their previous interactions with TM matches, since the TM’s segment-focused database logic is typically extended to the PEMT activity. At the same time, material differences that have an impact on knowing, as noted above, include differences in metadata available for TM matches and MT suggestions.

6 Changing meanings

Meaning is used as an overarching term to encompass “symbolic meanings, ideas and aspirations” of a practice ([Shove et al. 2012](#): 14), alternatively thought of as forms of understanding, states of emotion and motivational knowledge ([Reckwitz 2002](#): 249). Competences, as discussed above, relate to the practical know-how required by the practice and some codifications or prescriptions pertaining to that know-how. Here we consider other organizing elements of the practice,

selectively focusing on changes in general understandings of PEMT activities in translation practice among LSPs and among translators, where general understandings are understood in Schatzkian terms (e.g., [Schatzki 2002](#)) as general senses of the nature of things which find expression in the doings and sayings of a practice.

[Welch & Warde \(2017\)](#) consider general understandings of practices as sometimes tacit in the background and sometimes discursively articulated. Discursive articulations related to the use of MT in language services are clearly shifting. MT, as a service offering, was much less visible in industry discourse just a couple of years ago than it is now. In their online promotional material of 2018, the world's largest LSPs were mostly concerned with assuring clients that their texts would be translated by human translators to the highest levels of quality, using CAT tools (not MT) for productivity gains ([Olohan 2021](#): 76). A small number of these LSPs still do not offer MT and do not acknowledge its existence but most of them now generally promote MT as bringing benefits to clients, usually due to the need to translate greater volumes faster. Some articulations that are representative of the largest LSPs (by revenue, as listed in CSA Research's annual LSP rankings) are as follows:

Linguistic computing has come a long way over the decades, and in recent years, the quality and cost of machine translation (MT) solutions has harmonized with demand and time-to-market requirements.

[Janus](#)

To meet tight deadlines for large translation volumes while keeping a critical eye on the long-term costs, a machine translation may be a perfect alternative.

[Yamagata](#)

Welocalize language automation like machine translation (MT) delivers translation and content transformation faster across a larger volume of content without compromising quality.

[Welocalize](#)

These and other LSPs offer post-editing as part of their customized MT services, as a means for clients to achieve a desired level of quality, related to specialist content in particular:

We select and onboard post-editors with linguistic and technical experience in your industry to edit the machine's output to your desired level of quality.

[RWS](#)

Maeve Olohan

The post-editing service complements machine translation. The translator, referred to as “post-editor” in this case, uses his [sic] knowledge to harmonise the pre-translated text in order to make it easier to understand and to respect the terminology used in your sector.

Acolad

With the aim of making Machine Translation (MT) work for each translation, **we always advise our clients** to use it under prior human supervision (training, personalization, adaptation) and/or subsequent human editing (human revision of the content produced by the translation machine.)

Linguaserve (emphasis in the original)

It should be noted that these LSPs tend to offer MT or PEMT as a distinct service offering or option, or as a standard approach for particular domains of activity. However, in most cases the promised PEMT end product is not depicted as qualitatively different from what they promise as the product of translation. The client is not to expect any discernible difference in their translations, regardless of the combinations of activities that produce them. An exception in this dataset is seen in the discursive articulation by Morningside Translations, which stresses the cost savings for high volumes but also explicitly tempers quality expectations:

Machine translation is a powerful tool for lowering costs and accelerating turnaround times for high-volume document translation projects, though its quality is still far from being on par with human translation. [...] It can help you get the “gist” of a document when subpar quality is sufficient.

Morningside Translations

Addressing the teleo-affective or motivational dimensions of the practice, LSP managers, when describing how they introduce PEMT activities to their workflows, often mention a reluctance on the part of their translators to be involved in PEMT projects (e.g., [Premoli et al. 2019](#); [Kosmaczewska & Train 2019](#)). In CSA Research’s large-scale survey ([Pielmeier & O’Meara 2020](#)), 8,794 translators were asked to choose the task that they “would prefer to do when given the choice” and 89% chose translation, while 8% chose editing human translation and only 3% chose editing machine translation. As with the introduction of translation memories a few decades ago, this reluctance is sometimes interpreted as a reluctance to embrace new technology but this is an overly simplistic interpretation; the same survey data shows that only 7% of respondents are not very confident

trying new language technology. Other motivational factors are therefore likely to be much more relevant. Focus group studies such as Cadwell et al.'s (2018) have uncovered several of these, including translators' expectations of poor quality MT output, the potential degrading of their translation abilities or creativity through PEMT, and the prospect of MT eventually replacing human translators. It should be noted that the same translators also gave several reasons in favour of working with MT. Through real-time logging of translation workflows and a follow-up survey at the Directorate-General for Translation of the European Commission, Macken et al. (2020) also identify some of the factors that motivate translators' preferences for working with MT. These were mostly related to their impressions that they worked faster with MT than without; and, for most but not all translators, this was backed up by the researchers' measurements.

Attitudes of salaried translators in institutional environments, where some of the risk factors associated with MT use are managed by the institution, can be understandably different from freelancers, in sometimes precarious work situations. Nunziatini (2019) reports on an MT implementation in the financial services domain in which translators' reluctance to engage in PEMT was overcome, to some extent, by continuing to pay the full word rate in the pilot phase of the implementation. The question of how translators are remunerated for PEMT should perhaps not be underestimated as playing a part in motivation; many language professionals are suffering from downward pressure on rates (Pielmeier & O'Meara 2020: 60) and this can be exacerbated by other practices in the sector. Finally, Kosmaczewska & Train (2019) note that translators' initial reservations were overcome by their interest in continuing to work on their client's content and to use their acquired experience, as they changed from a human translation to a PEMT workflow. These observations serve to highlight the need for translation research to consider more closely those other practices, such as the management of resources, when seeking to understand the complexities of the translation practice.

7 Conclusion

Looking through a practice-theoretical lens, this paper has illustrated some of the changes in materials, competences and meanings that have recently reshaped and continue to transform the translation practice as it is expanded to include the activity of PEMT. The interdependencies of these elements has also come to the fore; changes in one element often bring about changes in others. In addition, thinking about translation in this way highlights the importance of connections

Maeve Olohan

between the translation practice and other practices, whether they are training NMT engines, buying and selling MT services, managing risk or balancing budgets in LSPs.

I conclude by reflecting on how this reconfiguration of translation practice can be studied empirically. Generally, practice research relies heavily on real-time observations of situated practice performances, often in combination with qualitative, ethnographic interviews. These methods allow practices to be made visible, articulated and reflected on by practitioners and then mediated and theorized through research practices. [Nicolini \(2009b\)](#) advocates an approach that involves “zooming in” on the accomplishment of a practice in a particular setting and then “zooming out” to focus on the texture of the practices with which it is connected. Translation practices incorporating PEMT activities were initially studied predominantly in experimental settings and sometimes among students, novice translators or those with little prior exposure to the PEMT activity. Increasingly, they are being observed by translation researchers in real time in their everyday occurrences (see, for example, [Macken et al. 2020](#)). There remains considerable scope for this kind of focus on a specific sited practice to be accompanied by a “zooming out” to the textures of connected practices.

Variations on ethnographic research (see [Katz 2019](#)) that can be transposed to practice research are also worth considering for the study of translation. Iconic ethnography, for example, focuses on a small number of practitioners or settings considered particular representative of a type. In the realm of PEMT and translation, an example would be the practice performed by translators designated as MT superusers or similar within LSPs, whose practice is held up as an example to others or who are responsible for instructing, guiding or supporting the practices carried out by others. Comparative analytical ethnography or multi-site ethnography, by contrast, focuses on teasing out the variations in translation practice as enacted at different sites.

Alongside conventional methods of observation and ethnographic interviewing, novel methods for practice research not yet used by translation scholars but offering some potential include Nicolini’s (2009a) “interview to the double”, a form of interview in which the practitioner gives an oral set of instructions to their hypothetical double, who will replace them in the workplace the next day but whose presence there should not be detected by others. This is intended to produce a detailed account of behaviour but is also likely to reveal the situated, normative influences on practices as the practitioner gives an insight into what is considered good practice, what should be done, said or prioritized, based on whose judgement, etc.

Many practice researchers do not see a role for quantitative data to be used alongside qualitative when studying practices but some studies have shown the benefits of a mixed-methods approach, for example, where time-use or diary records have been a useful source of information about the spatio-temporal organization of eating practices in past decades (Warde et al. 2007). As seen in this paper, survey and focus group methods have been helpful in prompting practitioners to report on aspects of their practices. Those *post-hoc* accounts can provide insights into the doings and sayings of specific, sited practices, especially for aspects such as motivations, expectations or preferences in practices. As illustrated by numerous researchers who have studied post-editing effort (e.g., Moorkens et al. 2015; Herbig et al. 2019; Macken et al. 2020), quantitative metrics are invaluable for understanding temporal organization and sequencing of activities within the practice, and technical effort is typically also captured through quantitative data on editing actions. I content that a reflexive, mixed-methods approach to the translation practice is possible, when such quantitative methods are used in conjunction with qualitative studies of practice performances. The imperative on the practice researcher is to resist the temptation to study the practice by attending to just one of its constituent elements, and to seek to understand the interdependencies of constituent elements and the interwoven nature of practices.

References

- Aharoni, Roei, Melvin Johnson & Orhan Firat. 2019. Massively multilingual neural Machine translation. *arXiv – Computation and Language* 1903.00089. <http://arxiv.org/abs/1903.00089v3> (12 August, 2019).
- Blue, Stanley & Elizabeth Shove. 2016. How social practices generate, carry and require knowledge and know-how. In Kevin Orr, Sandra Nutley, Shona Russell, Rod Bain, Bonnie Hacking & Clare Moran (eds.), *Knowledge and practice in business and organisations*, 188–191. London & New York: Routledge.
- Bräuchler, Birgit & John Postill (eds.). 2010. *Theorising media and practice*. New York & Oxford: Berghahn Books.
- British Standards Institution. 2015. *ISO 17100:2015 translation services - requirements for translation services*.
- British Standards Institution. 2017. *ISO 18587:2017 translation services - post-editing of machine translation output - requirements*.

Maeve Olohan

- Buch, Anders. 2015. Studying engineering practice. In Steen Hyldgaard Christensen, Christelle Didier, Andrew Jamison, Martin Meganck, Carl Mitcham & Byron Newberry (eds.), *Engineering identities, epistemologies and values*, 129–145. Cham: Springer. DOI: [10.1007/978-3-319-16172-3_7](https://doi.org/10.1007/978-3-319-16172-3_7).
- Cadwell, Patrick, Sheila Castilho, Sharon O'Brien & Linda Mitchell. 2016. Human factors in machine translation and post-editing among institutional translators. *Translation Spaces* 5(2). 222–243. DOI: [10.1075/ts.5.2.04cad](https://doi.org/10.1075/ts.5.2.04cad).
- Cadwell, Patrick, Sharon O'Brien & Carlos S. C. Teixeira. 2018. Resistance and accommodation: Factors for the (non-)adoption of machine translation among professional translators. *Perspectives* 26(3). 301–321. DOI: [10.1080/0907676X.2017.1337210](https://doi.org/10.1080/0907676X.2017.1337210).
- Canfora, Carmen & Angelika Ottmann. 2020. Risks in neural machine translation. *Translation Spaces* 9(1). 58–77. DOI: [10.1075/ts.00021.can](https://doi.org/10.1075/ts.00021.can).
- Cronin, Michael. 2017. *Eco-translation: Translation and ecology in the age of the anthropocene*. London & New York: Routledge.
- Daems, Joke & Lieve Macken. 2019. Interactive adaptive SMT versus interactive adaptive NMT: A user experience evaluation. *Machine Translation* 33(1). 117–134. DOI: [10.1007/s10590-019-09230-z](https://doi.org/10.1007/s10590-019-09230-z).
- Devlin, Jacob. 2017. Sharp models on dull hardware: Fast and accurate neural machine translation decoding on the CPU. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics.
- Directorate-General for Communications Networks, Content and Technology. 2020. *SME consultation on eTranslation and language technologies*. Brussels: European Commission. <https://ec.europa.eu/digital-single-market/en/news/sme-consultation-ettranslation-and-language-technologies> (5 July, 2020).
- ELIA, EMT, EUATC, FIT Europe, GALA & LIND. 2018. *2018 language industry survey - expectations and concerns of the European language industry*.
- EMT Board. 2017. *European master's in translation competence framework 2017*. Brussels: European Commission.
- Faes, Florian. 2019. *Slator 2019 language industry market report*. Slator AG.
- Girletti, Sabrina, Pierrette Bouillon, Martina Bellodi & Philipp Ursprung. 2019. Preferences of end-users for raw and post-editing NMT in a business environment. In *Translating and the computer* 41, 47–59. London: AsLing.
- Góis, António & André F. T. Martins. 2019. Translator2vec: Understanding and representing human post-editors. *arXiv – Computation and Language* 1907.10362. <http://arxiv.org/abs/1907.10362v1> (28 August, 2019).

- Göpferich, Susanne. 2013. Translation competence: Explaining development and stagnation from a dynamic systems perspective. *Target* 25(1). 61–76. DOI: [10.1075/target.25.1.06goe](https://doi.org/10.1075/target.25.1.06goe).
- Graciet, Céline. 2018. *A translator reviews TAUS post-editing course*. TAUS Blog. <https://elearning.taus.net/tausblog/a-translator-reviews-taus-post-editing-course> (26 August, 2019).
- Gupta, Rohit, Patrick Lambert, Raj Nath Patel & John Tinsley. 2019. Improving robustness in real-world neural Machine translation engines. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 142–148. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).
- Herbig, Nico, Santanu Pal, Mihaela Vela, Antonio Krüger & Josef van Genabith. 2019. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation* 33(1). 91–115. DOI: [10.1007/s10590-019-09227-8](https://doi.org/10.1007/s10590-019-09227-8).
- Hui, Allison, Theodore Schatzki & Elizabeth Shove (eds.). 2017. *The nexus of practices: Connections, constellations, practitioners*. London: Routledge.
- Hurtado Albir, Amparo (ed.). 2017. *Researching translation competence by PACTE group*. Amsterdam & Philadelphia: John Benjamins.
- Jonas, Michael & Beate Littig (eds.). 2017. *Praxeological political analysis*. London & New York: Routledge.
- KantanMT. 2019. *Post-editing guidelines*. http://kantanmt.com/documents/Post-Editing_Guidelines.pdf (25 August, 2019).
- Karimova, Sariya, Patrick Simianer & Stefan Riezler. 2018. A user-study on on-line adaptation of neural machine translation to human post-edits. *Machine Translation* 32(4). 309–324. DOI: [10.1007/s10590-018-9224-8](https://doi.org/10.1007/s10590-018-9224-8).
- Katz, Jack. 2019. On becoming an ethnographer. *Journal of Contemporary Ethnography* 48(1). 16–50. DOI: [10.1177/0891241618777801](https://doi.org/10.1177/0891241618777801).
- Kemmis, Stephen, Christine Edwards-Groves, Jane Wilkinson & Ian Hardy. 2012. Ecologies of practices. In Paul Hager, Alison Lee & Ann Reich (eds.), *Practice, learning and change*, 33–49. Dordrecht: Springer Netherlands. DOI: [10.1007/978-94-007-4774-6_3](https://doi.org/10.1007/978-94-007-4774-6_3).
- Kosmaczewska, Kasia & Matt Train. 2019. Application of post-edited Machine translation in fashion eCommerce. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 167–173. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).
- LeBlanc, Matthieu. 2014. Les mémoires de traduction et le rapport au texte: Ce qu'en disent les traducteurs professionnels. *TTR: traduction, terminologie, rédaction* 27(2). 123–148. DOI: [10.7202/1037748ar](https://doi.org/10.7202/1037748ar).

Maeve Olohan

- Macken, Lieve, Daniel Prou & Arda Tezcan. 2020. Quantifying the effect of Machine translation in a high-quality human translation production process. *Informatics* 7(2). 1–19. DOI: [10.3390/informatics7020012](https://doi.org/10.3390/informatics7020012).
- Moorkens, Joss & Sharon O’Brien. 2017. Assessing user interface needs of post-editors of machine translation. In Dorothy Kenny (ed.), *Human issues in translation technology*, 109–130. London & New York: Routledge.
- Moorkens, Joss, Sharon O’Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3). 267–284. DOI: [10.1007/s10590-015-9175-2](https://doi.org/10.1007/s10590-015-9175-2).
- Nicolini, Davide. 2009a. Articulating practice through the interview to the double. *Management Learning* 40(2). 195–212. DOI: [10.1177/1350507608101230](https://doi.org/10.1177/1350507608101230).
- Nicolini, Davide. 2009b. Zooming in and out: Studying practices by switching theoretical lenses and trailing connections. *Organization Studies* 30(12). 1391–1418. DOI: [10.1177/0170840609349875](https://doi.org/10.1177/0170840609349875).
- Nicolini, Davide. 2011. Practice as the site of knowing: Insights from the field of telemedicine. *Organization Science* 22(3). 602–620. DOI: [10.1287/orsc.1100.0556](https://doi.org/10.1287/orsc.1100.0556).
- Nicolini, Davide. 2012. *Practice theory, work, and organization: An introduction*. Oxford: Oxford University Press.
- Nitzke, Jean, Silvia Hansen-Schirra & Carmen Canfora. 2019. Risk management and post-editing competence. *The Journal of Specialised Translation* 31. 239–259.
- Nunziatini, Mara. 2019. Machine translation in the financial services industry: A case study. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 57–63. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).
- Olohan, Maeve. 2017. Knowing in translation practice: A practice-theoretical perspective. *Translation Spaces* 6(1). 160–181. DOI: [10.1075/ts.6.1.08olo](https://doi.org/10.1075/ts.6.1.08olo).
- Olohan, Maeve. 2021. *Translation and practice theory*. London & New York: Routledge.
- Pielmeier, Hélène & Paul O’Meara. 2020. *The state of the linguist supply chain*. Cambridge, MA: CSA Research.
- Postill, John. 2010. Introduction: Theorising media and practice. In Birgit Bräuchler & John Postill (eds.), *Theorising media and practice*, 1–32. New York & Oxford: Berghahn Books.
- Premoli, Valeria, Elena Murgolo & Diego Cresceri. 2019. MTPE in patents: A successful business story. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 36–41. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).

- Pym, Anthony. 2003. Redefining translation competence in an electronic age: In defence of a minimalist approach. *Meta* 48(4). 481–497.
- Reckwitz, Andreas. 2002. Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory* 5(2). 243–263. DOI: [10.1177/1368431022225432](https://doi.org/10.1177/1368431022225432).
- Schatzki, Theodore R. 1996. *Social practices: A Wittgensteinian approach to human activity and the social*. Cambridge: Cambridge University Press.
- Schatzki, Theodore R. 2001. Introduction: Practice theory. In Theodore R. Schatzki, Karin Knorr-Cetina & Eike von Savigny (eds.), *The practice turn in contemporary theory*, 1–14. London & New York: Routledge.
- Schatzki, Theodore R. 2002. *The site of the social: A philosophical account of the constitution of social life and change*. University Park, PA: Pennsylvania State University Press.
- Schatzki, Theodore R. 2012. A primer on practices. In Joy Higgs, Ronald Barnett, Stephen Billett, Maggie Hutchings & Franziska Trede (eds.), *Practice-based education: Perspectives and strategies*, 13–26. Rotterdam: SensePublishers. DOI: [10.1007/978-94-6209-128-3_2](https://doi.org/10.1007/978-94-6209-128-3_2).
- Schatzki, Theodore R. 2016. Practice theory as flat ontology. In Gert Spaargaren, Don Weenink & Machiel Lamers (eds.), *Practice theory and research: Exploring the dynamics of social life*, 28–42. London & New York: Routledge.
- Schatzki, Theodore R. 2019. *Social change in a material world*. London & New York: Routledge.
- Schmidtke, Dag & Declan Groves. 2019. Automatic translation for software with safe velocity. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 159–166. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).
- SDL. 2020. *Post-Editing Machine Translation Training*. <https://www.sdltrados.com/learning/training/post-editing-machine-translation.html> (7 July, 2020).
- Shove, Elizabeth. 2017. Matters of practice. In Allison Hui, Theodore Schatzki & Elizabeth Shove (eds.), *The nexus of practices: Connections, constellations, practitioners*, 155–168. London & New York: Routledge.
- Shove, Elizabeth, Mika Pantzar & Matt Watson. 2012. *The dynamics of social practice*. London: SAGE.
- Shove, Elizabeth, Matt Watson & Nicola Spurling. 2015. Conceptualizing connections: Energy demand, infrastructures and social practices. *European Journal of Social Theory* 18(3). 274–287. DOI: [10.1177/1368431015579964](https://doi.org/10.1177/1368431015579964).
- Silva, Catarina. 2019. Improving domain adaptation for Machine translation with translation pieces. In *Machine translation summit XVII. Volume 2: Translator,*

Maeve Olohan

- project and user tracks, 204–212. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).
- Slator. 2019. *Neural machine translation report: Deploying NMT in operations*. Slator.
- Spaargaren, Gert, Don Weenink & Machiel Lamers (eds.). 2016. *Practice theory and research: Exploring the dynamics of social life*. London & New York: Routledge.
- Stahlberg, Felix. 2019. Neural Machine translation: A review. *arXiv – Computation and Language* 1912.02047. <http://arxiv.org/abs/1912.02047v1> (8 July, 2020).
- TAUS. 2015. *MT post-editing guidelines*. <https://www.taus.net/think-tank/reports/postedit-reports/taus-post-editing-guidelines> (25 August, 2019).
- Teixeira, Carlos S. C. 2014. *The impact of metadata on translator performance: How translators work with translation memories and machine translation*. Tarragona: Universitat Rovira i Virgili. (Doctoral dissertation).
- Vardaro, Jennifer, Moritz Schaeffer & Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* 6(3). 41. DOI: [10.3390/informatics6030041](https://doi.org/10.3390/informatics6030041). <https://www.mdpi.com/2227-9709/6/3/41> (20 March, 2020).
- Warde, Alan. 2016. *The practice of eating*. Cambridge: Polity Press.
- Warde, Alan, Shu-Li Cheng, Wendy Olsen & Dale Southerton. 2007. Changes in the practice of eating: A comparative analysis of time-use. *Acta Sociologica* 50(4). 363–385. DOI: [10.1177/0001699307083978](https://doi.org/10.1177/0001699307083978).
- Welch, Daniel & Alan Warde. 2017. How should we understand “General Understandings”? In Allison Hui, Theodore Schatzki & Elizabeth Shove (eds.), *The nexus of practices: Connections, constellations, practitioners*, 183–196. London & New York: Routledge.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. Google’s neural Machine translation system: Bridging the gap between human and machine translation. *arXiv – Computation and Language* 1609.0814. <http://arxiv.org/abs/1609.0814v2> (10 August, 2019).
- Zaretskaya, Anna. 2019a. Optimising the Machine translation post-editing workflow. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 136–139. Varna, Bulgaria: Incoma Ltd. DOI: [10.26615/issn.2683-0078.2019_018](https://doi.org/10.26615/issn.2683-0078.2019_018).

Zaretskaya, Anna. 2019b. Raising the TM threshold in neural MT post-editing: A case study on two datasets. In *Machine translation summit XVII. Volume 2: Translator, project and user tracks*, 213–218. Dublin. <https://www.mtsummit2019.com/proceedings> (7 July, 2020).

Chapter 2

Testing the Gravitational Pull Hypothesis on modal verbs expressing obligation and necessity in Catalan through the COVALT corpus

Josep Marco

Universitat Jaume I

This chapter aims to test Halverson's Gravitational Pull Hypothesis (GPH) on the Catalan modal verb *caldre*. Data from two comparable parallel sub-corpora from the COVALT corpus (English-Catalan and French-Catalan) are used to that end. However, *caldre* only serves as a starting point for hypothesis formulation and data analysis, as it is part of the wider network of modal verbs and verbal periphrases conveying obligation or necessity in Catalan. On the basis of networks of obligation and necessity in Catalan, English and French, three hypotheses are put forward: 1) *caldre* will be under-represented in the English-Catalan subcorpus when compared to Catalan non-translations; 2) *caldre* will be either over- or under-represented in the French-Catalan sub-corpus when compared to Catalan non-translations; and 3) *caldre* will be significantly more frequent in the French-Catalan than in the English-Catalan sub-corpus. Results confirm hypotheses 1 and 3, and for hypothesis 2 the scales are tipped in favour of over-representation, mainly as a result of strong connectivity between *caldre* and its French source text triggers. Connectivity, in the present study, seems to be favoured by formal similarity, which comes in two forms: syntactic isomorphism and phonological/graphological similarity.

1 Introduction

The main aim of this chapter is to test out Halverson's Gravitational Pull Hypothesis (2003; 2010; 2017) on the Catalan modal verb *caldre* in two compara-



Josep Marco

ble parallel sub-corpora from the COVALT corpus: English-Catalan and French-Catalan. The Gravitational Pull Hypothesis (GPH) was first put forward as an attempt to account for different hypotheses about translated language by anchoring them in cognitive linguistics and bilingualism. The Catalan modal verb *caldre* is arguably a suitable testing ground for the hypothesis because it may be considered a unique item (UI) in the English-Catalan language pair but not in the French-Catalan. Focusing on an item that fulfils that condition in that kind of environment (two comparable parallel corpora) is one of the methods suggested (Hareide 2017) to put the GPH to the test. However, *caldre* will only serve as a starting point in hypothesis formulation and corpus data analysis. It is part of the wider network of modal verbs and verbal periphrases conveying obligation or necessity in Catalan; therefore, other nodes in that network in the languages involved will be brought under scrutiny.

The outline of the study is as follows. §2 will present the main tenets of the GPH. §3 will provide an account of the main modal verbs and periphrases used to convey obligation and necessity in Catalan, English and French. §4 will spell out the aims of the study, the particular hypotheses to be tested on corpus data, together with the assumptions on which they are based, and the different steps into which the method followed can be broken. §5 will present data from the two parallel sub-corpora and the component of Catalan non-translations, which will be analysed and discussed. Finally, §6 will offer some conclusions.

2 The Gravitational Pull Hypothesis

Halverson's Gravitational Pull Hypothesis (GPH) aims to account for different hypotheses about translated language by anchoring them in cognitive linguistics. As is well-known, Baker (e.g. 1993) postulated a number of so-called *translation universals*, or features of translated language purported to occur independently of the language pair involved and the particular factors defining the translation situation (brief, genre, register, gender, ideology, etc.). One of these features allegedly was over-representation in translated text of typical target language (TL) elements. This claim may be said to be akin to Toury's law of growing standardisation, which states that "in translation, source-text textemes tend to be converted into target-language (or target-culture) repertoremes" (Toury 1995: 267–268) – textemes being special (perhaps unique) textual relations and repertoremes, conventional elements of the target repertoire. On the other hand, Tirkkonen-Condit (2002; 2004) argued a few years later that typical TL elements or structures tended to be under-represented (not over-represented) in translated language

when their use was not triggered by a formal equivalent or counterpart in the source language (SL). These elements lacking correspondence in the other member of a particular language pair were called *unique items* – even if uniqueness in this case must be understood as a relative concept concerning two particular languages, not in the absolute sense of a given item existing only in one human language and being unheard-of in the rest. How can these two opposing claims be true? On the face of it, the truth of one would preclude the truth of the other. However, both have been attested and are supported by (at least partial) evidence.

In this context, the main tenet of Halverson's GPH is that features of translated language (including over- and under-representation of TL typical items) can be explained on the basis of general characteristics of human cognition. Cognitive linguistic theory claims that linguistic units are integrated into higher-order structures, such as schematic networks, made up of nodes and links between nodes. Networks are characterised by asymmetry: some nodes are more salient than others. The most salient elements are usually the high-level schema (because of its high level of generality) and the prototype (understood as the best representative of a category). On the other hand, empirical research on bilingualism has identified two layers of representation in the bilingual brain: the lexical memory, where the forms of spoken and written lexical items are stored, and the conceptual level, where word meanings are stored. Links between words in different languages are set up through connections either at the conceptual or the lexical level, and such connections may rest upon total or partial overlap. These are then the two basic elements of human cognition which are brought to bear on translational behaviour: degree of cognitive salience (of particular nodes in a network) and degree of overlap between nodes and networks across languages, which will favour connectivity or otherwise. The theory merges concepts from cognitive grammar (especially [Langacker 1987](#)) and bilingualism ([De Groot 1992a,b](#)). The implications of all this for translation are spelt out by [Halverson \(2003: 218\)](#) as follows:

The basic idea is straightforward: in a translation task, a semantic network is activated by lexical and grammatical structures in the ST. Within this activated network, which also includes nodes for TL words and grammatical structures, highly salient structures will exert a gravitational pull, resulting in an overrepresentation in translation of the specific TL lexical and grammatical structures that correspond to those salient nodes and configurations in the schematic network.

In a later work ([Halverson 2010](#)), some problems with the 2003 original model were identified. The main one was theoretical: the GPH as initially formulated

Josep Marco

conflated two separate factors, content of the schematic network and specific pattern of connectivity, which need to be teased apart for a more valid explanation of translational effects. The model was accordingly revised to posit three possible cognitive causes of translational effects (Halverson 2010: 356; Hareide 2017: 192): patterns of prototypicality in the target language, conceptual structures or the representation of the source language item, and patterns of connectivity. One effect is predicted for each potential cause, or factor. The effect of factor 1 will be over-representation; the effect of factor 2 will be over-representation too; and the effect of factor 3 may be over- or under-representation. A further revision of the GPH model was proposed by Halverson in 2017. The term *gravitational pull* was now reserved for translational effects stemming from the representation of SL items (factor 2), i.e. for “the cognitive force that makes it difficult for the translator to escape from the cognitive pull of highly salient representational elements in the source language” (Halverson 2017: 14). Salience in the target language (factor 1) “may be more clearly captured by the metaphorical term *magnetism*” (2017: 14). And the third factor is called *connectivity* and defined as “the nature and strength of links between elements in a bilingual’s two languages” (2017: 14). As in earlier versions of the model, Halverson (2017: 28) stresses the fact that it is not possible at present to predict how salience patterns and connectivity interact, whether the different factors act independently or jointly, etc. Like any scientific theory, the GPH is open to refinement and modification on the basis of evidence.

Pending availability of more evidence, the choice of this hypothesis as a framework can only be justified on theoretical grounds. Firstly, it may be said to be a move away from binary formulations towards a more complex, multi-factorial analysis. Corpus-based studies of features of translated text typically set out to test a particular hypothesis (e.g. simplification, explicitation, normalisation) in isolation on a given set of data, so that the hypothesis is confirmed or refuted on a binary, yes/no basis. The GPH, in contrast, brings together several causes and attempts to find out the effects brought about by a particular configuration of such causes and the interaction between them. And secondly, it is rooted in cognition – cognition being one of the two possible causes of translation features proposed so far (Halverson 2017: 10): “there are two main approaches taken to the problem of explaining translational patterns. These two are socially and cognitively oriented, respectively”. Malmkjær (2008: 57) had gone further than that and suggested that the term *universal* (which is less and less used nowadays in Malmkjær’s absolute sense) be reserved for phenomena that can be cognitively explained. The reader is referred to Halverson (2017) for a more detailed illustration of the two approaches to the explanation of translation properties.

3 Modal verbs expressing obligation or necessity in Catalan, English and French

There are two major verbal constructions expressing obligation or necessity in Catalan: *haver de* + infinitive and *caldre*. Of course the language has many other resources to convey these meanings, but these two are fully grammaticalised – in fact, [Badia i Margarit \(1994: 611\)](#) refers to them as “grammatical formulas”. The former is a modal periphrasis and shares some features of syntactic behaviour with other modal periphrases. According to [Gavarró & Laca \(2002: 2714\)](#), it can convey both necessity (interpreted as epistemic) and obligation (interpreted as deontic). *Caldre* + infinitive (one of the possible constructions in which *caldre* can occur), on the other hand, is excluded by these authors (2002: 2710) from the list of Catalan modal periphrases on grounds of syntactic behaviour, even if earlier grammarians (cf. [Badia i Margarit 1994](#)) had treated it as such.

Be that as it may, there seems to be general agreement that *caldre* is a modal verb (see e.g. [Payrató 2002: 1192](#); [Rigau 2005](#)). Rigau sees *caldre* as belonging to the category of relative impersonal verbs, i.e. verbs used impersonally in that they refer to a person who does not feature in the sentence as agent but recipient ([Rigau 2005: 242](#)). This verb, together with similar ones belonging to the same category, follows the syntactic pattern of such Latin verbs as *licet* (‘it is licit/permitted’) or *oportet* (‘it is proper/necessary’, ‘it behoves’). These verbs were only used in the third person and took two kinds of complements: an infinitive or a subordinate clause introduced by *ut* (similar to a *that*-clause). Relative impersonal verbs exist in all Romance languages, even if their syntactic behaviour shows some variation. The list provided by [Rigau \(1999: 324\)](#) includes (relevantly to our purposes, as will be seen) French *falloir*. *Caldre* comes from Latin *calēre* (‘to be hot’) – hence the sense of urgency, of necessity. It has cognates in such neighbouring languages as Occitan and Aragonese, and it used to have them in medieval Spanish, Old French and Old Italian ([Rigau 1999: 331](#)) as well. Syntactically, it may take three kinds of complements: an infinitive, a subordinate clause introduced by *que* (i.e. a *that*-clause) and a noun phrase. It may also take a zero complement, with any of the three types of complements just mentioned left implicit. Examples (1–4) illustrate these four patterns, respectively.

- (1) Cal tenir molta força per moure aquesta taula.
‘It takes a lot of strength/a very strong person to move this table.’
- (2) No cal que t’amoïnes tant.
‘You needn’t worry so much.’; more literally, ‘it is not necessary that you worry so much.’

Josep Marco

- (3) Et caldrà molta paciència si vols convèncer-lo.
'It will take you a lot of patience if you want to persuade him.'
- (4) Pots fer servir el meu cotxe, si cal.
'You can use my car if necessary.'

To sum up, *haver de* + infinitive is a modal periphrasis that can convey both epistemic and deontic meaning. It can occur both in personal (as in 5) and impersonal (as in 6) constructions.

- (5) Ha de lliurar el CV si vol que l'entrevisten.
'He must/has to submit a CV if he wants to be interviewed.'
- (6) S'ha de tenir en compte que jo no hi era.
'It must be taken into account that I was not there.'

Caldre is a modal verb, but grammarians do not agree on the kind of modality it conveys, whether epistemic or deontic. It is almost exclusively used in impersonal constructions. It is here assumed, since there is no evidence to the contrary in the literature, that the meaning of *caldre* remains stable across the four constructions it occurs in. No meaning variation is observed depending on the kind of complement it takes.

As to modality in English, various accounts have been provided on the basis of widely differing theoretical assumptions. Cognitive accounts start from the basic epistemic vs. deontic distinction, even if Langacker (1991: 272) claims that this distinction "is not always easy to maintain", as most English modals can be used both epistemically and deontically. Langacker draws on Talmy and Sweetser to suggest that "the English modals are best analyzed in terms of force dynamics" (1991: 273). Force-dynamic values are applicable either to the domain of social interaction (deontic modality) or reasoning (epistemic modality). Radden & Dirven (2007) take a much more comprehensive view, which can only be briefly summarised here. These authors define modality as "an assessment of potentiality, depending either on the speaker's judgement of the reality status of a state of affairs (epistemic modality) or on the speaker's attitude towards the realisation of a desired or expected event (root modality)" (Radden & Dirven 2007: 246). Under root modality are subsumed three sub-types: deontic, intrinsic and disposition. Deontic modality is typically realised in two ways: obligation (e.g. *You must be home by nine*) and permission (e.g. *You may leave now if you wish*). Intrinsic modality is concerned with "intrinsic qualities of a thing or circumstances" (Radden & Dirven 2007: 246), as in *Pros and cons must be weighed up before a decision is made*. Finally, disposition modality encompasses the notions

of ability, propensity and willingness (2007: 246). Radden & Dirven (2007) make another distinction that cuts across the previous one, that between compelling and enabling modalities – the two main paths of grammaticalisation leading from lexical to deontic and epistemic meanings. Compelling modalities “involve a compelling force; they comprise obligations, prohibitions and intrinsic and epistemic necessities” (Radden & Dirven 2007: 247), whereas enabling modalities convey possibilities, abilities and permissions.

The focus of the present chapter is on the first group, once epistemic necessity has been removed, i.e. on obligation (whether positive or negative) and intrinsic necessity, because those are the senses present in the Catalan modals *caldre* and *haver de*. (Epistemic necessity is conveyed by a different modal verb, *deure*.) These modalities are expressed by “the central modals *must*, *need (to)* and *should* and the semi-modals *ought to*, *have to* and *have got to*” (Radden & Dirven 2007: 247). Differences among these verbs are set up on the basis of two criteria: the source of the compelling force and the degree of strength. The source of the compelling force may be the speaker (subjective) or external circumstances (external). And according to its degree of strength, the compelling force may be strong, neutral or weak. Obligation is always subjective, with *must* and *have got to* as strong and *should* and *ought to* as weak indicators of modality. Intrinsic necessity is external, with *have (got) to* and *must* as strong, *need to* as neutral and *should* and *ought to* as weak indicators of modality. The central (in the sense of most commonly used) modal verbs for these categories would be *must* for strong obligation, *have (got) to* for strong intrinsic necessity and *should* for weak obligation or intrinsic necessity. Radden & Dirven (2007: 249) further add that shifts in the system of compelling modals have occurred in American English due to democratisation and colloquialisation, to the extent that *must* has become much less common than *have (got) to*.

Catalan *caldre* conveys both obligation and intrinsic necessity. In French, these meanings are mainly conveyed by the verbs *devoir* and *falloir*. According to Lewis (2015: 159), “*devoir* is said to be more solemn or more insistent than *falloir*, while *falloir* is more often used in ‘subjective contexts’ where *devoir* might be interpreted as epistemic or as expressing futurity”. Both are polysemic in that they can express obligation, whether from an internal or an external source (i.e. the source of the obligation may be either the speaker or otherwise, respectively), and “non-deontic necessity” (what we have referred to here as intrinsic necessity). Moreover, *devoir* can also convey epistemic necessity (like English *must* or Catalan *deure*) and futurity. Beyond their semantic values, Lewis emphasises the syntactic differences between the two verbs (2015: 158–159): “deontic *devoir* typically takes a human subject while *falloir* can only be used with dummy subject

Josep Marco

il". In other words, while the former occurs in personal constructions, the latter is impersonal, like *caldre*. Furthermore, *devoir* is regarded as more formal than *falloir*.

4 Aim and methodology

As explained at the beginning, the main aim of this chapter is to test out the GPH on the Catalan modal verb *caldre* in two comparable parallel sub-corpora from the COVALT corpus: English-Catalan and French-Catalan. The corpus used will be both parallel and comparable, as data will also be retrieved from a component of Catalan non-translations. *Caldre* is used as a starting-point for the analysis because it may be said to be a unique item for the English-Catalan but not for the French-Catalan language pair. English does not have a syntactic counterpart for *caldre*, as possible candidates, such as the expression *it + take + X* (as in *It takes a lot of courage to rise to that challenge*), are not frequent or grammaticalised to the same extent as *caldre* is.¹ French, on the other hand, has the verb *falloir*, as seen above, which also conveys obligation or necessity and typically occurs in impersonal constructions. As seen in §3, both *caldre* and *falloir* fall under the category of relative impersonal verbs, which cuts across all Romance languages, and take the same kind of complements to a large extent – the only difference being that *falloir* cannot take a noun phrase as a complement. They share the other two complements (infinitive and *that*-clause), and that is the basis of their syntactic similarity. Formal similarity comes under many guises. The most obvious one is phonological or graphological similarity, especially when it concerns two words with a common origin, e.g. English *hound* and German *Hund*. But there may be formal similarity at other levels, such as that of syntax. Catalan *caldre* and French *falloir* are not cognates, but they share two syntactic patterns in addition to their semantic common ground. The rationale behind using an item that is unique for a certain language pair but not for another is the same as in Hareide (2017), which serves here as a source of methodological inspiration: the basic contrastive fact around which the study pivots may well give rise to

¹The query [lemma="it"][lemma="take"] in the ST component of the English-Catalan sub-corpus in COVALT yields 44 matches, 15 of which are false positives, the remaining 29 often featuring a time complement, as in *It took three days to...* The normalised frequency (f) of this construction is 0.024 per 1,000 words. Just for the sake of comparison, the normalised frequency of *must* as an indicator of obligation and intrinsic necessity is 0.55 – over 20 times as high as the frequency of *it + take*. (This value is based on a projection of the results yielded by the manual analysis of a random sample of 300 instances, out of the total 993 matches found for the query [lemma="must"].)

different configurations of factors related to salience and connectivity that may impact translation outcomes.

One of the pre-requisites for this methodology is having two parallel corpora that can be regarded as comparable in all relevant respects, i.e. textual genre, date and place of publication, and type of readership. The COVALT corpus fulfils such a requirement. COVALT (Valencian Corpus of Translated Literature) is a multilingual corpus made up of the translations into Catalan of narrative works originally written in English, French, and German published in the autonomous region of Valencia from 1990 to 2000, together with their corresponding source texts. The English-Catalan sub-corpus comprises 36 English source texts, amounting to 1,201,757 words, and their corresponding target texts in Catalan (1,343,631 words). The French-Catalan sub-corpus comprises 21 French source texts, amounting to 551,869 words, and their corresponding target texts (566,998 words). COVALT also includes non-translated components for both target languages, Catalan and Spanish. The Catalan non-translated component is a set of narrative works originally written in Catalan intended to be comparable to the translated component in all relevant respects: place of publication (Valencian Community), date of publication (1990-2000), language (Catalan) and genre (narrative fiction). The non-translated component amounts to 1,551,521 tokens. These corpora were compiled at the Translation and Communication Department, Universitat Jaume I (Castelló, Spain) and can be accessed for research purposes upon request (<http://www.covalt.uji.es>).

Before formulating hypotheses, we need at least some basic information on the relative salience of the main verbal indicators of obligation and intrinsic necessity in the three languages involved in this study. Since salience is operationalised as frequency (as will be seen later on), corpus data will be used when suitable. Grammars tell us that both *caldre* and *haver de* + infinitive are central as regards the expression of obligation and necessity in Catalan, and the same applies to *must* and *have (got) to* for English; but which member of the pair is the more frequent? In the component of Catalan non-translations in COVALT, *caldre* occurs 735 times, with a normalised frequency of 0.47 per 1,000 words, whereas *haver de* + infinitive features 1,924 occurrences, with a normalised frequency of 1.24 per 1,000 words. *Haver de* + infinitive is about 2.5 times as frequent as *caldre* (in terms of normalised frequency) and we may assume, therefore, that it is more salient.

In the case of English, it would not make much sense to compare corpus frequencies because the meanings of *must* and *have to* do not overlap to such an extent as *caldre* and *haver de* in Catalan. As seen in §3, *must* is the central verb for strong obligation and *have (got) to* for strong intrinsic necessity (Radden &

Josep Marco

Dirven 2007). We can rely on these assumptions in order to formulate hypotheses. It must also be borne in mind that absence of obligation is usually conveyed by the negative forms of *have to* and *need (to)* – not by the negative form of *must*, which expresses negative obligation, i.e. prohibition.

As to French, we saw above that the main modal or semi-modal verbs conveying obligation and intrinsic necessity are *devoir* and *falloir*. As in the case of English, it would not make much sense here to compare corpus frequencies of these two verbs (e.g. in the ST component of the French-Catalan sub-corpus in COVALT) because their meanings overlap only to a certain extent. *Devoir* is more polysemous than *falloir*, as it also conveys epistemic necessity and futurity, and it would be necessary to discard these meanings manually. Lewis (2015) reports on a previous study by Labbé & Labbé (2013) according to which *falloir* is much more frequent than *devoir* in spoken and literary French, whereas the opposite is true for a corpus of presidential speeches. Lewis claims that her own results from a corpus-based analysis of political speeches in English and French are consistent with Labbé and Labbé's findings.

We may therefore assume that: a) *haver de* + infinitive is a more salient indicator of obligation and intrinsic necessity than *caldre* in Catalan; b) the main *prima facie* equivalents of *caldre* (and *haver de*) in English (*must* and *have to*) and French (*falloir*) are also salient in their respective modality networks; and c) patterns of connectivity between *caldre* and those *prima facie* equivalents will be stronger for French than for English. The first two assumptions were justified in the previous paragraph. The third assumption is based on the formal similarity between *caldre* and *falloir*, i.e. on their syntactic overlap (explained above), which is not paralleled by *caldre* and any of its English equivalents. On the basis of these assumptions, the following three hypotheses can be formulated:

1. *caldre* will be under-represented in the English-Catalan subcorpus when compared to Catalan non-translations, as neither factor 1 (magnetism) nor factor 3 (high degree of connectivity) will be at play – factor 2 (gravitational pull) being the only factor that might pull towards over-representation;
2. *caldre* will be either over- or under-represented in the French-Catalan subcorpus when compared to Catalan non-translations, depending on which factor prevails (gravitational pull and a high degree of connectivity will pull towards over-representation whereas magnetism will pull towards under-representation);
3. *caldre* will be significantly more frequent in the French-Catalan than in the English-Catalan sub-corpus, as over-representation will be favoured

by two factors (gravitational pull and a high degree of connectivity) in the former and only one (gravitational pull) in the latter.

The method employed to verify these hypotheses will consist of the following steps:

1. data retrieval with CQP (Corpus Query Processor), a tool that allows to query corpora on the basis of regular expressions containing words, lemmas and part-of-speech tags. Both the translated components of the English- and French-Catalan sub-corpora, and the Catalan non-translated component will be queried on the lemma *caldre*;
2. manual sifting in order to tell apart true from false positives. Corpus queries usually yield matches that do not conform to the criteria the analyst had in mind. If false positives are not removed, the data on which quantification draws will be distorted;
3. quantification + testing for significance. Raw and relative frequencies of *caldre* in the three components mentioned in step 1 will be established and tested for significance;
4. searching for triggers (i.e. ST segments matching the query word) of *caldre* in the English and French STs;
5. searching for TT segments matching the main triggers of *caldre*. Query matches will be thinned if their number proves unmanageable. Thinning is the standard method used by CQP for random sampling, and it can be based on a raw figure or a percentage;
6. manual sifting (again), in order to tell apart true from false positives;
7. establishing degrees of connectivity between ST and TT items. The measure to be used for that purpose will be introduced below;
8. repeating the whole process for *haver de* (the main alternative to *caldre* in Catalan, as seen above) in the English-Catalan and French-Catalan sub-corpora and the Catalan non-translated component.

As these steps suggest, for the big picture to emerge as regards patterns of salience and connectivity in the two language pairs it is necessary to go beyond the initial pivot of the study (*caldre*) and look at the main nodes in the monolingual and

Josep Marco

bilingual networks of which *caldre* is a part. This kind of analysis is extremely time-consuming. The *big* picture may not be the *full* picture, but it is hoped it will include enough relevant information not only to test the hypotheses but also to understand why they are confirmed or refuted.

Before moving on to results and discussion, the thorny question of the relationship between frequency, on the one hand, and salience and connectivity, on the other, must be addressed. Schmid (2010) poses the question in the most explicit possible manner when he wonders whether frequency in text instantiates entrenchment in the cognitive system. Entrenchment is defined as “the degree to which the formation and activation of a cognitive unit is routinized and automated” (Schmid 2010: 115). It is fostered by repetitions of cognitive events. Schmid (2010: 116) refers to the “considerable body of evidence from psycholinguistic experiments suggesting that frequency is one major determinant of the ease and speed of lexical access and retrieval”, and goes on to argue that, since speed of access and retrieval correlates with routinisation, “this indeed supports the idea that frequency and entrenchment co-vary” (Schmid 2010: 116). But this is not as straightforward as it seems.

Drawing on previous authors, Schmid (2010: 116) claims that “it is not frequency of use as such that determines entrenchment, but frequency of use with regard to a specific meaning or function, in comparison with alternative expressions of that meaning or function”. The former type of frequency is called *absolute* and the latter *relative*. Schmid observes that, even though the correlation between frequency and cognitive significance is far from unproblematic, cognitively-oriented corpus linguists “try to correlate the frequency of occurrence of linguistic phenomena (as observed in corpora) with their salience or entrenchment in the cognitive system” (Schmid 2010: 101). Indeed, it seems difficult to proceed otherwise. Schmid’s caveats are very much in place in methodological terms, but he provides no alternative to frequency as an operationalisation of salience and entrenchment, as no direct access to the cognitive system seems to be available at present. Halverson advocates a mixed-methods approach with different types of data (elicitation data and analysis of keystroke logs) in addition to corpus data, but, regardless of the type of data under scrutiny, both salience and entrenchment are operationalised as frequency. The same procedure will be followed here, even though most analyses (as in Halverson 2017) will be based on relative rather than absolute frequency.

5 Results and discussion

The lemma *caldre* was inserted in the query box of CQPweb for the three relevant sub-corpora: English-Catalan (EN-CAT), French-Catalan (FR-CAT) and Catalan non-translations (NTR). Query matches were manually checked and the number of false positives found to be rather low: 7 (out of 386 hits) for English-Catalan, 9 (out of 524 hits) for French-Catalan and 50 (out of 785 hits) for Catalan non-translations. All false positives are related to the contraction *cal(s)*, meaning ‘at somebody’s (house)’ and the adjective *calent*, meaning ‘hot’. Once these unwanted matches have been removed, results are as shown in Table 2.1.

Table 2.1: Query results for *caldre* in English-Catalan, French-Catalan and Catalan non-translations (f = normalised frequency per 1,000 words)

	n (words)	Query matches	f
Translations from English	1,343,631	379	0.28
Translations from French	566,998	515	0.91
Catalan non-translations	1,551,521	735	0.47

The figures for normalised frequency per 1,000 words strongly hint at significant differences across corpora. The log-likelihood (LL) test was applied to each pair of corpora and the differences turned out to be extremely significant in all three cases, with LL values at 70.33 for EN-CAT/NTR, 121.59 for FR-CAT/NTR and 299.56 for EN-CAT/FR-CAT.² The implications of these results for the three hypotheses formulated in the previous sections can be spelt out as follows:

1. *caldre* is under-represented in English-Catalan translations, when compared to Catalan non-translations;
2. *caldre* is over-represented in French-Catalan translations, when compared to Catalan non-translations, in accordance with one of the two possibilities foreseen in hypothesis 2;
3. *caldre* is significantly more frequent in French-Catalan translations than in English-Catalan translations.

²The critical value of the log-likelihood test is 3.84 for a 95% level of confidence (i.e. for a p value of <0.05) and 6.63 for a 99% level of confidence (p<0.01). Therefore, any LL value lower than 3.84 indicates that differences do not reach the threshold of statistical significance.

Josep Marco

Thus, hypotheses 1 and 3 are confirmed, and for hypothesis 2 the scales are tipped in favour of over-representation, which suggests that gravitational pull and a high degree of connectivity between *caldre* and its French triggers prevail over the relatively low magnetism posited for this verb in the Catalan modality network (in comparison with *haver de* + infinitive).

But at this point we know nothing yet about connectivity patterns between *caldre* and its triggers, as we have only looked at the translated component of the parallel corpora, not at the bilingual concordances. Bilingual concordance analysis for each parallel corpus is expected to provide: a) a list of ST triggers for *caldre*; b) the source concentration for those triggers; c) starting from the ST pole, a list of matching TT segments for the main triggers of *caldre*; d) the target concentration of those TT segments; e) a quantitative measure of the degree of connectivity between *caldre* and its ST triggers, based on source and target concentration. Let us see how this unfolds step by step, first for EN-CAT and then for FR-CAT. But before looking at results we need to dwell on the concepts of source and target concentration.

Schmid (2010) put forward two statistical measures to gauge the interaction between nouns and different kinds of shell-content constructions. One of these measures was the so-called *attraction-reliance method*. If we take, for instance, the construction Noun + *that* + clause, we may be interested in calculating the strength of the relationship between the noun *fact* and that construction (i.e. *the fact that...*). The attraction-reliance method allows us to do just that by calculating first the frequency of *fact* in that construction in proportion to the total frequency of the construction (attraction) and then the frequency of *fact* in that construction in proportion to the total number of occurrences of the noun in the corpus (Schmid 2010: 107). The attraction-reliance method “captures to some extent the intuition that some nouns are more important for certain constructions than others, and that some constructions are more important for certain nouns than others” (Schmid 2010: 111). Halverson (2017: 30ff) draws on Schmid’s method to introduce two statistical measures intended to gauge the strength of translation relationships between items in a parallel corpus: source concentration and target concentration. Source concentration is “the percentage of all occurrences of a TL item that are translations of a specific SL item” (Halverson 2017: 30), whereas target concentration is “the percentage of a set of translations of an SL item that is comprised by a given TL item” (Halverson 2017: 30). Both measures are expressed as percentages. There is no need to provide examples here as plenty of them will come up in what follows.

Table 2.2 shows the ST triggers for *caldre* in EN-CAT both in terms of raw frequency and source concentration. Since the list of trigger types was rather

long, triggers with fewer than 10 occurrences were grouped under “Other” for the sake of convenience. That is why this category yields such a comparatively large figure. It includes such heterogeneous triggers as imperatives, *ought to* + infinitive, *-ly* adverbs, *require/be required*, *it + take*, *want*, *have got to* + infinitive, and several others. \emptyset accounts for triggers with no overt expression of obligation or necessity. The figures for source concentration are relatively low in all cases, which means that no single ST trigger is responsible for the activation of a large percentage of occurrences of *caldre*. The three triggers with source concentration values higher than 10% (apart from \emptyset and “Other”) are *need*, *have to* + infinitive and *must* + infinitive, and they range from 11.1% to 15.30%. This suggests low connectivity, as assumed at the stage of hypothesis formulation, but only from the perspective of source concentration. We need to look at the main triggers for *caldre* in order to have the full picture of connectivity patterns.

Table 2.2: ST triggers for *caldre* in EN-CAT (n = raw frequency, s.conc = source concentration)

	n	s.conc
need	58	15.30
\emptyset	56	14.78
have to + inf	47	12.40
must + inf	42	11.1
infinitive	32	8.44
other solutions with <i>need</i>	22	5.80
be/become necessary	16	4.22
should + inf	10	2.64
other	94	24.80
misalignments	2	0.52
Total	379	100

Table 2.3 shows the TT matching segments of the three main ST triggers for *caldre* (*need*, *have to* + infinitive and *must* + infinitive) in EN-CAT both in terms of raw frequency (n) and target concentration (t.conc). When the number of hits for these three triggers was deemed manageable, all results were manually analysed, as in the case of *need*; when the number was deemed too high for manual analysis, results were thinned, as in the cases of *have to* + infinitive and *must* + infinitive.

The query for *need* (as a verb) yielded 227 matches, with *necessitar* (‘need’) as the top-ranking match with a high target concentration (48.02%). *Caldre* comes

Josep Marco

Table 2.3: TT matches for *need*, *have to* and *must* in EN-CAT (n = raw frequency, t.conc = target concentration)

	need		have to		must	
	n	t.conc	n	t.conc	n	t.conc
caldre	56	24.67	16	6.67	15	7.46
necessitar	109	48.02	–	–	–	–
fer falta	21	9.25	–	–	–	–
haver de	14	6.17	166	69.17	142	70.65
no modality	–	–	28	11.67	14	6.97
other	22	9.69	26	10.83	26	12.93
misalignments	5	2.20	4	1.66	4	1.99
Total	227	100	240	100	201	100

second with a target concentration of 24.67%. The query for *have to* + infinitive yielded 523 matches, which were thinned to 250. These 250 were manually sifted and 10 of them were seen to convey meanings other than obligation or intrinsic necessity and consequently removed. Analysis of the remaining 240 hits shows that the top-ranking TT match for *have to* + infinitive is by far the modal periphrasis *haver de* + infinitive, with a high target concentration of 69.17%. That means that *have to* + infinitive is translated as *haver de* + infinitive in over two thirds of the cases. *Caldre* is a poor match for *haver de* + infinitive, with a target concentration of just 6.67%. A similar picture emerges for *must* + infinitive. This query yielded 993 results, which were thinned to 300. These were again manually sifted and 99 of them were discarded because they were instances of *must* conveying strong possibility (i.e. epistemic modality), not obligation or intrinsic necessity. Manual analysis of the remaining 201 instances shows *haver de* + infinitive as the top-ranking match for *must* + infinitive, with a high target concentration value of 70.65%, with *caldre* again a poor second with a target concentration of merely 7.46%.

To sum up, the source concentration of English ST triggers for *caldre* is never too high (15.30 for *need*, 12.40 for *have to* + infinitive, 11.10 for *must* + infinitive), and nor is the target concentration of *caldre* as a Catalan TT match for its English triggers (24.67 for *need*, 6.67 for *have to* + infinitive, 7.46 for *must* + infinitive). But how can these two measures, source and target concentration, be brought together under a single formula that operationalises degree of connectivity, or strength of translation relationships, between items across the two components

of a parallel corpus? Both Schmid's attraction-reliance method and Halverson's adaptation in the form of source and target concentration are conceived as measures offering complementary views on connections between two items, but no suggestions for combining these measures are offered. A possible way of bringing them together is through an adaptation of Altenberg's (1999) concept of Mutual Correspondence.³ The concept is intended to measure the strength of the translation relationship between an item A in a given language and an item B in a different language in a parallel bi-directional corpus. It is defined as "the frequency with which different (grammatical, semantic and lexical) expressions are translated into each other" and formulated as follows:

$$\frac{(A_t + B_t) \times 100}{(A_s + B_s)}$$

where A_t and B_t = the number of times the compared items (A and B) are translated into each other, and $A_s + B_s$ = the total number of occurrences of the compared items in the source texts. Since the situation is different here, as the corpus we are using is parallel but not bi-directional, the formula is adapted as follows:

$$\frac{(A_b + B_a) \times 100}{(A_t + B_s)}$$

where A_b and B_a = the number of times A is the translation of B and B is translated as A (it will be the same figure, of course), and $A_t + B_s$ = the total number of occurrences of A in TT and of B in ST. Moreover, a different name needs to be found, as using the term *mutual* for a translation relationship that is not bi-directional may be misleading.⁴ I suggest the alternative term *Unidirectional Translation Correspondence* (UTC), which has the twofold advantage of drawing a parallel with Altenberg's term through the preservation of *correspondence* and explicating the unidirectional nature of the translation relationship.⁵ Let us take

³I would like to thank Sandra Halverson (personal communication) for suggesting this option.

⁴I am indebted to Sandra Halverson (personal communication) for this suggestion.

⁵A different possibility might have been the use of Dyvik's (e.g. 2002) *semantic mirrors method*, which allows the analyst to establish translation correspondences across languages by generating "images" of one word in the other language and then proceeding the other way around with a view to setting up (partly overlapping) semantic fields in both languages. Vandevoorde (2020) put the method to good use with the help of sophisticated statistics-based visual representations. But I can see two reasons for not using it in my research. Firstly, Dyvik's method is intended for use with bi-directional corpora, whereas mine are unidirectional. And secondly, while Vandevoorde aims at the visual representation of semantic fields (more particularly, the field of inchoativity in Dutch translated and non-translated language), my aim is to test a hypothesis on a particular modal indicator. True, in order to do that I need to look at other items

Josep Marco

the pair *caldre/need* as an example. If *need* is translated as *caldre* 58 times, since *caldre* occurs 379 times in the Catalan TTs and *need* 227 in the English STs, the UTC of *caldre* and *need* in the English-Catalan sub-corpus will be as follows: $(58+58) \times 100 / (379+227)=19,14\%$. The results of applying the same formula to the other two pairs are 10.66% for *caldre/have to* and 8.04% for *caldre/must*.⁶

These figures clearly suggest that the degree of connectivity (operationalised as UTC) between the Catalan modal verb *caldre* and its three main ST triggers in the English-Catalan sub-corpus of COVALT is rather low, which (together with the relatively low magnetism of *caldre*) accounts for its under-representation. The results in Table 2.3 also suggest that *haver de* + infinitive, which shows a high target concentration as a TT segment matching *have to* + infinitive and *must* + infinitive, is likely to display a high degree of connectivity with those two triggers. At the stage of network modelling prior to hypothesis formulation in §4 it was established that *haver de* + infinitive is about 2.5 times as frequent as *caldre* in Catalan non-translations, which suggests that the former is more salient than the latter as an indicator of obligation and necessity. For a full comparison between the two, we now need to look at the source concentration of the main triggers of *haver de* + infinitive with a view to determining the UTC of *haver de* + infinitive and each of these triggers. The whole process carried out for *caldre* must be repeated for *haver de*.

Table 2.4 shows the results for *haver de* + infinitive in the three sub-corpora. As seen above, *haver de* + infinitive is much more frequent than *caldre* in NTR (1.24 vs. 0.47 in normalised frequency per 1,000 words), and the same is valid for EN-CAT (1.55 vs. 0.28) and FR-CAT (1.04 vs. 0.91). Differences are huge indeed in the first two cases, but not so much in FR-CAT. However, when the log-likelihood test is applied, they turn out to be significant in all cases, with LL values at 1,303.55 for EN-CAT (extremely significant), 550.98 for NTR (extremely significant) and 5.36 for FR-CAT (significant at $p<0.05$).

in the network, especially as onomasiological salience can only be determined by comparing frequencies of synonyms and near-synonyms. But taking account of the whole semantic field of obligation/necessity in the three languages involved falls outside the scope of my study.

⁶On the basis of intuition alone I should have thought that there is no correlation between ST triggers and the four constructions *caldre* can occur in. However, this intuition needed to be confirmed by corpus data. A second manual analysis of the bilingual concordances for *caldre* shows that its distribution across types of construction is not symmetrical, as it occurs 223 times with an infinitive, 74 with a *that*-clause, 45 with a noun phrase and 37 with a zero complement. In relative terms, that amounts to 58.84%, 19.53%, 11.87% and 9.76%, respectively. If this analysis is replicated for each individual trigger (*need*, *must* + infinitive, *have to* + infinitive, etc.), frequency distributions do not exactly match the one just given, but differences are not marked enough to suggest a correlation between the two variables (type of trigger and type of construction *caldre* occurs in).

Table 2.4: Query results for *haver de* + infinitive in English-Catalan, French-Catalan and Catalan non-translations (f = normalised frequency per 1,000 words)

	n (words)	Query matches	f
Translations from English	1,343,631	2,088	1.55
Translations from French	566,998	592	1.04
Non-translations	1,551,521	1,924	1.24

As to degree of connectivity between *haver de* + infinitive and the ST triggers analysed above (*need*, *have to* + infinitive and *must* + infinitive), we already have data for queries in the English-to-Catalan direction. The next (and last) step will be to insert *haver de* + infinitive as query and to look at its ST triggers in order to determine their source concentration for the Catalan modal periphrasis. Table 2.5 offers such information. The 2,088 hits for *haver de* + infinitive were thinned to 250, four of which were manually discarded. On the basis of the remaining 246 matches, the source concentration of ST triggers of *haver de* + infinitive is found not to be very high in any case; that of *must* + infinitive is 20.73% and that of *have to* + infinitive is 17.48%. That means that the occurrence of *haver de* + infinitive in translations from English is not largely dependent on any particular trigger. But, as seen above, the target concentration of *haver de* + infinitive as a TT match for *have to* and *must* is very high. The Unidirectional Translation Correspondence value is 30.95% for *haver de/must* and 27.72% for *haver de/have to*, which is considerably higher than the UTC values for *caldre* and its main ST triggers. Therefore, the connectivity patterns of *haver de* + infinitive with its main ST triggers are stronger than those of *caldre* with its main triggers. That, together with its higher salience, makes *haver de* + infinitive a likelier match than *caldre* for English items conveying obligation or intrinsic necessity.

The data for *caldre* and its triggers retrieved from the French-Catalan corpus are much more straightforward. Hypothesis 2 predicted that *caldre* would be either over- or under-represented in FR-CAT as compared to NTR because gravitational pull and a high degree of connectivity would pull towards over-representation whereas magnetism would pull towards under-representation. Hypothesis 3 predicted that the frequency of occurrence of *caldre* in FR-CAT would be higher than in EN-CAT because over-representation would be favoured by two factors (gravitational pull and a high degree of connectivity) in the former and only one (gravitational pull) in the latter. Hypothesis 3 was confirmed, and for hypothesis 2 over-representation was the case, which suggests that gravi-

Josep Marco

Table 2.5: ST triggers for *haver de* in EN-CAT (n = raw frequency, s.conc = source concentration)

	n	s.conc
no modality	57	23.17
must	51	20.73
have to	43	17.48
should	28	11.38
be + inf	12	4.88
other	52	21.14
misalignments	3	1.22
Total	246	100

tational pull and a high degree of connectivity prevail over the relatively low magnetism of *caldre*.

Table 2.6 shows results for the ST triggers of *caldre* in FR-CAT both in terms of raw frequency and source concentration. The 515 hits for *caldre* in FR-CAT were thinned to 250 and manually analysed. The top-ranking trigger is by far the modal verb *falloir*, with a high source concentration of 68.4%. None of the remaining triggers individually reaches the value of 10%. That means that, when *caldre* occurs in FR-CAT, its occurrence is triggered by *falloir* in over two thirds of the cases. Data for the translation relationship between *falloir* and *caldre* from the source pole are shown in Table 2.7.

Table 2.6: ST triggers for *caldre* in FR-CAT (n = raw frequency, s.conc = source concentration)

	n	s.conc
falloir	171	68.4
no modality	17	6.8
devoir	12	4.8
other	46	18.4
misalignments	4	1.6
Total	250	100

The query for *falloir* yields 607 matches, which are thinned to 200 and manually sifted. Two are manually discarded and, for the remaining 198 instances,

Table 2.7: TT matches for *falloir* in FR-CAT (n = raw frequency, t.conc = target concentration)

	n	t.conc
caldre	117	59.1
haver de	20	10.10
no modality	11	5.55
other	34	17.17
unclear	10	5.05
misalignments	6	3.03
Total	198	100

caldre is by far the best represented Catalan match for *falloir*, with a high target concentration of 59.1%, with *haver de* + infinitive a poor second at 10.10%. This suggests strong translation links between *falloir* and *caldre* from both perspectives – a suggestion confirmed by their UTC, which stands at 62.74%. Connectivity patterns between *caldre* and its main French trigger, *falloir*, are very strong. That seems to be the main reason for over-representation of *caldre* in FR-CAT, together with salience of *falloir* in the French modal network for obligation and necessity, which was established on the basis of previous studies (Labbé & Labbé 2013; Lewis 2015).

For the analysis based on FR-CAT to be parallel in all respects to that based on EN-CAT, it would now be the time to look at the ST triggers of *haver de* + infinitive in FR-CAT. However, in EN-CAT that step was justified by the fact that *haver de* + infinitive was better represented as a target match for *have to* + infinitive and *must* + infinitive than *caldre*, whereas the case is otherwise for *falloir* in FR-CAT, with *caldre* as the top-ranking target match and *haver de* + infinitive with a relatively low target concentration of 10.10%. Therefore, it is not necessary to perform that query, which would probably show a higher source concentration of *devoir* (the other major verb conveying obligation and intrinsic necessity in French) than was the case with *caldre*.

It may be in place at this point to recapitulate the results of the corpus analysis reported on in this section. It was initially established that *caldre* is less salient in the TL than its main alternative in the obligation/intrinsic necessity network, *haver de* + infinitive. Therefore, magnetism can only be expected to play a minor role in the creation of translation effects. Even so, it is over-represented in FR-CAT as compared both to NTR and EN-CAT (hypotheses 2 and 3). This may

Josep Marco

be accounted for by strong connectivity between *caldre* and *falloir* (attested by the data) and, perhaps, by the gravitational pull of *falloir* (not tested for but reflected in the literature). On the other hand, connectivity between *caldre* and its main English ST triggers (*need*, *have to* + infinitive and *must* + infinitive) is low, which, added to the relatively low magnetism of *caldre*, results in its under-representation in EN-CAT. *Haver de* + infinitive, on the contrary, is over-represented in EN-CAT. This may be accounted for by the relatively high salience of *haver de* and the relatively strong connectivity between *haver de* and two of its ST triggers (*must* and *have to*).

6 Conclusions

The case of *caldre* shows that connectivity may tip the scales in favour of over- or under-representation. Through the use of two comparable parallel corpora with the same target language, TL salience is controlled for, as there is no reason to think that a certain TL item will be more salient in one corpus than in the other. Connectivity, in the present study, seems to be favoured by formal similarity, which comes in two forms. The first is syntactic isomorphism. In the French-Catalan combination, both *caldre* and *falloir* are mainly used in impersonal constructions that share two possible kinds of complements – infinitives and *that*-clauses. In the English-Catalan combination, *haver de* + infinitive can be used in both personal and impersonal constructions; and, whenever *must* + infinitive, *have to* + infinitive or *need* are used in a personal construction, there is a higher degree of overlap with *haver de* than with *caldre*.

The second factor is phonological/graphological similarity: *haver de* and *have to* display that kind of similarity, which would seem to foster connectivity at a very basic level. They are not cognates, as Latin *habere* and the Proto-Germanic root of English *have* are not etymologically related; but they could easily pass for cognates on the basis of phonological/graphological similarity. Cognate status is often deployed as an independent variable in psycholinguistic experiments on word translation. De Groot (1992b) is a case in point. This author sets out to measure translation performance (operationalised as reaction time, number of omissions and number of translation errors) under varying conditions. With regard to cognate status, her results lead her to conclude that “in addition to being translations, cognates have an extra reason to be linked in lexical memory. This could be reflected in relatively strong T1 links” – T1 links being links between lexical nodes at the level of lexical memory, without resorting to conceptual memory. Translation between cognates, then, would be favoured by strong connectivity

of a special kind; and there is no reason to suppose that this cannot hold true for *false* cognates too, since links between lexical nodes cannot be expected to reflect expert etymological knowledge.

The GPH is not incompatible with other models of the translation process. Carl et al. (2019) present a model based on the concept of *entropy*, borrowed from the fields of physics and information theory. Entropy describes “the amount of disorder in a system” (Carl et al. 2019: 217). In a context of translation, the more possible translations are activated in word and phrase translation systems, the higher the entropy. When entropy is high, the translator needs to invest a great deal of effort to find a solution. When a translator finds a complex word or structure for the first time, the information available is low and the degree of entropy at its highest, so much cognitive energy is required. Finding a satisfactory solution creates internal structure and reduces the degree of entropy, so less cognitive energy will need to be spent when the same word or structure recurs a second or third time. The process of entropy reduction over time is captured by the concept of *entropic gravity*. Entropy may arise from variability both in lexical and syntactic choices. In the model presented by these authors, activation of translation solutions in a system is non-selective for language, as elements are activated in the system on the basis of phonological and semantic associations in both languages. This initial stage is followed by a task-dependent decision process in which elements activated solely on the basis of phonological similarity, or belonging to the source language, are discarded and a satisfactory translation solution reached.

Carl et al. (2019: 226) claim that their model “relates to Halverson’s (2003) *gravitational pull hypothesis*”. However, they think it “unfortunate” (Carl et al. 2019: 227) that Halverson should have split her initial concept of gravitational pull into the three causes of translational effects mentioned above, among other reasons because “each of Halverson’s salience, link and connectivity effects might be more simply and coherently described in terms of entropic gravity, which assumes similar underlying mechanisms for producing the various translational effects” (Carl et al. 2019: 227). They further claim that there may be more than just three causes of translational effects (Carl et al. 2019: 227), although they do not mention any. Pending specification of such causes, it may be safe to stick to the three posited by Halverson. However, *factors* may be introduced that favour the activation of these causes. The research reported on in this paper suggests that syntactic isomorphism and phonological/graphological similarity strengthen connectivity – or, alternatively, entropic gravity by reducing the degree of entropy. This should not be seen as an attempt to alter the GPH in any

Josep Marco

fundamental way, but to refine it by introducing the notion of factor. Only further research will determine whether the attempt is worth pursuing or not.

Acknowledgements

This work was supported by Universitat Jaume I [UJI-B2017-58] and by the Spanish Ministry of Science and Innovation [PID2019-103953GB-I00].

References

- Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hilde Hasselgård & Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 249–268. Amsterdam: Rodopi.
- Badia i Margarit, Antoni. 1994. *Gramàtica de la llengua catalana: Descriptiva, normativa, diatòpica, diastràtica*. Barcelona: Enciclopèdia Catalana.
- Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.
- Carl, Michael, Andrew Tonge & Isabel Lacruz. 2019. A systems theory perspective on the translation process. *Translation, cognition and behavior* 2(2). 211–232.
- De Groot, Annette M. B. 1992a. Bilingual lexical representation: A closer look at conceptual representations. In Ram Frost & Leonard Katz (eds.), *Orthography, phonology, morphology, and meaning*, 389–412. Amsterdam: North Holland.
- De Groot, Annette M. B. 1992b. Determinants of word translation. *Journal of experimental psychology: learning, memory, and cognition* 18(5). 1001–1018.
- Dyvik, Helge. 2002. Translations as semantic mirrors: From parallel corpus to wordnet. In Karin Aijmer & Bengt Altenberg (eds.), *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, 311–326. Amsterdam: Rodopi.
- Gavarró, Anna & Brenda Laca. 2002. Les perífrasis temporals, aspectuals i modals. In Joan Solà, Maria Rosa Lloret, Joan Mascaró & Manuel Pérez-Saldanya (eds.), *Gramàtica del català contemporani*, vol. 3, 2663–2726. Barcelona: Empúries.
- Halverson, Sandra. 2003. The cognitive basis of translation universals. *Target* 15(2). 197–241.
- Halverson, Sandra. 2010. Cognitive translation studies: Developments in theory and method. In Gregory M. Shreve & Erik Angelone (eds.), *Translation and cognition*, 349–70. Amsterdam: John Benjamins.

- Halverson, Sandra. 2017. Developing a cognitive semantic model: Magnetism, gravitational pull, and questions of data and method. In Gert de Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies: New methods and theoretical traditions*, 9–45. Berlin: Mouton de Gruyter.
- Hareide, Lidun. 2017. Is there gravitational pull in translation? A corpus-based test of the Gravitational Pull Hypothesis on the language pairs Norwegian-Spanish and English-Spanish. In Meng Ji, Michael Oakes, Li Defeng & Lidun Hareide (eds.), *Corpus methodologies explained: An empirical approach to translation studies*, 188–231. London: Routledge.
- Labbé, Cyril & Dominique Labbé. 2013. La modalité verbale en français contemporain: Les hommes politiques et les autres. In David Banks (ed.), *La modalité, le mode et le texte spécialisé*, 33–61. Paris: L'Harmattan.
- Langacker, Ronald W. 1987. *Foundations of cognitive grammar*. Vol. 1: Theoretical prerequisites. Stanford: Stanford University Press.
- Langacker, Ronald W. 1991. *Foundations of cognitive grammar*. Vol. 2: Descriptive Applications. Stanford: Stanford University Press.
- Lewis, Diana. 2015. A comparable-corpus based approach to the expression of obligation across English and French. *Nordic Journal of English Studies* 14(1). 152–173.
- Malmkjær, Kirsten. 2008. Norms and nature in translation studies. In Gunilla Anderman & Margaret Rogers (eds.), *Incorporating corpora: The linguist and the translator*, 49–59. Clevedon: Multilingual Matters.
- Payrató, Lluís. 2002. L'enunciació i la modalitat oracional. In Joan Solà, Maria Rosa Lloret, Joan Mascaró & Manuel Pérez-Saldanya (eds.), *Gramàtica del català contemporani*, vol. 2, 1149–1220. Barcelona: Empúries.
- Radden, Günter & René Dirven. 2007. *Cognitive English grammar*. Amsterdam: John Benjamins.
- Rigau, Gemma. 1999. Los predicados impersonales relativos en las lenguas románicas. *Revista española de lingüística* 29(2). 317–355.
- Rigau, Gemma. 2005. Estudi microsinàctic del verb *caldre* en el català antic i en l'actual. *Caplletra* 38. 241–258.
- Schmid, Hans-Jörg. 2010. Does frequency in text really instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 101–133. Berlin: Walter de Gruyter.
- Tirkkonen-Condit, Sonja. 2002. Translationese—a myth or an empirical fact? *Target* 14(2). 207–220.

Josep Marco

- Tirkkonen-Condit, Sonja. 2004. Unique items—over- or underrepresented in translated language? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist?*, 177–184. Amsterdam: John Benjamins.
- Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.
- Vandevoorde, Lore. 2020. *Semantic differences in translation: Exploring the field of inchoativity*. Berlin: Language Science Press.

Chapter 3

Stylistic normalisation, convergence and cross-linguistic interference in translation: The case of the Czech transgressive

Olga Nádvoříková

Charles University

The Czech transgressive is a non-finite verb form belonging to the cross-linguistic converb category. In contrast with other converbs (e.g. Romance *gerundio* or the Russian

деепричастие), the Czech transgressive has a strong stylistic mark and is very rare in contemporary language. Using a parallel (multilingual) corpus and a comparable corpus of translated and non-translated Czech, the paper investigates the differences in the frequency of the transgressive in translated and non-translated fiction and non-fiction. The data shows the effect of stylistic normalisation in fiction, but not in non-fiction. The results of the potential effect of cross-linguistic interference are less conclusive, indicating that a thorough contrastive analysis of different language pairs is required first. Finally, the effect of convergence was observed neither in fiction nor in non-fiction.

1 Introduction

The Czech transgressive is part of the cross-linguistic category of converb, i.e. “a non-finite verb form whose main function is to convey adverbial subordination” (Haspelmath 1995: 3). Therefore, the Czech transgressive belongs to the same category as the Romance *gerundio*, English participial adjuncts in -ing, the Russian



Olga Nádvořníková

деепричастие and the Polish *imiesłów przysłówkowy*, which is also reflected in translations, as illustrated in example (1):

- (1) a. Czech
 “Bliju, soudruhu četaři,” odpověděl
 puke.1.SG.PRS comrade.VOC sergeant.VOC reply.M.SG.PST.PTCP
 jsem **opíraje** se rukou o
 be.1SG.AUX **lean.CONVERB.PS.IMPF.M.SG** REFL hand.INSTR against
 zeď.
 wall.ACC
 “I am puking, Comrade Sergeant,” I answered, leaning with one hand
 against the wall.’ (Milan Kundera, *Žert/The Joke*, 1991/1969)
- b. English
 “Puking, Comrade Sergeant,” I replied, **leaning** against the wall with
 one hand. (transl. David Hamblyn; Oliver Stallybrass, 1992)
- c. French
 Je suis en train de dégueuler, camarade sergent, expliquai-je **en**
m’appuyant d’une main au mur. (transl. Marcel Aymonin, 1975)
- d. Polish
 – Rzygam, towarzyszu plutonowy – odpowiedział em, **opierając się**
 ręką o mur. (transl. Emilia Witwicka, 1999)
- e. Russian
 “Блюю, товарищ сержант”, – ответил я, **опираясь** рукой о стену.
 (transl. Нина Шульгина, 1999)
- f. Spanish
 “Vomito, camarada sargento”, le respondí **apoyándome** con la mano
 en la pared. (transl. Fernando de Valenzuela, 1994)

However, in contrast with the other converbs, the Czech transgressive has a strong stylistic mark (bookish or even archaic), is used rarely and only in written texts.

This paper aims to investigate whether translated and non-translated Czech texts differ in the frequency of the transgressive. We assume that a higher frequency of transgressives in translations in comparison with non-translated texts may be attributed to the cross-linguistic influence (in translations from languages using converbs). The opposite result, i.e. a lower frequency of transgressives in translations than in non-translated texts, may be explained by the tendency to normalisation. We assume also that a higher tendency to convergence may be

manifested by a lower coefficient of variation of the frequency of the transgressive in translations.

The analysis is conducted on the InterCorp parallel (multilingual) corpus; a comparable corpus of translated and non-translated Czech, named Jerome; and a large monolingual synchronic corpus of Czech (SYNv8). All the corpora are limited to fiction and non-fiction. InterCorp and Jerome are used for comparison of translated and non-translated texts; the SYNv8 corpus (restricted to non-translated texts only) serves as a reference corpus for the other two corpora under analysis.

The paper is organised as follows. §2 introduces the twofold theoretical background of the research: the morphological, semantic, syntactic and stylistic properties of the Czech transgressive in the framework of the cross-linguistic category of converbs (§2.1) and the theory of special features of translated language, namely normalisation, convergence and cross-linguistic interference (§2.2). After presenting the corpora used in this research (§3), we introduce the results (§4): the analysis of the potential influence of normalisation and convergence in translations, and subsequently, the potential impact of the cross-linguistic interference. In the final part of the paper, we summarise the main outcomes of the research and suggest some open questions for future research.

2 Theoretical background

2.1 The Czech transgressive

2.1.1 The Czech transgressive as a converb

According to Nedjalkov (1998: 431), most European languages have one or two converbal forms, whereas languages outside Europe often have several converbal forms (e.g. Japanese, see Alpatov & Podlesskaya 1995).¹ According to Nedjalkov (1998: 430), polyconverb languages within Europe are Basque, Finnish and Lithuanian with six converbs each. Converb-free languages in Europe are rare, for example, Maltese and Romani (Nedjalkov 1998).

Due to their non-finite form (see the definition in §1), converbs convey the adverbial meaning in a more condensed way than the corresponding finite subordinate adverbial clause (see Vachek 1955; Nosek 1964; for Romance, for instance, Čermák et al. 2020; for Czech Bečka 1977). Because of their verbal character, they are also richer in information than complete nominalisations (verbal nouns,

¹For instance, “the average number of converbs in languages spoken within the territory of Russia is more than seven converbal forms per language” (Nedjalkov 1995: 431).

prepositional phrases, etc.). Moreover, in contrast with coordinate finite clauses, converbs allow for hierarchisation of information; in other words, the event conveyed by the converb is considered secondary (see 1a).

Converbs may differ regarding their position in the system of the given language (see Nedjalkov 1995: 104–106; 1998: 433). Strict, canonical converbs are specialised in the converbal (adverbial) function (e.g. Russian and Czech converbs, the French *gérondif* and Estonian forms in *–des*, etc.). By comparison, non-strict converbs fulfil, apart from the converbal function, other functions, e.g. participial or infinitival, such as the English forms in *–ing*, German *–end* or Spanish and Italian *–ndo* (see Nedjalkov 1998: 425; for Romance, see Čermák et al. 2020).

From the syntactic point of view, it should be noted that as non-finite verb forms, converbs do not have a valence position for the subject. In some languages, the subject (controller) of the converb has to be coreferential with the subject of the main clause, like in Slavic languages, including Czech (*same-subject converbs*, see Nedjalkov 1998: 425; Komárek 1986: 153)² or French (*gérondif*, see Grevisse & Goosse 2016: 1252). In other languages, the controller of the converb may be non-coreferential, as in Portuguese, Spanish (see Čermák et al. 2020: 111) or in Modern Greek, Armenian and Albanian (Nedjalkov 1998: 425).

Finally, concerning the semantic interpretation of converbs, we can distinguish contextual and specialised converbs (Nedjalkov 1998: 431). Specialised converbs only have one or two specific meanings (e.g. Finnish has a specialised converb conveying manner, see Nedjalkov 1998: 443). However, most European languages, including Czech, belong to the contextual converb type, i.e. their converbs are semantically vague, the potential meanings are numerous and given by the context (for the factors influencing the semantic interpretation of converbs, see for example König 1995; König & van der Auwera 1990: 337; Dvořák 1983: 29–41 for Czech; and for French Nádvořníková 2012).³

The meanings conveyed by contextual converbs can be divided into two large groups: temporal (simultaneity, anteriority and posteriority) and non-temporal (manner/means, cause, concession, condition etc.). The meaning of simultaneity proper or concomitance/attendant circumstance is the default meaning: “if a European language has only one converb, then it is a mixed converb of contextual Simultaneity” (Nedjalkov 1998: 432; see a similar observation for French *gérondif* in Kleiber (2007: 117; 2009: 19)). This observation is corroborated by a diachronic trend: “If a language moves from the group of two-converb languages

²Dvořák (1970: 37–45), in his diachronic study of Czech, points out that 30% of transgressives in the 17th century were non-coreferential.

³Moortgat (1978: 157) considers the French gerund to be a “semantic chameleon” (see also Halmøy 2003).

to the group of mono-converb languages then the remaining converb will belong to the contextual Simultaneity type” (Nedjalkov 1998: 437).

Despite the aforementioned variety in the subtypes, all converbs share the basic definition as a “non-finite verb form conveying the adverbial subordination” (Haspelmath 1995: 3, see §1). Most European converbs share other properties, in particular, the contextual semantic interpretation and the use as a means of syntactic condensation (*he said and he smiled* – *he said, smiling* – *he said, with a smile*). More specifically, Romance and Slavic converbs are considered typical (see Haspelmath 1995: 45 for the former and Nedjalkov 1998: 422 for the latter). According to Nedjalkov (1998), Slavic converbs are prototypical for the category, especially the Russian converb. Germanic languages, conversely, make, except for English, “only parsimonious use of converbs” (König 1995: 72, and a similar statement in Kortmann (1997: 192)). According to Kortmann, in English, “free adjuncts are far from playing a minor role” and the frequency of adverbial participial clauses is five times higher in English than in German (Kortmann 1997: 192). Similar differences in the use of converbs can be observed in the Slavic and Romance language families. In Slovak, the frequency of the converb is much lower than in Russian (Brtková 2004: 25). Similarly, in French, the frequency of *gérondif* is several times lower than the frequency of the corresponding forms in Italian, Spanish and Portuguese (1,571 instances per million words (ipm) against 4,098 ipm, 4,886 ipm and 6,939 ipm respectively, see Čermák et al. 2020: 116).

The Czech transgressive displays most of the properties considered by Nedjalkov as defining the prototypical (Slavic) converb: it is syntactically strict, i.e. it may be used only in the adverbial (converbal) function; it is formally simple, i.e. its formation involves suffixes, not adpositions; it has two converbal forms, one of which is a converb of contextual Simultaneity and the other as a converb of contextual Anteriority; and it is referentially the same-subject (co-referential). However, the Czech converb shows one important particularity: it maintains a very complex, archaistic morphology, involving adjectival agreement markers (in gender and number) (see §2.1.2), and, as a consequence, it acquires a strong stylistic mark and is used very rarely in contemporary language (see §2.1.3).

2.1.2 Morphological features of the Czech transgressive

As mentioned above, the Czech transgressive has two main forms (Karlík 2017):

- the “present transgressive” (*přechodník přítomný* or *-c transgressive*), formed with imperfective verbs only and conveying the meaning of simultaneity.

Olga Nádvorníková

- the “past transgressive” (*přechodník minulý* or *-š transgressive*), formed with perfective verbs only and conveying the meaning of Anteriority.

When influenced by contextual factors (see 2.1.1), these basic meanings may acquire other nuances, such as manner, cause, condition, etc. (Dvořák 1983: 33). Each form (present and past) has a different set of morphemes, varying according to the type of verb stem and as the consequence of the agreement with the subject (controller) of the transgressive in gender and number, as mentioned above. Table 3.1 summarises this complex morphology of the Czech transgressive.

Table 3.1: Morphology of the Czech transgressive

Czech transgressive forms		Form	
		Present	Past
M.SG		-a / -e / -ě	-Ø / -v
F.SG + N.SG		-ouc / -íc	-ši / -vši
PL (M+F+N)		-ouce / -íce	-še / -vše
Aspect	Imperfective	Present (Simultaneity) CONV.PS.IMPF	Simultaneity/Anteriority CONV.PT.IMPF
	Perfective	Futurate CONV.PS.PF	Past (Anteriority) CONV.PT.PF

Table 3.1 demonstrates four possible combinations of aspectual and formal characteristics of the Czech transgressive. Conv.ps.impf and Conv.pt.pf are the dominant forms, mentioned by most of the Czech grammars (e.g. Komárek 1986: 154; Cvrček 2010: 148–249; or Karlík et al. 1995: 335–337).⁴ For Conv.ps.impf, see 1a; for Conv.pt.pf, see 2.

- (2) **Uslyšev**, že Švejka naznačil plukovník ordonanci u 11. kumpanie, zvolal: “Pomoz nám pánbůh.” (Jaroslav Hašek, *Osudy dobrého vojáka Švejka za světové války*, 1921–1923/1996)
hear-CONV.PT.PF.M.SG
‘**Having heard** that the colonel marked Švejk to be the ORDONANC with the 11th company, he yelled out: “Help us Lord God.”’ (transl. Zdenek K. Sadloň)

⁴Nedjalkov (1998: 437) gives for the Czech converb conveying simultaneity the endings -a/-je/-oic, which is incorrect.

Conv.pt.impf was already rare in Old Czech (Dvořák 1970: 115); in contemporary Czech, it is not in use (Komárek 1986: 154). Finally, the form of the futurate transgressive (Conv.ps.pf) is supposed to be used only to convey anteriority in the future, i.e. combined with the main verb in the future tense (see Nedjalkov 1995: 126; Komárek 1986: 154). However, these forms were replaced by past converbs (Conv.pt.pf, see Oktábec 1953: 261) and are not in use either.

Even when limited to Conv.ps.impf and Conv.pt.pf, the morphology of the Czech converb is very complex and contrasts with the converbal systems in other Slavic languages, where the converb went through the process of adverbialisation and the forms were simplified, in particular by dropping the agreement with the subject (controller). In contemporary Czech, the only non-coreferential converbs are the grammaticalised ones: as much as in other languages (see Haspelmath 1995: 27–41), Czech converbs may be reanalysed in other categories, mainly adverbs (e.g. *chtě nechtě* ‘reluctantly’) and prepositions (e.g. *počínaje* ‘starting with’ or *nemluvě* ‘notwithstanding’, see Komárek 1986: 156).

Some languages maintained different sets of suffixes for the past (perfective) and the present (imperfective) forms (e.g. Russian and Polish). Other languages, for instance Slovak (Dvonč et al. 1966: 487), went further in the process of simplification and use the same set of suffixes for the perfective and imperfective verbs. The same tendency can be observed in Czech dialects (Dvořák 1983: 55–56; Michálková 1963), which demonstrates that spoken, non-standard Czech also adverbialised the transgressive.

This difference between standard literary Czech and its dialects (and other Slavic languages) is caused by a normative intervention made by grammarians during the Czech National Revival movement in the first third of the 19th century. At that time, the use of the Czech language was limited, since the language of economic and cultural elite was German, and Czech was spoken mostly by the rural population and the poorer inhabitants of cities. Therefore, while choosing the norm for the Czech language to be resuscitated, the grammarians and lexicographers of the National Revival movement did not opt for the language of their time (the 19th century), which was considered unprestigious and decayed, but the norm of the flourishing period of the Czech state, culture and language at the end of the 16th century, i.e. the language that was more than 200 years old at that time.

More importantly for our topic, the newly defined norm of standard literary Czech re-introduced in the transgressive its old complex morphology of the literary norm of the end of the 16th century. Since the transgressive was mostly used in written texts, especially for its advantages as a means of syntactic condensation, the norm was respected. Nevertheless, the transgressive gradually ceased

Olga Nádvorníková

to be part of the internalised, unconscious competence of the speakers; its frequency was in constant decline and the form acquired its stylistic mark.

2.1.3 Stylistic features and frequency of the Czech transgressive

Extensive research into transgressives conducted by Dvořák (1983: 60) demonstrates the constant decrease in the frequency of this form between 1781 and 1978 (from 6.49% of all verb forms in the period 1781–1830 to only 0.14% in 1971–1978). He observes the decrease in the frequency of the transgressive in the 18th century already (Dvořák 1970: 142), which indicates that the normative intervention during the National Revival movement may not have been the main factor triggering the decrease of the frequency of this form in Czech. Nevertheless, it is plausible to assume that in the 20th century, the archaistic morphology and stylistic mark resulting from the normative intervention contributed considerably to the retreat of this form. The most recent grammar of Czech, published in 2010 and based on corpus data, states that the transgressive is “very rare” and that it represents less than 1% of the verb forms in Czech (Cvrček 2010: 249).

It is worth noting that there is a neat difference in the frequency of the two main forms of the transgressive: Cvrček & Kovářiková (2011: 130) indicate that the frequency of *Conv.ps.impf* is nowadays less than 0.1%, but the frequency of *Conv.pt.pf* is even less than 0.01% of all verb forms, which means that the ratio of the two forms is 10. A similar difference in the frequency of the present and the past transgressive was already observed by Dvořák (1983: 60): 0.34% *Conv.ps.impf* and only 0.04% *Conv.pt.pf* of all verb forms in texts published between 1960 and 1970 (ratio 8.5). Conversely, in the period of 1781–1830, Dvořák observed 4.17% of *Conv.ps.impf* and 1.39% of *Conv.pt.pf*, i.e. only the ratio of 3.⁵ Even though the exact figures given by Dvořák for the different time spans may not be fully reliable, due to the lack of comparability of the sub-corpora under analysis, the tendency is clear: *Conv.pt.pf* is systematically less frequent than *Conv.ps.impf*.

The difference in frequency between the two main forms of the transgressive may be ascribed not only to the specific morphology of *Conv.pt.pf* (see Table 3.1), but also to the differences in the meaning of the two forms and the availability of concurrence forms in the language. *Conv.pt.pf*, conveying the meaning of anteriority, is strongly concurred by other forms, especially finite subordinate clauses of a temporal or a specific adverbial meaning (e.g. the cause, as in example 2). *Conv.ps.impf*, by contrast, mostly conveys a simple accompanying circumstance

⁵Dvořák also indicates the frequencies of the two remaining forms of the transgressive in 1781–1830: 0.03% for *Conv.pt.impf* and 0.9% for *Conv.ps.pf* (Dvořák 1983: 69).

(see example 1a or “řekl jsem *usmívaje se*” - CONV.PS.IMPF.M.SG ‘I said *smiling*’), which cannot be expressed by a subordinate clause, but only by a coordinate one (“řekl jsem *a usmíval jsem se*” ‘I said *and I was smiling*’) or a simple SP (“řekl jsem *s úsměvem*” ‘I said *with a smile*’).

Kortmann (1997: 281) made a similar observation for most European languages: they do not explicitly encode the meaning of concomitance (by an adverbial subordinator), so this meaning is mostly conveyed by converbs or a simple juxtaposition of two finite clauses. Even though the replacement of the converb by a subordinate clause moves the form from non-finite to finite and explicates its meaning by a subordinator (see Nádvorníková 2017), in contrast with the coordinate clause, it maintains the adverbial subordination relation and hence the hierarchisation of events typical for converbs (see the definition in §1). As a consequence, the coordinate clause is a less obvious concurrent of the verb than a subordinate one, and the meaning of the accompanying circumstance is more likely to persist in this form than more specific adverbial meanings. Furthermore, as remarked by Nedjalkov (1995), the accompanying circumstance is the most frequent meaning conveyed by converbs in general (see the same observation for Romance languages in Čermák et al. (2020: 122) and for Czech in Dvořák (1983: 33)).

As mentioned above, the archaistic morphology of the transgressive is also the source of its specific stylistic mark. Most Czech grammars consider the transgressive as bookish (Conv.ps.impf) or even archaic (Conv.pt.pf), and limited to the written language (Komárek 1986: 154; Cvrček 2010: 249; Karlík et al. 1995: 337). The stronger stylistic mark of Conv.pt.pf correlates with the aforementioned lower frequency.

The bookish/archaistic stylistic mark of the transgressive also influences its frequency in different text registers. Most sources agree that the transgressive is typical for fiction (Dvořák 1983: 105; Bečka 1977: 24; Čechová et al. 1997: 102), in particular because of its ability to convey in a condensed way the accompanying circumstance in narrative sequences and introductory clauses (Dvořák 1983: 107; Bečka 1977: 19 and example 1a). The stylistic mark in fiction is also exploited in historical novels or as a means of irony or parody (Čechová et al. 1997: 102–103; Komárek 1986: 154). However, in fiction intended for children or young readers, transgressives are less frequent than in fiction for adults (see Jelínek et al. 1961: 90).⁶ In non-fiction, the transgressive is considered less frequent than in fiction and conveys more specific adverbial meanings than a simple accompanying cir-

⁶Bečka (1977: 23) also mentions the potential influence of a specific author’s idiolect (e.g. the Czech author Vladislav Vančura, 1891–1942, is known for his penchant for transgressives).

Olga Nádvorníková

cumstance (see Dvořák 1983: 33; Bečka 1977: 21; Karlík et al. 1995: 337).⁷ Dvořák (1983: 106 and 108) points out that the transgressive is more frequent in social sciences than in natural or technical sciences. Finally, in journalistic texts, the transgressive is the least frequent, in comparison with fiction and non-fiction (Dvořák 1983: 106; Jelínek et al. 1961: 90).

2.2 The transgressive in translations

To our knowledge, only a few researchers have focused specifically on the use of transgressives in translation, apart from two rather dated studies (Bečka 1977; Dvořák 1972; 1983). However, the topic is occasionally addressed in contrastive studies exploring Czech equivalents of converbs.

In his quantitative study, Dvořák (1972; 1983) analysed various Czech translations of the same source texts (four source texts in Russian, one in French and one in English). The translations were published between 1863 and 1975 and six different translations in average were analysed for each text. The results confirmed the decrease in the frequency of the transgressive observed in non-translated texts (see §2.1.3), but the normalised frequency of transgressives was almost always higher in translations than in non-translated texts from the corresponding period. For instance, in the Czech translation of Charles Dickens' *The Posthumous Papers of the Pickwick Club* (chapters 1-5) published in 1925, transgressives represent 5.5% of all verb forms, whereas the average for the given period in non-translated texts is only 1.384% (Dvořák 1983: 94).

These results suggest that in translations from languages using converbs (i.e. most European languages, see below), the effect of cross-linguistic interference (or *shining through*, see §2.2.1) may be expected and the frequency of transgressives may be higher in translated than in non-translated texts. However, other studies indicate the opposite conclusion.

First, in their contrastive research of Czech equivalents of Romance converbs, Čermák et al. (2015; 2020) show that in Czech translations from four Romance languages (French, Italian, Portuguese and Spanish), the transgressive represents the least frequent counterpart (from 2.0% in translations from French to 9.6% in translations from Portuguese), despite the presumed systemic equivalence. In comparison, the finite counterparts (coordinate and subordinate clauses) form about 70% of the whole. Malá & Šaldová (2015: 240) present a similar result in translations from English: the transgressive represents only 2.1% of the Czech counterparts of English adverbial participles; the overwhelming majority of counterparts being

⁷Karlík (2017) and Čechová et al. (1997) consider the transgressive also appropriate in highly formal, e.g. diplomatic or legal, documents.

finite verbs (73%). Finally, in translations from Russian (Kocková 2011), transgressives constitute less than 1% of the equivalents of (past) *деепричастия*. These results suggest that in translations, the frequency of the transgressive may be as low as in non-translations, or even lower.

Second, Bečka (1977: 26) in his (non-quantitative) analysis of the transgressive in translations points out that translators had been warned against the use of the transgressive and that they avoid it because of its stylistic mark. Similarly, Levý (2011: 51) states that the frequency of the transgressive in Czech is lower in translations than in non-translated texts, because translators are over-concerned to avoid stylistically marked features. These observations indicate that, on the contrary, the frequency of transgressives may be influenced by the effect of stylistic normalisation and, therefore, be lower than in non-translated texts.

2.2.1 Specific translation features

Toury (1995) states that translations are governed by two universal laws: the law of interference and the law of growing standardisation (or *normalisation*, according to Baker (1993; 1996)). Cross-linguistic interference (or *shining through*, according to Teich (2003)) consists of transferring linguistic features of the source language into a target language (see Toury 1995: 274–279). Normalisation, by contrast, may be defined as “the tendency to conform to patterns and practices that are typical of the target language, even to the point of exaggerating them” (Baker 1996: 176–177).

Various studies have shown the effect of cross-linguistic interference in translation in various language pairs. For instance, Dai & Xiao (2011), when analysing Chinese texts translated from English, found that passive voice is more frequent in Chinese translated from English than in non-translated Chinese texts. Similarly, Cappelle (2012) shows that English texts translated from French contain fewer manner-of-motion verbs than English texts translated from German. He explains this effect by the typological differences between the two source languages: German and English are satellite-framed languages, whereas French is a verb-framed language.

As for the normalisation, this is defined by the linguistic properties as well as the sociocultural norms of the target language (see Lefer & Vogeleeer 2013: 17). Alongside the explicitation, the simplification and the levelling out (convergence, Laviosa 2002) it is one of the specific features of translation (“translation universals”, according to Baker (1993; 1996)) that is addressed the most in literature. In Chesterman’s (2004: 39) terms, it can be conceived either as an S-universal, causing differences “between translations and their source texts”, or as a T-universal,

Olga Nádvorníková

giving rise to differences between translations and comparable non-translated texts in the target language. In this study, we focus on the effect of normalisation as a T-universal, by comparing non-translated Czech texts with translations in the same language. We also partially focus on the convergence (levelling out).

Normalisation as a T-universal was analysed by [Delaere et al. \(2012\)](#): their results confirmed the tendency to normalisation (standardisation) in translated Dutch, in comparison with non-translated texts of the same language. Similarly, [Chlumská \(2017\)](#) observed the effect of normalization in translation in the choice of two forms of the verb *say* in Czech: *řici* (formal, stylistically marked form) and *říct* (standard, stylistically neutral). In her corpus, translations showed a higher frequency of the latter form than non-translations, which suggests the effect of (stylistic) normalisation ([Chlumská 2017](#): 65). [Lapshinova-Koltunski \(2018\)](#), who compared translations from English into German (in six different text registers), observed that the effect of normalisation is sensitive to two factors: text register (the highest score of normalisation was in translations of fiction) and the translator's proficiency (the normalisation score was higher in student translations than in professional translations).

Levelling out (or convergence, [Laviosa 2002](#)) is sometimes considered as a sub-type of normalisation. [Baker \(1996](#): 177) defines levelling out as “the tendency of translated text to gravitate around the centre of any continuum rather than move towards the fringes”. [Laviosa \(2002](#): 71) is more specific and points out that the convergence implies a relatively higher level of homogeneity of translated texts concerning certain linguistic features, such as lexical density, sentence length, etc. As stated by [Baker \(1996](#): 184) and by [Chlumská \(2017](#): 104), less attention has been paid to this feature than to the other translation universals as it is more difficult to operationalise. [Lapshinova-Koltunski \(2015\)](#) confirmed the tendency to convergence in several translation variants in German (translated from English). [Chlumská \(2017](#): 104–121) analysed various potential indicators of the convergence (sentence length, TTR, etc.) in Czech and observed its effect in translations of fiction but not in non-fiction.

2.3 Hypotheses and research questions

Our main research question aims to find out the differences in the frequency of transgressives in translated and in non-translated texts. Based on the theory of the interplay between the cross-linguistic interference and the normalisation in translation (see §2.2), we can formulate the following hypotheses (H_1 and H_2 being in opposition):

H_0 Translated and non-translated texts of the same text register do not differ in the frequency of transgressives.

H_1 Due to the effect of cross-linguistic interference, in translations from languages using converbs, the frequency of transgressives is higher than in non-translated texts of the same text register. Based on typological observations made in the literature (Haspelmath 1995, Nedjalkov 1995, see §2.1.1), we expect more transgressives in translations from Romance and Slavic languages, especially Russian, than in translations from Germanic languages (with the potential exception of English). Transgressives resulting from this interference can also be expected in translations from poly-converbal Latvian, Finnish and Japanese.

H_2 Due to the effect of (stylistic) normalisation, the frequency of transgressives is lower in translations than in non-translated texts of the same text register (independently of the source language and the text register).

Based on the theory of convergence (see §2.2.1), we can formulate the third hypothesis:

H_3 Due to the tendency of translations to the convergence (greater homogeneity), the coefficient of variation of the frequency of transgressives in translations is lower than in non-translated texts of the same text register.

Our second research question aims to find out what other factors influence the frequency of transgressives in translated and non-translated texts. From a strictly linguistic point of view, we expect an important difference in frequency between the two forms of the transgressive since the past form (Conv.pt.pf) is stylistically more marked than the present form (Conv.ps.impf). Among the extra-linguistic factors, we expect the greatest influences to be the date of publishing of the text (the older the text, the higher the frequency) and the text register (more transgressives in fiction, exploiting its stylistic mark and the ability to convey accompanying circumstance, and fewer transgressives in non-fiction, more stylistically neutral than fiction).

3 Data and Methods

As mentioned in §2, the main source of data for our research is corpora including translated texts: the comparable corpus of translated and non-translated Czech

Olga Nádvorníková

named Jerome (Chlumská 2013), and the InterCorp parallel (multilingual) corpus (Rosen et al. 2019). The data obtained from these corpora is confronted with the data extracted from the large monolingual synchronic corpus of Czech named SYNv8, limited to non-translated texts (Křen et al. 2019).

All these corpora were created by the Institute of the Czech National Corpus and are freely available using the same corpus interface (KonText; www.korpus.cz and <http://kontext.korpus.cz>). InterCorp was annotated using the POS-tagger named Morče (see <http://ufal.mff.cuni.cz/morce/index.php>); SYNv8 and Jerome were annotated by an in-house developed hybrid system (combining stochastic and rule-based disambiguation, see Hnátková et al. 2011).

The error rate of the transgressive POS-tagging is 3.6% for the Conv.ps.impf and 6.8% for the Conv.pt.pf (tested on a sample of 250 occurrences for each form).⁸ Even though these POS-tagging errors can influence the resulting frequencies, we consider our observations reliable, since the main information for our research is not the absolute frequencies of the transgressive, but the comparison of the relative ones.

3.1 The Jerome comparable corpus of translated and non-translated Czech

Jerome (Chlumská 2013; 2017) is a monolingual corpus specifically designed for the research of translation features (in terms of T-universals, see Chesterman (2004) and §2.2), in fiction and non-fiction. The corpus comprises translated and non-translated texts in Czech, the two subcorpora being comparable in size and other relevant factors. For instance, all the texts were published between 1992 and 2009 and the same author/translator may be represented by a maximum of three texts to prevent the risk of the influence of a specific idiolect.

The representation of source languages in the translation part of the corpus reflects the situation on the publishing market in the Czech Republic, where translations from English are three times more frequent (i.e. probably three times more read) than from any other language.⁹ In total, it includes 22 different source

⁸In the corpus queries, we excluded from the analysis the most frequent grammaticalised forms of the transgressive in the two sub-corpora. The resulting queries are [tag="V(e|m).*" & word!="((N|n)e)?(C|c)ht(ě|íc)((N|n)emluvě|(P|p)očinaj(e|íc)|(K|k)onč(e|íc)|(N|n)evyjímaj(e|íc)|(T|t)ak říkájíc|(S|s)oudě"] for fiction and [tag="V(e|m).*" & word!="((N|n)e)?(C|c)ht(ě|íc)((N|n)emluvě|(P|p)očinaj(e|íc)|(K|k)onč(e|íc)|(N|n)epočítaj(e|íc)"] for non-fiction.

⁹According to the Czech National Library statistics of translated books, available (in Czech) at <http://text.nkp.cz/sluzby/sluzby-pro/sluzby-pro-vydavatele/vykazy>, for more details, see Cvrček & Chlumská (2015: 313).

languages in fiction and 15 source languages in non-fiction. The potential interference effect can be explored using a smaller balanced subcorpus including an equal amount of texts translated from 14 different languages in fiction and 6 languages in non-fiction. Table 3.2 summarises the composition of the Jerome comparable corpus:

Table 3.2: Composition of the Jerome comparable corpus of translated and non-translated Czech

Jerome corpus	Nº of tokens (incl. punctuation)		Source languages in translation
	translated	non-translated	
Fiction (non-balanced)	26,617,523	26,551,540	da nl en fi fr de el(new) he hu is it ja no pl pt ru sr sk sl es sv
Non-fiction (non-balanced)	15,946,319	15,949,930	ar en fr de el(old) hu it la pl ro ru sr sk es sv
Fiction (balanced)	1,765,433	1,768,079	da nl en fi fr de is it ja pl pt ru es sv
Non-fiction (balanced)	774,610	779,288	en fr de it pl ru

In the whole (non-balanced) corpus, translations from English represent 69% of the subcorpus of translations (by the number of tokens). The other languages represented by at least 500,000 tokens are French and German (8% each) and Russian and Polish (3% and 2% respectively). In non-fiction, the composition is similar: translations from English represent 55% of the sub-corpus, followed by German (25%) and French (8%). The other languages usually do not exceed 1%.

As mentioned above, all the texts included in the Jerome corpus were published between 1992 and 2009; nevertheless, some of them were first published earlier. Since the frequency of the transgressive is highly likely to be sensitive to the date of the creation of the text (see §2.1.3), we eliminated these texts from our corpus research (14 texts in translated fiction, 24 texts in non-translated fiction, one text in translated non-fiction and three texts in non-translated non-fiction). Thus, in translated fiction, we excluded, e.g. the Czech translation of William Faulkner’s novel *The Wild Palms*, first published in 1960. In non-translated fiction, the set of eliminated texts includes not only texts first published before 1992,

Olga Nádvorníková

for instance four novels by Vladislav Vančura,¹⁰ first published in the 1920s, but also texts first published in the given interval (1992–2009), but written earlier, e.g. a posthumous edition of a novel by Lev Blatný (*Servus, Ser-vá-ci*), an author who died in 1930 and the memoirs of an Habsbourg Empire army officer in the Great War (*Z turecké armády do britského zajetí*).

It is very likely that, especially in translations, new editions of texts published earlier were revised and adapted by editors in publishing houses. However, our data shows that the inclusion of these texts in the research may have skewed the results, if not excluded. For instance, all the aforementioned eliminated texts show the highest normalised frequencies of the transgressive in our subcorpora of fiction, which corroborates the hypothesis of the influence of the date of creation/publication of the text on the frequency of the transgressive.

Furthermore, to maintain the comparability with the InterCorp parallel corpus (see below), we limited the Jerome corpus to novels and short stories in fiction and scientific (SCI) and popular (POP) texts in non-fiction (eliminating e.g. textbooks and encyclopaedias, not included in InterCorp). For these reasons, the size of the Jerome sub-corpora introduced in the results of our research (see §4) is smaller than that given in Table 3.2.

3.1.1 The InterCorp multilingual corpus

InterCorp (<https://wiki.korpus.cz/doku.php/cnk:intercorp:verze12>) is a large multilingual (parallel) corpus currently involving 41 languages, with Czech as pivot language (Čermák & Rosen 2012; Nádvorníková 2016). The corpus is composed of the so-called core, which comprises fiction and partially non-fiction, and collections (movie subtitles, the Bible, journalistic texts, *Acquis communautaire* and EuroParl). Our research exploits only the core of the corpus, because, in contrast with the collections, the quality of translations is higher in the core texts and all the metadata necessary for research in translation studies is available (date of publication, source language, name and sex of the author/translator, different text sizes in tokens, etc.).

The main advantage of the InterCorp parallel corpus, in comparison with the Jerome comparable corpus, is its larger size, i.e. the larger number of texts and different authors/translators, which reduces the risk of the influence of a specific text style or an author's/translator's idiolect. This is also the reason why we do not use the Jerome corpus in the interference hypothesis testing (H_1), as it is limited to one to three texts per language, but instead, we use the InterCorp parallel corpus (see §2.1.3).

¹⁰ *Amazonský proud*, *Pekař Jan Marhoul*, *Pole orná a válečná* and *Poslední soud*.

However, the translated and non-translated sub-corpora of the InterCorp corpus are not comparable, neither in size nor composition. As can be seen in Table 3.3, the size of the non-translation sub-corpora in the InterCorp corpus is quite small, which is due to the limited availability of translations from Czech into foreign languages. This issue is even more pronounced in non-fiction than in fiction. In addition, the sub-corpus of non-translations in InterCorp is not limited solely by size, but also by composition, as foreign publishing houses particularly choose texts by well-known and established authors for translations from Czech. As a consequence, the non-translation subcorpora of InterCorp are not a reliable source of data for real language use in Czech.

For this reason, the data for the comparison of translations with non-translated texts was not extracted from InterCorp, but from the largest corpus of contemporary Czech – SYNv8 (Křen et al. 2019; Hnátková et al. 2014), limited to non-translated fiction (novels and short stories) and non-fiction (scientific and popular texts). Table 3.3 demonstrates the resulting size of the sub-corpora.

Table 3.3: Composition of the InterCorp parallel corpus and the SYNv8 reference corpus of Czech

Corpus		InterCorp		SYNv8
		translated	non-translated	non-translated
Fiction	texts (n)	1,179	286	496
	tokens (n)	107,375,278	19,208,622	30,527,709
	SLs (n)	32	–	–
Non-fiction	texts (n)	80	13	650
	tokens (n)	6,803,832	881,833	33,878,274
	SLs (n)	5 (de,it,fr,en,sv)	–	–

The overwhelming majority of texts in the InterCorp parallel corpus and SYNv8 were published after 1950, with the majority after 1980. However, some texts were first published much earlier, e.g. *Osudy dobrého vojáka Švejka* by Jaroslav Hašek (1921–1923, see example 2) and the Czech translation of *The Jungle Book* by Rudyard Kipling (1911).

To maintain the comparability with the results obtained on the Jerome corpus and to reduce the influence of the date of publishing, we limited the InterCorp parallel corpus and the SYNv8 to texts (first) published after 1992. The whole corpora, including older texts, are only used to analyse the evolution of the frequency of the transgressive (see Figures 3.1 and 3.2). The normalisation, conver-

Olga Nádvorníková

gence and cross-linguistic hypotheses are thus tested only on the texts published after 1992 (inclusive). Similar to the Jerome corpus, the resulting sub-corpora provided in §4 are smaller than those in Table 3.3. Furthermore, as much as in the Jerome corpus, even in the sub-corpora limited to texts published after 1992, we identified and eliminated some of the texts first published or written earlier (e.g. the novel *Nesmrtelnost/Immortality* by Milan Kundera, written in 1987–1988 and the Czech translation of *Les Mots* by Jean-Paul Sartre, first published in 1967).

As for the source languages, the whole fiction/non-fiction sub-corpus of InterCorp involves 31 different source languages: Arabic (ar), Belarussian (be), Bulgarian (bg), Catalan (ca), Croatian (hr), Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), German (de), Hindi (hi), Hungarian (hu), Italian (it), Japanese (ja), Lithuanian (lt), Latvian (lv), Macedonian (mk), Norwegian (no), Polish (pl), Portuguese (pt), Romany (rn), Romanian (ro), Russian (ru), Slovak (sk), Slovene (sl), Serbian (sr), Spanish (es), Swedish (sv), Turkish (tr) and Ukrainian (uk).¹¹ The most represented languages are German and English (more than 30 million tokens each, i.e. more than 10% of the corpus each). The source languages representing between 5% and 10% of the corpus (i.e. more than 20 million tokens) are Polish, Spanish, Croatian and French (for detailed information, see <https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze12>).

In the corpus limited to texts published after 1992 (inclusive), the number of source languages is only twenty (da, de, en, es, fi, fr, hr, it, ja, lv, nl, no, pl, pt, ro, ru, sk, sl, sr, sv). Translations from English prevail (36% of the sub-corpus), followed by German, Spanish and Swedish (see Table 3.5 for more details). It can be observed that all the languages included in this sub-corpus belong to the European area (except for Japanese) and except for Finnish and Latvian, they all belong to one of the three prevailing language families in Europe (Romance, Slavic and Germanic). The corpus thus allows to test the normalisation and convergence hypotheses (§4.1) and investigate the potential cross-linguistic interference effect (§4.2).¹²

¹¹Nine source languages are available in collections only and not in the core of the corpus: Greek (el), Estonian (et), Hebrew (he), Icelandic (is), Malay (ms), Maltese (mt), Albanian (sq), Chinese (zh) and Vietnamese (vi).

¹²The number of source languages in the Jerome corpus is higher than in the InterCorp parallel corpus, because InterCorp includes only source languages for which source texts are really available in the corpus, whereas the Jerome corpus simply includes all translated texts available in Czech.

4 Analysis

Even though our main analysis focusses on the potential effects of normalisation, convergence and cross-linguistic interference in translation (see Sections 4.1 and 4.2), we will first briefly examine the evolution of the frequency of the transgressive, in translated and non-translated texts. By doing so, we intend to verify the soundness of the limitation of the data for our analysis of the texts published after 1992 (inclusive). As mentioned above (§3.1), the sub-corpus of translations includes all the texts in the translated sub-corpus of InterCorp (limited to fiction and non-fiction), regardless of the date of publication or the source language. The non-translated texts are extracted from the reference corpus SYNv8.

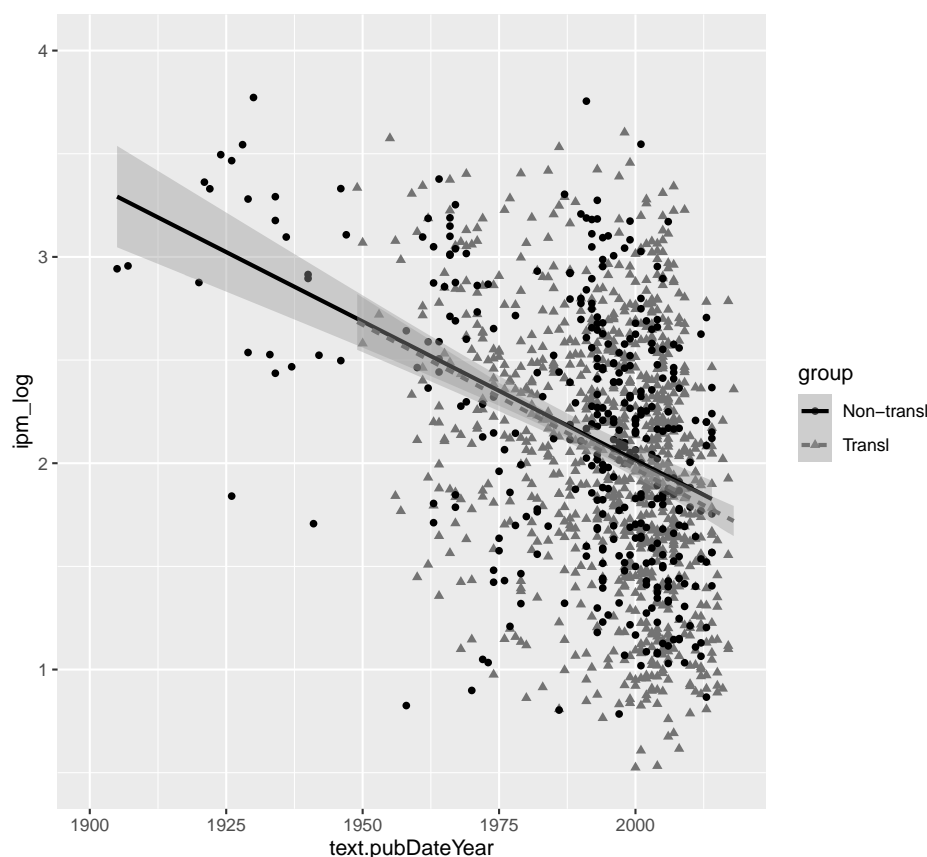


Figure 3.1: Normalised frequency of the transgressive in translated and non-translated fiction (InterCorp vs. SYNv8)

Olga Nádvořníková

As shown in Figures 3.1 and 3.2, the time span for non-translated texts is larger than that of translations: the first texts in non-translated sub-corpora were published at the beginning of the 20th century in fiction (*Pověsti vyšehradské* by Popelka Biliánová, 1905) and in the 1920s in non-fiction. The first translations, by contrast, start in 1949 in fiction (Jorge Amado's novel *Suor*, a translation from Portuguese) and in non-fiction (*Wstęp do semantyki* by Adam Schaff, a Polish author) in 1963. Since the language of translations becomes obsolete faster than that of non-translated texts, this difference is well-founded.

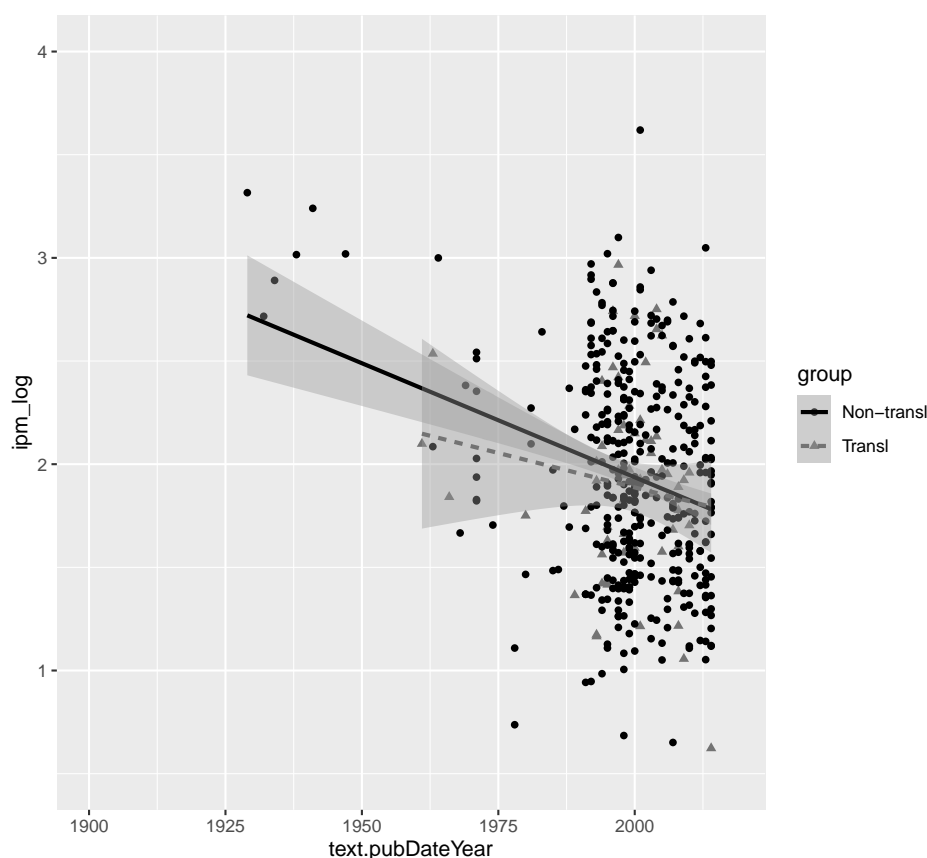


Figure 3.2: Normalised frequency of the transgressive in translated and non-translated non-fiction (InterCorp vs. SYNv8)

Figures 3.1 and 3.2 fully confirm the tendency observed in previous research (see §2.1.3): the normalised frequency of the transgressive constantly decreases in both translated and non-translated texts. It is worth noting that the decrease

is more important in fiction than in non-fiction and that the frequency of the transgressive is lower in non-fiction than in fiction. It is also necessary to point out that the actual decrease may be less dramatic than suggested by these figures, since the form of the graph is influenced by the few texts at the beginning of the observed period showing very high frequencies of the transgressive.

The data also indicates that the decrease in the frequency of the transgressive also continues after 1992, which suggests that there may be differences due to the time factor between texts within the time span of the limited corpora used in the main research. However, a further limitation of the corpus to more recent texts would have reduced the reliability of the results; hence we maintain the 1992 limit.

As for the differences between the translated and non-translated texts, Figures 3.1 and 3.2 suggest that they are only very slight, with a tendency to differentiation in the latest years in fiction and to a similarity in non-fiction. In what follows, we will investigate the statistical significance of these differences.

4.1 Normalisation and convergence in translation

Tables 3.4 and 3.5 show the absolute (n) and normalised (f) frequencies in instances per million words of the transgressive in translated and non-translated fiction and non-fiction, and the standard deviation (SD) and the coefficient of variation (CV, (SD/f)*100) for all the subcorpora. Even though the coefficient of variation is mostly higher in non-translations than in translations, with the exception of the fiction part of the Jerome corpus, the differences are very slight. This means that the convergence hypothesis (H₃, see §2.2.1) is not confirmed by our data, and with regard to the frequency of the transgressive, translations do not show more homogeneity than non-translated texts.

Table 3.4: Frequencies of the transgressive (both forms) in fiction (n = absolute frequency, f = normalised frequency in instances per million words, CV = coefficient of variation)

Fiction	corpus	texts	tokens	n	f	SD	CV
transl	Jerome	380	23,301,169	2,538	108.92	228.23	209.54
non-transl	Jerome	247	15,692,373	2,795	178.11	368.50	206.89
transl	InterCorp	774	71,063,940	9,268	130.42	262.94	201.61
non-transl	SYNv8	328	20,663,102	3,090	149.54	343.43	229.66

Olga Nádvorníková

Table 3.5: Frequency of the transgressive (both forms) in non-fiction (n = absolute frequency, f = normalised frequency in instances per million words, CV = coefficient of variation)

Non-fiction	corpus	texts	tokens	n	f	SD	CV
transl	Jerome	221	15,904,500	754	47.41	113.73	239.89
non-transl	Jerome	242	15,719,462	813	51.72	126.67	244.91
transl	InterCorp	78	6,591,970	720	109.22	160.85	147.27
non-transl	SYNv8	592	30,988,911	3,447	111.23	166.90	150.05

As for the normalisation hypothesis, Tables 3.4 and 3.5 show that the normalised frequency of the transgressive is indeed higher in non-translations than in translations, regardless of the corpus (Jerome translated/non-translated or InterCorp/SYNv8) and the text register (fiction or non-fiction). However, the differences in the normalised frequency of the transgressive are statistically significant in fiction only ($p < .0001$), as demonstrated in Figures 3.3–3.6. The differences observed in non-fiction are not significant even at $p < .05$. This means that the normalisation hypothesis is confirmed in fiction, but not in non-fiction. From the methodological point of view, this result also indicates that the investigation of specific features of translation may be strongly text-type dependent.

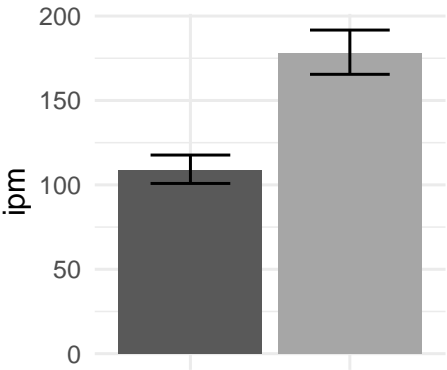


Figure 3.3: Jerome translated vs. non-translated (fiction)

The difference between the two text registers (fiction and non-fiction) regarding the tendency to normalisation may be due to various factors, especially because in fiction, translators are more likely to exploit the stylistic mark of the

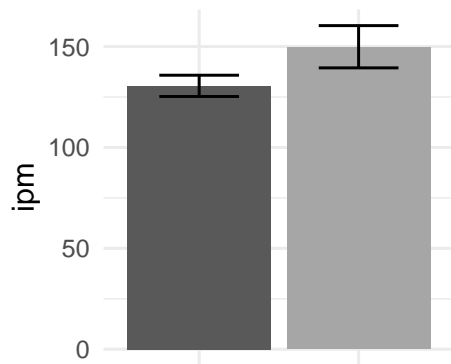


Figure 3.4: InterCorp vs. SYNv8 (fiction)

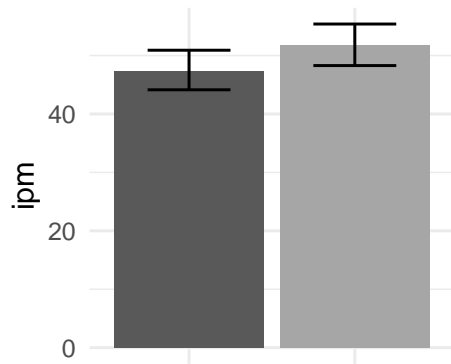


Figure 3.5: Jerome translated vs. non-translated (non-fiction)

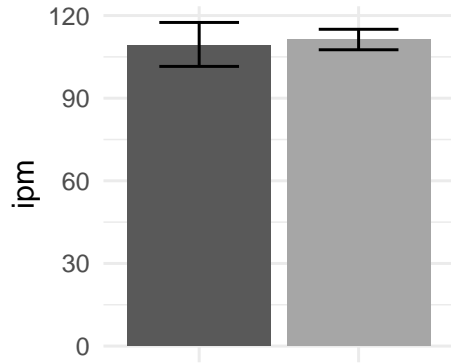


Figure 3.6: InterCorp vs. SYNv8 (non-fiction)

Olga Nádvorníková

transgressive, which may increase their awareness about the pitfalls of this form and cause the stylistic normalisation. Among the texts at the top of frequency list of the transgressive in fiction are (in both translated and non-translated sub-corpora) texts overtly exploiting the archaistic stylistic mark of the transgressive, in particular historical novels and fantasy stories (e.g. Andrzej Sapkowski's fantasy novel *Miecz przeznaczenia* tops the list of translations in InterCorp – 2,652.29 ipm). A similar motivation is found in translations of old source texts, first published in the 19th century. Even if they do not belong to the category of historical novels and the translations are recent, these texts show a high normalised frequency of the transgressive (e.g. Eça de Queiros' novel *A Cidade e as Serras*, 1,900.2 ipm, and two novels by Honoré de Balzac – *Le colonel Chabert* and *Gobseck*, 2026.87 ipm). Especially in non-translated texts at the top of frequency lists, the transgressive is used also in order to create a humoristic or ironic effect.¹³ In some texts, the transgressive reflects a specific, syntactically complex style of the author of the source text, e.g. in the translation of the novel *Trans-Atlantyk* by the Polish author Witold Gombrowicz (2,196.12 ipm) and in a collection of short texts by the Belgian (French-writing) author Jean-Philippe Toussaint *Autoportrait* (1,817.00 ipm).

Nevertheless, a much more thorough analysis of the types of the use of transgressives (in translated as well as in non-translated texts) is needed. For instance, various factors triggering the use of transgressives may combine in one text,¹⁴ and in texts in the middle or at the bottom of the frequency list the use of the transgressive may be less motivated by its stylistic properties than by its use as a means of syntactic condensation. However, the aforementioned types extracted from the top of the frequency lists indicate that the frequency of the transgressive in fiction is probably closely related to the specific style of individual texts and authors.

In contrast, in non-fiction, not only the overall frequency of the transgressive is lower than in fiction (see Tables 3.4 and 3.5), but it appears more governed not by the specific stylistic norm of the text sub-type rather than the individual style of texts and authors. Most texts containing transgressives in non-fiction sub-corpus belong to the domain of social sciences (both in translated and non-translated

¹³For instance, at the most top of the frequency list in non-translated fiction, we find a short text by Michal Šanda (*Obecní radní Stoklasné Lhoty vydražíví za 37 Kč vycpaného jezevce pro potřeby školního kabinetu*), with 3,517.69 ipm of the normalised frequency of the transgressive.

¹⁴In the *Autoportrait*, for instance, the high frequency of the transgressive may be the result of a combination of complex syntax and irony in the source text (personal communication of the Czech translator of the text Jovanka Šotolová). The age and personal style of the author (in non-translated texts) and the translator (in translations) may also come into play.

texts), especially philosophy and religious studies (Radim Palouš *Totalismus a holismus*, 756.93 ipm or *Cogitata metaphysica* by Benedict de Spinoza), literary studies (e.g. Roland Barthes' *Mythologies*, 632.16 ipm) and history (e.g. *Každodennost renesančního aristokrata* by Marie Šedivá 717.52 or Ferdinand Seibt's *Deutschland und die Tschechen*). In technical and natural science books, by contrast, the transgressives are much less frequent or even completely missing.¹⁵

It is important to point out that in non-fiction, the proportion of texts containing zero transgressives is higher than in fiction (one quarter of texts are without any transgressive in the latter and one third in the former). More importantly for our topic, in both corpora (Jerome and InterCorp), more texts show zero transgressives in translations than in non-translated texts, and the maximum frequencies are higher in non-translated texts than in translations.¹⁶

Figures 3.7 and 3.8 show density plots of the normalised frequencies of the transgressive in the fiction part of InterCorp/SYNv8 (Figure 3.7) and the Jerome corpus (Figure 3.8) in translated and non-translated texts. It can be seen that in both corpora, even though the number of texts showing higher normalised frequencies of the transgressive is higher in non-translated texts than in translations, the differences are not extensive. Thus, the main difference between the translated and non-translated texts consists mainly in “category zero”: the number of texts containing no transgressive at all is higher in translations than in non-translated texts. This is also the main cause of the normalisation effect in translations.

Figures 3.7 and 3.8 suggest that if translators decide to use transgressives, they do so in a way similar to non-translated texts. However, more translators than authors of original Czech texts decide not to use transgressives at all. In the Jerome corpus, for instance, 31% of translations do not contain any transgressive, and 13% only one, i.e. 44% of texts with an extremely low frequency of the transgressive. In the non-translated texts, only 20% of texts show no transgressive and 11% only one occurrence, i.e. only 31% of texts without (or almost without) transgressives. These results suggest that translators could use more transgressives without be-

¹⁵This difference, already observed in previous studies (Dvořák 1983: 106 and 108, see §2.1.3), may also explain the difference in the normalised frequency of the transgressive in the non-fiction sub-corpora of Jerome on the one hand, and SYNv8 on the other hand: the former is a mix of various text register sub-types, whereas the latter contains more books from the domain of humanities.

¹⁶By contrast, in all the subcorpora, regardless of the text register or the translated/non-translated distinction, about a quarter of texts show the normalised frequency of the transgressive to be higher than the average of the whole sub-corpus (25% in SYNv8 fiction and 25% in all the other sub-corpora).

Olga Nádvorníková

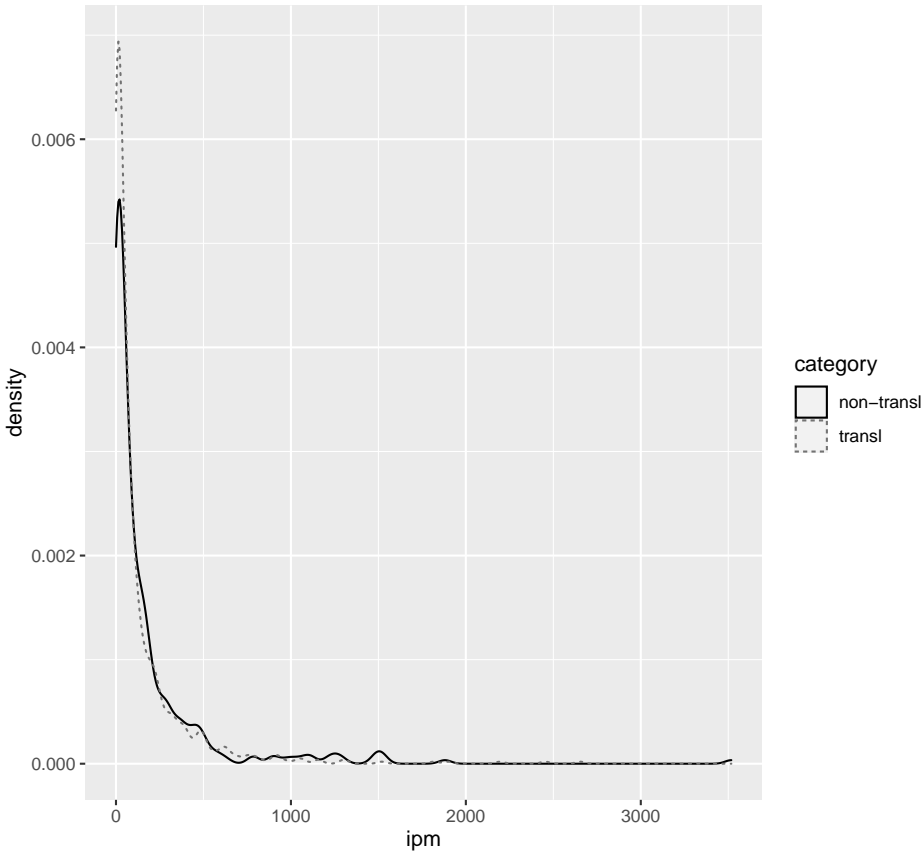


Figure 3.7: InterCorp/SYNv8 translated vs. non-translated density plot (fiction)

ing afraid to violate the norm of the target language (with respect to the style of the source text, of course).

It is also worth noting that in fiction, normalisation and convergence, are more pronounced in the past transgressive forms (Conv.pt.pf) than in the present forms (Conv.ps.impf). As expected, the frequency of the past transgressive is much lower than that of the present form (Conv.pt.pf represents 6% of all transgressives in translations and 14% in non-translations, see Table 5). However, the rate of the difference between translated and non-translated texts is higher in Conv.pt.pf than in Conv.ps.impf (3.24 and 1.52 respectively). The tendencies are similar in both text registers and both corpora (Jerome and InterCorp/SYNv8); therefore, we illustrate these with the numbers for the fiction part in the Jerome

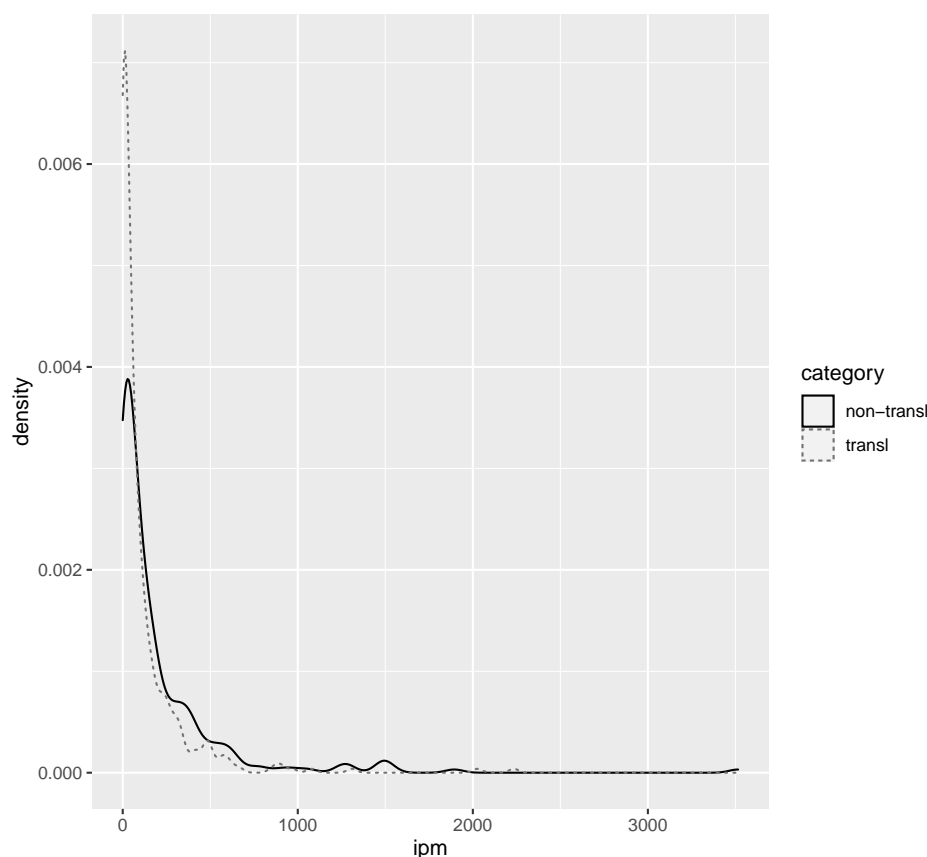


Figure 3.8: Jerome translated vs. non-translated density plot (fiction)

corpus only, in Table 3.6.

All the differences between translated and non-translated texts observed in Table 3.6 are statistically significant ($p > .0001$), and the comparison of Figures 3.9 and 3.10 demonstrates that the difference is more pronounced in Conv.pt.pf (Figure 3.10) than in Conv.ps.impf (Figure 3.9).

The greater tendency to normalisation of Conv.pt.pf is due to the more important stylistic mark of this form, in comparison with Conv.ps.impf. We can recall that Conv.ps.impf is considered bookish, whereas the Conv.pt.pf is assigned an archaistic stylistic mark. Since translators normalise, it is natural that they tend to avoid the form manifesting a stronger stylistic mark.

Olga Nádvořníková

Table 3.6: Frequency of the transgressive (present and past form) in Jerome (fiction) (n = absolute frequency, f = normalised frequency in instances per million words, CV = coefficient of variation)

Jerome corpus	form	n	f	SD	CV
translated	Conv.ps.impf	2,376	108.92	101.97	93.62
non-translated	Conv.ps.impf	2,441	155.55	282.66	181.72
translated	Conv.pt.pf	162	6.95	39.74	571.80
non-translated	Conv.pt.pf	354	22.56	118.03	523.18

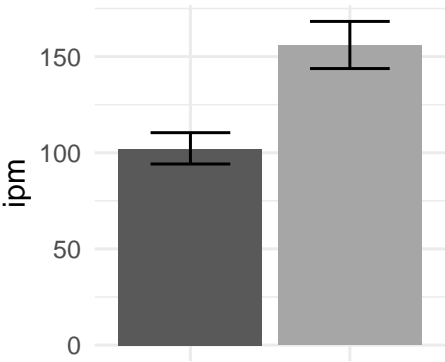


Figure 3.9: Jerome translated vs. non-translated (fiction) frequency of Conv.ps.impf

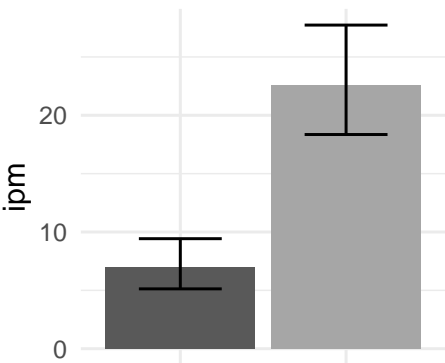


Figure 3.10: Jerome transl vs. non-transl (fiction) frequency of Conv.pt.pf

4.2 Cross-linguistic interference in translations

Since normalisation is considered a universal phenomenon, in §4.1 we analysed its potential effect in translations for all the source languages together. Conversely, cross-linguistic interference is intrinsically related to the linguistic properties of the different source languages. Concerning converbs, hypothesis H_1 expects more transgressives in translations from Romance and Slavic languages than in translations from Germanic languages. As introduced in §3.1.1, we conduct this analysis on the fiction part of the InterCorp parallel corpus (texts published after 1992 including), which contains a larger amount of texts than the Jerome comparable corpus and the non-fiction sub-corpus of InterCorp.

Table 3.7 shows the absolute and the normalised frequencies of the transgressive (both forms together) in translations from the 20 source languages available in the sub-corpus of fiction translated into Czech in the InterCorp parallel corpus. At first sight, the results confirm the H_1 , since Slavic and Romance source languages are grouped at the top of the frequency list (except for Italian in Romance and Slovenian in Slavic), whereas the Germanic languages are found mostly in the lower part of the table (except for German, which is ranked 9 in the table). English, considered exceptional among the other Germanic languages, is found in the middle of the list. It is important to note that only eight source language sub-corpora show a normalised frequency higher than 140.54 ipm, i.e. the frequency in the reference non-translated fiction corpus in SYNv8 (see Table 3.4). This confirms the tendency to normalisation observed in 4.1.

However, upon closer examination, the results introduced in Table 3.7 appear much less reliable. For instance, it is true that within the group of Romance languages, the lower frequency of the transgressive in translations from Italian may be explained by the lower frequency of the Italian converb (*gerundio*, see Čermák et al. 2020) in comparison with Portuguese and Spanish (ranking 2nd and 5th). The French *gérondif*, however, is even less frequent than the Italian *gerundio* (ibid.), but translations from French rank 6th, just after Spanish. This brief observation reveals the first methodological pitfall of the analysis of the potential effect of cross-linguistic interference based only on frequencies: without understanding the *valeur* of the converb in the system of the source language and without a detailed analysis of parallel concordances in the individual language pairs, all the cross-linguistic observations are potentially unreliable.

Similarly, a closer look at the group of Slavic languages reveals other discrepancies of the purely frequential approach to the cross-linguistic interference. Polish, for instance, using its two converb forms extensively, is likely to be found at the top of the list, which is the case in Table 3.6. However, the position of Russian in

Olga Nádvořníková

Table 3.7: Frequency of the transgressive (present and past form) in different source language sub-corpora of InterCorp (fiction)

Rank	src.lang	positions (n)	texts (n)	abs.fq.	ipm
1	pl	2,436,840	35	891	365.64
2	pt	1,250,080	16	398	318.38
3	sr	366,940	6	108	294.33
4	ro	372,404	5	95	255.10
5	es	8,393,499	101	1,762	209.92
6	fr	5,009,729	73	988	197.22
7	hr	1,242,178	19	209	168.25
8	sk	994,572	16	165	165.90
9	de	8,920,552	91	1,154	129.36
10	ru	1,306,704	11	154	117.85
11	en	25,810,495	226	2,597	100.62
12	it	1,044,540	14	103	98.61
13	fi	1,355,134	23	124	91.50
14	nl	1,657,687	23	151	91.09
15	lv	228,997	5	17	74.24
16	sl	835,792	11	37	44.27
17	da	1,023,334	9	44	43.00
18	sv	6,604,972	69	207	31.34
19	no	1,498,553	16	46	30.70
20	ja	710,938	5	18	25.32
	total	71,063,940	774	9,268	130.42

Table 3.6 is surprising: even though its converb is considered prototypical (see §2.1.1) and its two converb forms are well attested, Russian only ranks 10th, even after Slovak, making only very limited use of its converb (see §2.1.1 and Brtková (2004: 25)). By its ranking, Russian is placed even below German, yet considered making only “parsimonious” use of converbs (see §2.1.1 and König (1995: 72)). Similarly, polyconverb Finnish, Latvian and Japanese surprisingly only rank 13th, 15th and even 20th.

The reliability of the results for the different language sub-corpora introduced in Table 3.7 is undermined by the same (external) factors as in the analysis of normalisation: the frequency of the transgressive may be influenced by the specific style and topic of the text, by the individual preferences of the translators, and

even by the date of publication of the source text. Moreover, since the corpus is divided into 20 sub-corpora, the risk of systematic bias is higher than in the normalisation testing. For example, in the small subcorpus of translations from Slovak, we find two fantasy novels showing very high frequencies of the transgressive, which may influence the results for the whole sub-corpus, containing only 16 texts. Similarly, the normalised frequency of the transgressive in translations from Portuguese is skewed by one translation of a text first published in the 19th century (Eça de Queiros' novel *A Cidade e as Serras*) and showing the normalised frequency of the transgressive more than 12 times higher than in the reference corpus SYNv8. In the subcorpus of translations from Romanian, it is not possible to say whether the sub-corpus reflects cross-linguistic interference or the personal preferences of the translator because all the five texts in this sub-corpus were translated by the same translator (Jiří Našinec).

Figure 3.11 summarises the tendencies in the frequency of the transgressive and the limitations of the reliability of the data extracted from our corpus (the confidence intervals).

We can see that for Danish, Japanese, Norwegian and Serbian, the data extracted from our corpus is not reliable. The rest of the data confirms the tendencies observed in Table 3.7, i.e. a higher frequency of transgressives in translations from Slavic and Romance languages (except for Slovenian, and partly Slovak and Russian) and a lower frequency in translations from Germanic languages.

Nevertheless, the analysis of the potential effect of the cross-linguistic interference between the converb in the source language and the Czech transgressive necessitates a thorough contrastive examination of individual language pairs. Subsequently, there needs to be a detailed analysis of the occurrences in parallel concordances, which takes into account the linguistic factors of the use of the transgressive (and its counterpart/s in the source language), and the potential influence of the style of the text, the translators' idiolects and other factors.

5 Conclusion

The Czech transgressive is a specific case of the cross-linguistic category of converb. On the one hand, it shows most properties of the prototypical converbs: it is strict, has two forms (present and past transgressive), is referentially same-subject (i.e. coreferential with the controller of the main clause) and, as with most European converbs, its semantic interpretation is contextual (with the prevailing meaning of accompanying circumstance). On the other hand, it has an archaistic morphology, requiring agreement with the controller in number and gender and

Olga Nádvorníková

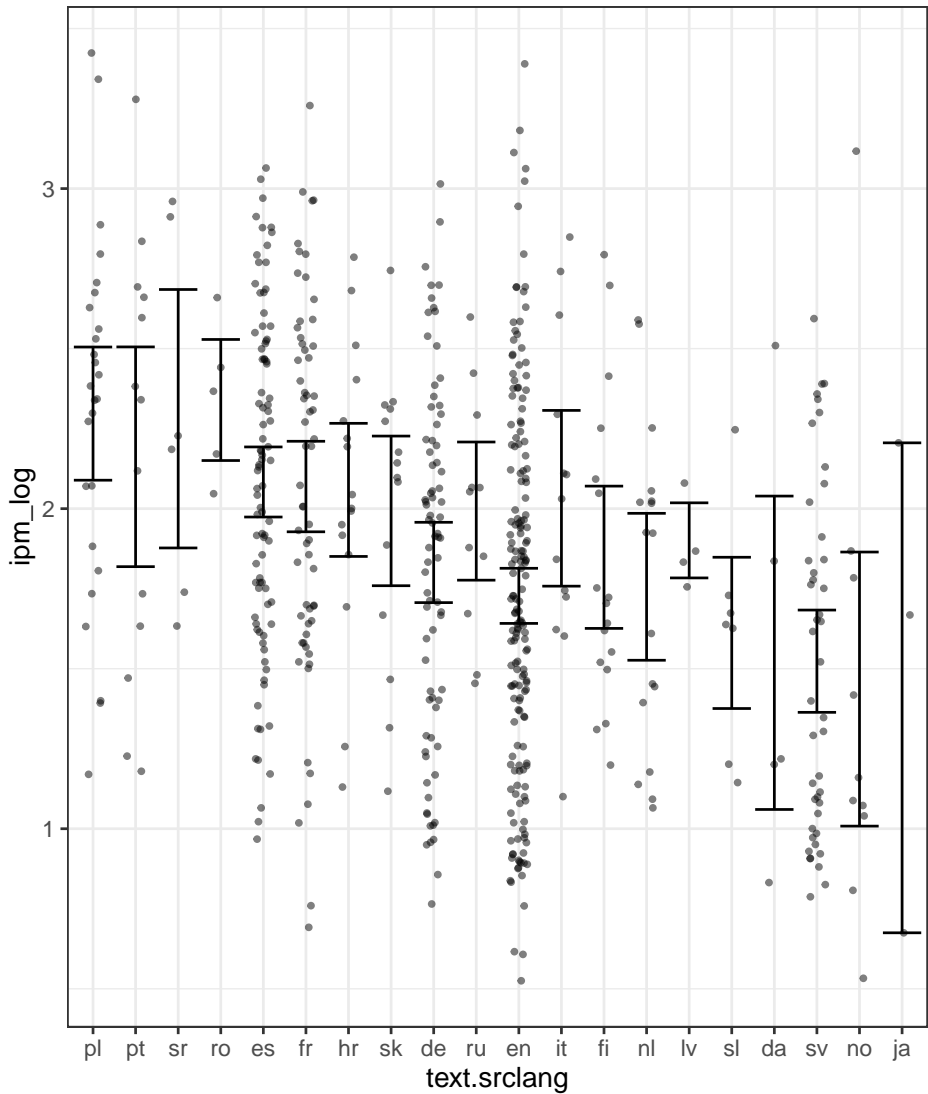


Figure 3.11: Normalised frequency of the transgressive in 20 different source language sub-corpora in the InterCorp corpus (fiction)

a strong stylistic mark: bookish for the present transgressive and archaistic for the past transgressive. Because of this stylistic mark, the transgressive is used rarely in contemporary language, and only in written texts.

In this study, we investigated the potential impact of these double-face characteristics of the Czech converb on translations of fiction and non-fiction in Czech.

Our preliminary frequential analysis confirmed the constant decrease in the frequency of the transgressives in both text registers and both translated and non-translated texts during the 20th and 21st centuries. This observation also justified the limitation of our corpora to texts published after 1992 (inclusive). In line with expectations, the frequency analysis revealed the strong dominance of the present transgressive over the past form, which corroborates the diachronic trend suggested for Czech by Nedjalkov (1995): Czech appears to be moving from a bi-converbal language to a mono-converbal one.

The main finding of our study is the confirmation of the normalisation effect in translations of fiction (but not in those of non-fiction), the absence of convergence in translations in comparison with non-translated texts, and the necessity of a thorough contrastive analysis of converbs before investigating the potential effect of the cross-linguistic interference.

As for the normalisation, the difference in the frequency of the transgressive between translated and non-translated fiction is not extensive but is statistically significant. Of greater interest, a detailed analysis of the distribution of the frequencies revealed that this normalisation effect is caused especially by the number of texts using zero transgressives: in translations 31%, in non-translated texts only 20% of the texts. This means that more translators decided to avoid transgressives than the authors of the original texts. Finally, the normalisation impact is stronger in the past transgressive, showing a stronger stylistic mark, than in the present transgressive. These results suggest that if translators decided to use more transgressives – with respect to the style of the source text, of course – they would not violate the norm of the target language.

In non-fiction, the effect of normalisation was not observed. This text-register difference may be explained either by the overall lower frequency of the transgressive in non-fiction than in fiction or precisely by the stylistic mark of the transgressive. In fiction, the authors and translators appear to exploit this characteristic of the transgressive, e.g. the use as a means of irony or parody (mainly in non-translated texts), as the reflection of a specific, very complex style and syntax of the source text in translations or to create the archaistic effect in historical novels or in fantasy stories. This last use was also observed in translations of source texts first published in the 19th century, even if the actual translation was recent. In non-fiction, the use of the transgressive appears to be governed

Olga Nádvorníková

not by the individual style of the text or the author, but by the norms of the text register sub-types. In line with observations in previous studies, the transgressive is more frequent in humanities (philosophy, history, literary studies, etc.) than in natural and technical sciences. Nevertheless, all these observations require a more thorough analysis of individual texts and concrete occurrences of transgressives in context.

Pertaining to the convergence hypothesis, based on the analysis of the coefficient of variation, it was observed neither in fiction nor in non-fiction. This means that translations are as heterogeneous in the frequency of the transgressive as non-translated texts. However, both in translations and in non-translated texts, the coefficient of variation is higher in the past form of the transgressive, considered archaistic, than in the present form, considered only bookish. This result indicates that the effect of convergence may vary according to the stylistic mark of the linguistic feature under investigation.

The results for the cross-linguistic hypothesis are the least conclusive. The comparison of the normalised frequency of the transgressive in twenty source language subcorpora showed a higher frequency of transgressives in translations from Slavic and Romance languages, where the converbs are considered prototypical, and a lower frequency in translations from Germanic languages, supposedly to make very limited use of converb (except for English). However, several partial results were not consistent with the hypotheses. In the Slavic languages, for instance, translations from Slovak show a higher frequency of transgressives than translations from Russian, although converbs in Slovak are rare but abundant in Russian. Similarly, translations from French contain more transgressives than those from Italian despite the much lower frequency of the French *gérondif* than the Italian *gerundio*.

These inconsistencies reveal two important pitfalls of the purely frequential analysis of the cross-linguistic interference effect in translations. First, since the use of the transgressive is intrinsically linked to its stylistic mark, the results are extremely sensitive to the composition of the different source language subcorpora and the style of the texts they contain. Second, and more importantly, these results reveal the necessity of a thorough contrastive analysis of the different language pairs, taking into account the frequency and the *valeur* of the different converbs in the language systems, and their specific uses in context.

Future research may provide not only a more fine-grained contrastive analysis of converbs in different language pairs but also a deeper understanding of the motivations of the normalisation and convergence in translation and various factors coming into play in the process of translation and the translation workflow. It is worth investigating, for instance, the potential effect of the translator's

proficiency (do experienced translators use the transgressive more than translators in the early stage of their career? What is the role of translators' training in their attitude to the transgressive? cf. [Lapshinova-Koltunski 2018](#)), the sex of the translator (preliminary results indicate female translators use transgressives less than their male colleagues; see the impact of the gender factor in [Magnifico & Defrancq 2018](#)), the target audience (is there a difference between translated and non-translated literature intended for children and young readers, with regard to the use of transgressives? cf. e.g. [Čermáková 2017](#)), and the attitude of text revisers in publishing houses to the transgressive and the impact of their interventions on its frequency in (translated as well as non-translated) texts (see also [Bisiada 2017; 2018; 2019; Kruger 2018](#)). Only this complex approach may help to fully conceive translation as a socially contexted behaviour and understand the norms to which the translator is supposed to adhere to.

Acknowledgements

I would like to express my gratitude to Adrian Zasina and Martin Vavřín from the Institute of the Czech National Corpus, for providing me with the data necessary for the research, and to Tomáš Bořil from the Institute of Phonetics of the Faculty of Arts of Charles University in Prague, for the statistical analysis of my data in R.

This work was supported by the European Regional Development Fund-Project “Creativity and Adaptability as Conditions of the Success of Europe in an Inter-related World” (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

This research was supported by the Charles University project Progres Q10, Language in the shiftings of time, space, and culture.

References

- Alpatov, Vladimir M. & Vera Podlesskaya. 1995. Converbs in Japanese. In Martin Haspelmath & Ekkehard König (eds.), *Converbs in cross-linguistic perspective: Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, 465–487. Berlin: Mouton de Gruyter.
- Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Olga Nádvorníková

- Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*, 175–186. Amsterdam: John Benjamins.
- Bečka, Josef Václav. 1977. Přechodníky jako překladatelský problém [Transgressives as a translation problem]. *Zprávy Kruhu přátel českého jazyka* 5. 17–27.
- Bisiada, Mario. 2017. Universals of editing and translation. In Silvia Hansen-Schirra, Oliver Czulo & Sascha Hofmann (eds.), *Empirical modelling of translation and interpreting*, 241–275. Berlin: Language Science Press.
- Bisiada, Mario. 2018. The editor's invisibility: Analysing editorial intervention in translation. *Target* 30(2). 288–309.
- Bisiada, Mario. 2019. Translated language or edited language? A study of passive constructions in translation manuscripts and their published versions. *Across Languages and Cultures* 20(1). 35–56.
- Brtková, Kamila. 2004. K otázce prekladu ruských prechodníkov do slovenčiny [On translation of Russian transgressives into Slovak]. *Opera Slavica* XIV(2). 20–31.
- Cappelle, Bert. 2012. English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures* 13(2). 173–195.
- Čechová, Marie, Jan Chloupek & Eva Minářová. 1997. *Stylistika současné češtiny* [Stylistics of contemporary Czech]. Praha: ISV – nakladatelství.
- Čermák, František & Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3). 411–427.
- Čermák, Petr, Dana Kratochvílová, Olga Nádvorníková & Pavel Štichauer. 2020. *Complex words, causatives, verbal periphrases and the gerund: Romance languages versus Czech (a parallel corpus-based study)*. Praha: Karolinum.
- Čermák, Petr, Olga Nádvorníková, Leontýna Bratánková, Štěpánka Černíková, Jan Hricsina, Jiří Jančík, Jaroslava Jindrová, Dana Kratochvílová, Zuzana Krinková, Petra Laufková, Daniel Petřík, Eliška Třísková & Pavel Štichauer. 2015. *Románské jazyky a čeština ve světle paralelních korpusů* [Romance languages and Czech in the light of parallel corpora]. Praha: Karolinum.
- Čermáková, Anna. 2017. Translating children's literature: Some insights from corpus stylistics. *Ilha do Desterro – A Journal of English Language, Literatures in English and Cultural Studies* 71(1). 117–134.
- Chesterman, Andrew. 2004. Beyond the particular. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals – do they exist?*, 33–49. Amsterdam: John Benjamins.

- Chlumská, Lucie. 2013. *Jerome corpus – comparable corpus of translated and non-translated Czech*. Praha: Institute of the Czech National Corpus. <https://kontext.korpus.cz>.
- Chlumská, Lucie. 2017. *Překládová čeština a její charakteristiky [Translated Czech and its features]*. Praha: Nakladatelství Lidové noviny.
- Cvrček, Václav. 2010. *Mluvnice současné češtiny 1 [Grammar of contemporary Czech]*. Praha: Karolinum.
- Cvrček, Václav & Lucie Chlumská. 2015. Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics* 39(3). 309–325.
- Cvrček, Václav & Dominika Kovářiková. 2011. Možnosti a meze korpusové lingvistiky [Opportunities and limitations of corpus linguistics]. *Naše řeč* 94(3). 113–133.
- Dai, Guangrong & Richard Xiao. 2011. SL “shining through” in translational language: A corpus-based study of Chinese translation of English passive. *Translation Quarterly* 62. 85–108.
- Delaere, Isabelle, Gert de Sutter & Koen Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target* 24(2). 203–224.
- Dvonč, Ladislav, Gejza Horák, František Miko, Jozef Mistrík, Ján Oravec, Jozef Ružička & Milan Urbančok. 1966. *Morfológia slovenského jazyka [Morphology of the Slovak language]*. Bratislava: Vydavateľstvo Slovenskej akadémie vied.
- Dvořák, Emil. 1970. *Vývoj přechodníkových konstrukcí ve starší češtině [Evolution of transgressive constructions in old Czech]*. Praha: Universita Karlova.
- Dvořák, Emil. 1972. Přechodníkové konstrukce v překladech beletrie do češtiny [Transgressive constructions in translation of fiction in Czech]. *Slavica Pragensia* 14(2–4). 101–114.
- Dvořák, Emil. 1983. *Přechodníkové konstrukce v nové češtině [transgressive constructions in modern Czech]*. Praha: Univerzita Karlova.
- Grevisse, Maurice & André Goosse. 2016. *Le bon usage*. Louvain-la-Neuve: De Boeck Supérieur.
- Halmøy, Odile. 2003. *Le gérondif en français*. Paris: Ophrys.
- Haspelmath, Martin. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath & Ekkehard König (eds.), *Converbs in cross-linguistic perspective: Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, 1–57. Berlin: Mouton de Gruyter.
- Hnátková, Milena, Michal Křen, Pavel Procházka & Hana Skoumalová. 2014. The SYN-series corpora of written Czech. 160–164.

Olga Nádvorníková

- Hnátková, Milena, Vladimír Petkevič & Hana Skoumalová. 2011. Linguistic annotation of corpora in the Czech National Corpus. In Viktor Zacharov (ed.), *Proceedings of the international conference "Corpus Linguistics – 2011"*, 15–20. St. Petersburg: St. Petersburg State University.
- Jelínek, Jaroslav, Josef Václav Bečka & Marie Těšitelová. 1961. *Frekvence slov, slovních druhů a tvarů v českém jazyce* [Frequency of words, word classes and word forms in Czech]. Praha: Státní pedagogické nakladatelství.
- Karlík, Petr, Marek Nekula & Zdenka Rusínová. 1995. *Příruční mluvnice češtiny* [Reference grammar of Czech]. Praha: Nakladatelství Lidové noviny.
- Karlík, Petr. 2017. *Přechodník* [Transgressive]. <https://www.czechency.org/slovník/P%C5%98ECHODN%C3%8DK> (12 August, 2020).
- Kleiber, Georges. 2007. La question temporelle du gérondif: Simultanéité ou non ? In Frédéric Lambert, Catherine Moreau & Jean Albrespit (eds.), *Les formes non finies du verbe 2*, 109–123. Rennes: Presses Universitaires de Rennes.
- Kleiber, Georges. 2009. Gérondif et relations de cohérence: Le cas de la relation de cause. In Elena Comes & Florica Hrubaru (eds.), *Relations de discours II – actes du XVe séminaire de didactique universitaire*, 9–24. Cluj: Editura Echinox.
- Kocková, Jana. 2011. Ekvivalenty ruských přechodníků sloves dokonavého vidu v češtině na základě paralelních korpusů [Equivalents of Russian perfective transgressives in Czech: A corpus-based study]. In František Čermák (ed.), *Korpusová lingvistika praha*, 251–261. Praha: NLN.
- Komárek, Miroslav. 1986. *Mluvnice češtiny 2* [Grammar of Czech]. Praha: Academia.
- König, Ekkehard. 1995. The meaning of converb constructions. In Martin Haspelmath & Ekkehard König (eds.), *Converbs in cross-linguistic perspective: Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, 57–97. Berlin: Mouton de Gruyter.
- König, Ekkehard & Johan van der Auwera. 1990. Adverbial participles, gerunds and absolute constructions in the languages of Europe. In Johannes Bechert, Giuliano Bernini & Claude Buridant (eds.), *Toward a typology of European languages*, 57–95. Berlin: Mouton de Gruyter.
- Kortmann, Bernd. 1997. *Adverbial subordination: A typology and history of adverbial subordinators based on European languages*. Berlin: Mouton de Gruyter.
- Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kovářiková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondříčka & Adrian Zasina. 2019. *Korpus SYN, version 8 from 12/12/2019*. Praha: Institute of the Czech National Corpus. <https://kontext.korpus.cz/>.

- Kruger, Haidee. 2018. The effects of editorial intervention: Implications for studies of the features of translated language. In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies*, 113–157. Berlin: Mouton de Gruyter.
- Lapshinova-Koltunski, Ekaterina. 2015. Variation in translation: Evidence from corpora. In Claudio Fantinoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies*, 93–114. Berlin: Language Science Press.
- Lapshinova-Koltunski, Ekaterina. 2018. Text classification for detection of translationese in novice and professional translations. In Sylviane Granger, Marie-Aude Lefer & Laura Aguiar de Souza Penha Marion (eds.), *Book of abstracts: Using corpora in contrastive and translation studies conference*, 108–110. Louvain: CECL/Université catholique de Louvain.
- Laviosa, Sara. 2002. *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.
- Lefer, Marie-Aude & Svetlana Vogeleer. 2013. Interference and normalization in genre-controlled multilingual corpora. *Belgian Journal of Linguistics* 27. 1–21.
- Levý, Jiří. 2011. *The art of translation*. Amsterdam: John Benjamins.
- Magnifico, Cédric & Bart Defrancq. 2018. Norms and gender in simultaneous interpreting: A study of self-repairs. In Sylviane Granger, Marie-Aude Lefer & Laura Aguiar de Souza Penha Marion (eds.), *Book of abstracts: Using corpora in contrastive and translation studies conference*, 119–120. Louvain: CECL/Université catholique de Louvain.
- Malá, Markéta & Pavlína Šaldová. 2015. English non-finite participial clauses as seen through their Czech counterparts. *Nordic Journal of English Studies* 1. 232–257.
- Michálková, Věra. 1963. K polovětným konstrukcím v nářečí [On semi-predicative constructions in dialects]. *Naše řeč* 46(3). 136–145.
- Moortgat, Bernadette. 1978. *Participe et gérondif: Étude de l'opposition entre la présence et l'absence de EN devant la forme en -ant*. Université de Metz: Thèse de 3e cycle.
- Nádvorníková, Olga. 2016. Le corpus multilingue InterCorp et les possibilités de son exploitation. In Éva Buchi, Jean-Paul Chauveau & Jean-Marie Pierrel (eds.), *Actes du XXVIIe congrès international de linguistique et de philologie romanes (Nancy, 15–20 juillet 2013)*, 223–237. Strasbourg: Société de linguistique romane/ÉLiPhi.
- Nádvorníková, Olga. 2017. Parallel corpus in translation studies: Analysis of shifts in the segmentation of sentences in the Czech-English-French part of the InterCorp parallel corpus. In Joseph Emonds & Markéta Janebová (eds.), *Language use and linguistic structure*, 445–461. Olomouc: Palacký University Olomouc.

Olga Nádvorníková

- Nádvorníková, Olga. 2012. *Korpusová analýza faktorů sémantické interpretace francouzského gérondivu* [A corpus-based analysis of factors influencing semantic interpretation of the French gérondif]. Univerzita Karlova v Praze. (Doctoral dissertation).
- Nedjalkov, Igor V. 1995. Some typological parameters of converbs. In Martin Haspelmath & Ekkehard König (eds.), *Converbs in cross-linguistic perspective: Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, 97–137. Berlin: Mouton de Gruyter.
- Nedjalkov, Igor V. 1998. Converbs in the languages of Europe. In Johan van der Auwera (ed.), *Adverbial constructions in the languages of Europe*, 420–455. Berlin: Mouton de Gruyter.
- Nosek, Jiří. 1964. Notes on syntactic condensation in modern English. *Travaux linguistiques de Prague* 1. 281–288.
- Oktábec, Vojtěch. 1953. O překládání ruských přechodníkových vazeb [On translation of Russian transgressive constructions]. In Jaroslav Moravec (ed.), *Kniha o překládání: Příspěvky k otázkám překladu z ruštiny*, 258–279. Praha: Nakladatelství Československo-sovětského institutu.
- Rosen, Alexandr, Martin Vavříň & Adrian Zasina. 2019. *InterCorp multilingual corpus – Czech, version 12 from 12/12/2019*. Praha: Institute of the Czech National Corpus. <https://kontext.korpus.cz/>.
- Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin: Mouton De Gruyter.
- Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.
- Vachek, Josef. 1955. Some thoughts on the so-called complex condensation in modern English. *Sborník prací Filozofické fakulty Brněnské univerzity* 4. 63–77.

Chapter 4

The internationalized text and its localized variations: A parallel analysis of blurbs localized from English into Arabic and French

Madiha Kassawat

Université Sorbonne Nouvelle (ESIT)

In an increasingly globalized world, accessibility to digital content has become indispensable for people around the world. This accessibility would not be possible without translation which plays an important role in linguistic and cultural mediation, as well as in marketing. As the majority of products is promoted for and sold on the internet, their web pages are often localized based on the market, including the language and the culture. The required speed in this type of work, its tools and process play a remarkable role which influences the quality of the localized texts. Therefore, it is necessary to analyze these texts, explore the different interpretations of a text in several languages and cultures, and the adaptation level which should convince the consumer to purchase the product. This pilot study is an attempt to compare the product descriptions provided in English and localized into Arabic and several French versions. The results show the relationship between the international text and the localized texts on the linguistic and cultural levels.

1 Introduction

We are living in a globalized world where translation is a daily-lived practice (Ladmiral 2014). The majority of translated texts nowadays are not literary but utilitarian (Le Dizez 2004). At the same time, the digital environment is considered a decisive channel of marketing. It absorbs increasingly between 20 to 30



Madiha Kassawat

percent of advertising expenses (Stenger & Bourliataux-Lajoinie 2014). From this perspective, localization plays an important role in enriching the digital content, particularly websites which are used for different purposes. This role is manifested by the various linguistic versions of the same product whose target audience is large and diversified. Such a big audience makes the localized product more visible to the public and more accessible on the internet than the printed products (Jiménez-Crespo 2013). This, in turn, makes the product (the website in this case) a cultural object which conveys cultural markers (Rémon 2005).

Localization started in the 1990s when software localization was most common, including the software content and its printed or online help text (Esselink 2003). The change and use of websites have resulted in the idea of content localization which focuses on the linguistic skills more than on the technical ones (Esselink 2003). In addition, the year 1995 witnessed the first official advertisements on commercial websites of corporates such as Amazon, eBay and Yahoo! (Stenger & Bourliataux-Lajoinie 2014). Digital marketing has become a popular tendency that the majority of international corporations have adopted since. It is even considered an indispensable strategy for product distribution at the global level.

First, this paper will attempt to discuss the *locale* as an elastic term which can represent a country or a territory, or a cluster of countries or territories. This should explain the necessity of cultural considerations in the target market as specified by the client. Second, the process of website localization will be briefly explained, while highlighting the internationalization phase and its goals regarding the adaptation of the localized text. Third, a brief discussion will take place on the applied theories of “traditional” advertising translation. This subject will be necessary to reframe these theories in a relatively recent industry such as localization. The paper will attempt to answer the question of whether and how the localized text is adapted for the different cultures. It examines adaptation through comparing how the linguistic and cultural elements have been treated in the Arabic and French versions, taking the international the English text as a reference.

2 Localization: A locale and a process

2.1 The amalgamation of the locale

The *locale* is a term used in the industry. It combines a language variety and cultural norms using the market criteria to resolve contradictions between sociolinguistic levels (Pym 2005). These criteria can include language, currency, and

the consumers' education level or their revenues, based on the communication nature (Pym 2011). By looking at the international corporate website addresses, each one has an identifier that is usually a combination of the country/region and the language. For example, the code of a website whose target country is France can include fr-fr in its URL, and an Arabic site for Egypt can include eg-ar, and en-uk for a site targeting the UK, etc.

However, this strategy is not always as specific or organized as it appears. Not all the countries have the same opportunity of being offered a culturally adapted website if they do not represent significant markets. This is where amalgamation of the different geographic denominations happens instead of using country codes. For example, Jiménez-Crespo (2010) found that the main target hispanic markets are Spain (42%), Mexico (32%), the United States (27%) and Argentina (27%). This tendency of generalizing the communication language of different cultural communities highlights the marketing approach which focuses on the "languages of consumption" (Pym 2000) instead of languages belonging to specific cultures. In addition, the geography that is associated to language is unspecific (Guidère 2000: 26–27). This geography is supposed to specify linguistic, cultural and economic peculiarities (Guidère 2000: 29).

2.2 From internationalization to localization

Localization is a necessary process for immigrating information to other sites, where languages other than the original language of the content are used (Cronin 2006: 28–29). A typical localization project passes through three phases: the project preparation, translation and quality assurance (Quah 2006: 114). However, *localization* is usually used as a general term without mentioning the general life cycle of the product. Localization is one of the GILT phases (Globalization, Internationalization, Localization and Translation) (Munday 2008). Globalization is the encompassing cycle of the product, where internationalization includes planning and preparing the product, while localization is the actual adaptation of the product for its target market (Anastasiou & Schäler 2010). As for internationalization, it is about adapting the products to facilitate their localization in the international markets (Esselink 2003). The central aspect of internationalization, as pointed out by Esselink (2000), is displaying the characters according to the local standards of the target locale. For example, double-byte character compatibility should be provided prior to the product translation (Esselink 2000: 3). In other words, the product should be "enabled" in order to be usable in certain countries and regions (Esselink 1998: 2). More generally, this phase necessitates removing

Madiha Kassawat

any cultural, linguistic, technical, religious, philosophical, value-related peculiarities from the product and its complements (Gouadec 2003).

This extraction of the language- and culture-dependent elements during the internationalization phase has also been highlighted by (Schäler 2007). Hence, internationalization is not only applied to the technical aspects of the product but also to its textual content. The internationalized text should facilitate the transfer to the maximum number of languages without producing any complications (Jiménez-Crespo 2013: 26). This phase includes the source text pre-editing as well. It is used as a form of quality assurance and limiting the cost to meet the need for translating into several languages. In his discussion of the notion of the internationalized text, Pym uses the term “one-to-many geometry” versus “language-into-language situations” which is adopted in translation to refer to the source text and the target text (2006). This strategy is adopted particularly with the product launch simultaneously or successively, in several languages and multiple countries, and within a very short time around (Quah 2006: 45).

Having internationalized the text, the localization phase consists of the linguistic and cultural adaptation of the text in order to distribute the digital products and services independently of the characteristics of the original country (Schäler 2007). The Localization Industry Standards Association, which was deprecated in 2011, provided a definition of localization that “involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used and sold” (LISA 2003: 13, in Jiménez-Crespo 2013), (cf. Yunker 2003). The Globalization and Localization Association (GALA) explains that “[T]he aim of localization is to give a product the look and feel of having been created specifically for a target market, no matter their language, culture, or location” (GALA 2019).

Localization is sometimes viewed as a practice that goes beyond translation due to the adaptation to the culture of the target text (Anastasiou & Schäler 2010), and to the fact that it includes technical aspects in addition to the “traditional translation” tasks (Austermühl 2006). However, adaptation is required in translation (Nord 2005). This is underlined in another definition of localization which shows that it is a type of functionalist translation whose goal is the communicative purpose:

Localization is therefore conceptualized as a target-oriented translation type and, in line with the functionalist notion of adequacy, emphasizes users’ expectations and achieving the communicative purpose for which the localization was commissioned, rather than equivalence relationships to source texts (STs). (Jiménez-Crespo 2013: 18)

Despite the importance of the technical aspects in localization, whether on the language agents' or engineers' side (concerning localization agents, cf. [Canum Alkan 2017](#)), this paper focuses on the linguistic and cultural adaptation in the localized text and how its function is treated in the frame of the aforementioned definitions of localization.

2.3 From an international to a local text: contradictory phases

It is important to point out the contradiction between the product globalization phases. The process starts with internationalization and filtering the cultural references of a product in order to make it look locally made. This area has been highlighted by Jiménez-Crespo who hints to this contradiction in the localization industry discourse. On the one hand, localization aims to make websites give the impression that they have been created in the target country. On the other hand, internationalization neutralizes the products in terms of language and culture ([Jiménez-Crespo 2010](#)). He finds that internationalizing a communication has direct consequences on the languages and the translation process itself ([Jiménez-Crespo 2013](#): 10). Therefore, it is intriguing to analyze the potential internationalization impact on the localized product, on the adaptation level in particular.

3 Advertising translation: Which function in localization?

3.1 The message effect

The localization definitions discussed earlier show that the text function is essential. It has a direct relationship with the communicative purpose of the products and the inter-linguistic and intercultural approach of marketing. In looking at the text as a marketing tool for a product, the appellative intention should be the keyword to persuade the recipient to adopt a certain opinion or perform a certain activity ([Nord 2005](#); cf. [Tatilon 1990](#); [Boivineau 1972](#)). In the case of advertising translation for example, the persuasive function of the message is pivotal and the distinction between the source text and the translated one becomes difficult ([Cruz-García 2018](#)) due to its reformulation. This intention should be well explained in the “brief” as it will be received and read before reading the source text ([Nord 2005](#)). Moreover, it is the client's desired effect which determines the translation strategy. As put by [Ladmiral \(2014: 76\)](#), the “sourciers” are those who translate based on the source text, while the “ciblistes” concentrate on the message and the effect to be translated. For the latter, they make use of

Madiha Kassawat

all the available tools and ways of the target language. Such an effect is only achieved through extracting all the elements which can shock the consumer regarding his/her beliefs, feelings, traditions, attitudes, customs and anything related to his/her cultural package [Tatilon \(1990\)](#). The advert effect has been studied by [Gully \(1996\)](#) who analyzed Egyptian Arabic advertisements of TV, radio and magazines. He found different strategies used for persuasion such as the use of metaphors, rhetorical expressions and the local dialect. Having said that, persuasion can be achieved through the use of references from the target language and culture.

4 The translator-localizer and the audience... cultural packages

Translation strategies differ based on the guidelines, and can vary more depending on the translator's preferences and knowledge as a first reader of the text ([Plassard 2007](#)). According to [Munday \(2008\)](#), the interpretative theory in translation identifies three phases of the process: comprehension, de-verbalization and re-expression. The result is an association of the linguistic and non-linguistic sets of knowledge. Therefore, this association can be understood as a cube: it can generate multiple possibilities of translations of the same text. This can be complemented by the fact that language is related to culture. In the Onion Model ([Hofstede et al. 2010](#)), culture consists of two layers. The first represents practices (symbols, heroes and rituals). The second represent values. Although words exist at the surface of culture, within symbols, they are considered the vehicles of cultural transfer ([Hofstede et al. 2010](#)). Hence, the translator's cultural package has an unavoidable impact on the translation, which is the final result that encompasses the linguistic and cultural knowledge.

The language-culture combination is done by the translator-localizer in this case. At the same time, readers (users of the localized products) "who want to 'understand', have to connect or associate the new information given by the text with the knowledge of the world already stored in their memories" ([Nord 2005: 96](#)). Here comes the importance of adapting the information which can be "trivial" for the source text recipients, depending on their own cultural package, but can be unknown to the target text audience ([Nord 2005: 107](#)). In other words, there is a need to fill in the gaps that exist in the recipient's knowledge ([Baker 2011](#)).

This adaptation necessitates a transformation, a mediation and a change ([Maitland 2017: 159](#)) as "the articulation of another's experience in one's own words re-

quires the importation of other ideas, other viewpoints, other worldviews” (Maitland 2017: 07). Moreover, and although a person belongs to a specific culture, they do not know all the aspects of that culture (Gudykunst 2004: 42). Their point of view will definitely differ from that of another. Such differences in viewing and understanding the world lead to multiple variations and possibilities in the translation.

5 The loyalty to the translation and the purpose of the product

Coming back to localization, adaptation is sometimes considered an additional element introduced by localization, as opposed to the literal nature of translation: “Adaptations are seen as the additional component that localization provides, as opposed to the textual or wordly nature of ‘translation’. The term *adaptation* is typically used to indicate the performative action of the localization process” (Jiménez-Crespo 2013: 15).

Nonetheless, adaptation is just a modification procedure, besides transposition, which tends towards the literality, and re-writing (Guidère 2000). As for adaptation, it can be formal, where it affects the structure of the original statement, or idea-based to meet the cultural expectations of the target recipients (Guidère 2000: 124). On the contrary, re-writing tends to provide a different expressive orientation to the message initial idea (Guidère 2000: 129). It should be indicated here that the term *re-writing* reminds us with a more commonly used term nowadays particularly by translation agencies but which has taken its position in Translation Studies as well: transcreation (cf. Pedersen 2014; Katan 2016). In discussing the *skopos* of multilingual communication, Guidère (2008: 17) sets two main rules: the coherence rule and the loyalty rule, which indicates the need to keep a sufficient relationship between the target text and the source message in order to not consider the translation as too literal.

Such a perspective seems limiting of the *skopos* which is based on the text function. For example, in the case of the localized text, the user, and even the client, would not necessarily be interested in the loyalty *per se*. It is the purpose of the product and its usage that determine the strategy. The translation strategy can require liberty in translating in a way to make the product suitable for the target culture and the message effect similar on the audience in question. This need for liberty becomes necessary when the target language and culture do not have the words which express certain concepts, or when these concepts are absent in the life of the other nation (Ranzato 2016: 54). Having a distance from

Madiha Kassawat

the source text is also important to avoid providing a target text which sounds like a translation and has a heavy style and is difficult to read (Boivineau 1972).

6 A trilingual localisation pilot study

In order to shed light on the variations which result from the different translation strategies, two variables will be taken into account: the language combination and the culture based on the country. The selected corpus of this pilot study is multilingual (Olohan 2004). It was selected from three international corporate web pages that are localized in several languages. Three websites are included in the corpus, one per industry. The texts are informative and commercial, and describe cosmetic, technology and furniture products. A text consists of a tagline, a subtitle and/or a short description. The structure of the text is similar across the analyzed websites, although its length can vary slightly.

Given the lack of access to the “internal knowledge” (Pym 2004), including the source language, the international English version is taken as a reference only. The term *reference text* will therefore be used instead of *source text*. The target locales in this study are Arabic (Saudi Arabia, as a representative target market on several websites) and French (Canada, France, Switzerland and Morocco). Analyzing texts in Arabic and French, besides country-based variations, should help in exploring the different cultural interpretations compared to the international text and the possible adopted approaches in the translation. This method has been also used in a study which analyzes the “uniformization” level through internationalizing the linguistic content or the affirmation of the cultural differences (Bouffard & Caignon 2006).

On the one hand, the analysis focuses on the adaptation level, the difference or similarity between the translation and the international version. On the other hand, it distinguishes the cultural points of view in each localized version and how the same message was interpreted. The analysis takes random samples from the selected websites and excludes the reasoning of the strategy, whether it is stylistic, cultural, intuitive or personal. It focuses on meeting the function of the text rather than accuracy, particularly that online versions might not be updated simultaneously, which can create discrepancies in meaning.

7 Results and discussion

The analysis of the selected websites shows a variety in the used strategies and procedures from a version to another, in relation to the reference text. The differ-

ences take place at the level of the linguistic and cultural elements. The localized versions sometimes meet the expectations of the studied genre and the required results in localization, such as the use of adaptation, locale-specific expressions and stylistic choices. These cases reflect the understanding and interpretation of the translator (or the agent dealing with the text) as a reader of the text (Plas-sard 2007). They also correspond to the knowledge of the world that the target readers have (Nord 2005: 96), and attempt to fill in the gaps that exist in their knowledge (Baker 2011).

Moreover, it has been noted that when the international text does not achieve the “*zéro spécifique*” (‘the specific zero’) (Gouadec 2003), but rather introduces culture-specific structures, the other versions introduce their own. However, when the international version is neutralized by extracting language- and culture-specific elements (Schäler 2007), it often leads to similar translations that can be considered comprehensible in the target locale but miss the local voice and the dynamism which is encouraged in this kind of translations (Tatilon 1990). This lack of dynamism can reduce the possibilities of convincing the consumer due to the gap between the localized text and the common advertising text features in the consumer’s language and culture. Furthermore, the appellative aspect (Nord 2005) is not always present in the translations, although it is a necessary one for achieving the desired effect (Ladmiral 2014).

Having said that, the use of linguistic and non-linguistic sets of knowledge has not been always observed. Sometimes, the translation did not go beyond the Practices layer in Hofstede’s Onion Model. It rather stayed limited to the Symbols layer (which includes words), conveying a wording similar to the international and other versions. From the localization industry perspective, adaptation and the offer of the “look and feel” of the target country are the “additional” characteristics in localization (Jiménez-Crespo 2013: 15). However, the translations seem to be influenced by several factors, including the localization process, the degree of internationalization, or whether this phase was applied or not. This interpretation corresponds to what Jiménez-Crespo pointed out with regard to the consequences of internationalization on the language and the translation process (2013: 10).

Regarding the locale amalgamation discussed earlier, creative variations have been observed among the French versions. This does not necessarily suggest that the variations are specifically done for the target country/culture. Many expressions used are common in different French-speaking countries. Their use depends on the intuitive of the translator-localizer and his/her interpretation associated to his/her cultural package. As for the Arabic translations, the same copies

Madiha Kassawat

were used for several Arab countries regardless of how many similarities and differences they may have, which reflects the notion of “languages of consumption” (Pym 2000). That also shows how the geographic location and language association remains unspecific (Guidère 2000: 26–27) and subject to the marketing strategy. The analysed translations can be generally considered neutral and suitable without introducing shocking linguistic or cultural aspects (Tatilon 1990). Even so, there are cases where more suitable and target-oriented adaptations could have been applied, and more culture-related aspects could have been included, as illustrated in the following paragraphs.

7.1 From untranslatability to creativity

The first example [T1] is a description of a mobile phone camera. The example includes the Arabic version of the Saudi Arabia page and the French version of Morocco:

International English version:

Super Slow-mo
The camera that slows down time,
making everyday moments epic.

Arabic version – Saudi Arabia

ميزة الحركة البطيئة جدا
تباطأ اللقطات،
لتعيش اللحظات.

‘myzat alḥaraka albaṭy’a ḡiddan
tatabaṭa’ allaqaṭāt,
lita’yš allahaḏāt’

French version – Morocco:

Super Slow-mo
La caméra qui ralentit le temps,
rendre les moments quotidiens épiques.

The Arabic translation does not have similar words to the English text: “the function of the very slow motion | the footage slows down, to let the moments

live”. On the contrary, the French version (Morocco) is more similar to the international one; even the function name “Super slow-mo” is in English. Having said that, the similarity is not viewed as a bad aspect but the translation lacks cultural references which can enrich the translation. The Arabic translation of this example seems to be creative in terms of the use of rhymes (footage/moments): *laqaṭāt* and *laḥazāt*. This stylistic choice is often used in Arabic audiovisual advertising as well (Gully 1996).

In the above example, creativity is demonstrated by the use of stylistic elements and cultural references from Arabic; while similarity to the international version was noticed in the French version. It is important to point out that the Moroccan version is available only in French on the website of this example.

The second example [T2] is a description of a mobile phone. Although culture- and language-specific references were used in both French versions, each one employed a different metaphor.

International English version:

It doesn't just stand out. It stands apart.
Completely redesigned to remove interruptions.
No notch, no distractions. Precise laser cutting and a Dynamic AMOLED screen
that's easy on the eyes make the Infinity Display our most innovative yet.

French version – France:

Il atteint de nouveaux sommets
Vous pensiez savoir à quoi ressemble un smartphone ? Écran Infinity nouvelle
génération, lecteur d'empreinte sous l'écran, technologie Dynamic AMOLED :
l'écran du Galaxy S10 est une fenêtre vers le futur.

French version – Switzerland:

Le téléphone qui sort résolument du lot
Un design entièrement repensé pour que rien ne vienne perturber votre vue.
Pas d'encoche, pas de distractions visuelles. Grâce à la découpe laser précise, au
dispositif de sécurité par reconnaissance digitale sous l'écran et à la technologie
Dynamic AMOLED qui est un régal pour les yeux, l'Infinity Display est l'écran
Galaxy le plus innovant jamais conçu.

In the French version of France, the image used for distinguishing the product from others is associated to the height and the progress achieved by the product.

Madiha Kassawat

As for the Swiss version, the distinction is represented with regard to a group of similar objects. This choice is linguistic metaphoric, stylistic and cultural. The metaphors and the expressions address the consumers directly through using references from their cultures, i.e. elements that exist in their knowledge and memory (Nord 2005: 96). Although both choices seem common in the French and Swiss cultures, the variations enrich both cultures and make the product closer to its consumers. It is noticed that the descriptions differ from the international version as well. There are additions in both versions as well, probably due to updating the content or the differences in the guidelines.

7.2 Simplicity: is it easier to translate?

The translations of the example below [T3] are semantically and syntactically close to the international text, besides using a similar style and cultural neutrality. The text is a short and simple description of a piece of furniture:

International English version:

SHOE STORAGE, COAT AND HAT RACKS

Coat, hat, shoes and go!

How do you get the hallway to be that stumble-free,
get-ready-in-the-morning-without-thinking part of your everyday? Our
different styles of shoe storage and coat and hat racks help make your outdoor
things easy to get at without using up too much of your space.

Arabic version – Saudi Arabia:

خزائن الأحذية ورفوف المعاطف والقبعات

معطف وقبعة وحذاء وانطلق!

كيف تحصل على مدخل خالي من الفوضى وتكون جاهزاً في الصباح دون أن تشغل بالك في جزء منه كل يوم؟ لدينا أشكال
مختلفة من خزائن الأحذية ورفوف المعاطف والقبعات تُساعد في ترتيب أشيائك الخارجية لسرعة الوصول إليها دون شغل
مساحة كبيرة من المكان.

‘haza’en al’aḥḍya wa rufūf alma‘ātef w alqubba‘āt

mi‘taf wa qubb’a wa ḥidā’ w ānṭaleq!

kayfa taḥṣal ‘ala madḥal ḥāly min alfawḍa wa takwn ḡāhizan fy ṣṣabāḥ dwna
an taṣḡal bālak fy

ḡuz’n minhu kulla yawm? lādynā aṣkālun muḥtalifa min ḥazāin ālaḥḍya wa
rufūf alm‘āṭif w

ālqubb‘āt tusā‘du fy tartyb aṣyā’ka alḥāriḡyya lisur‘at ālwuṣūl ilayhā dūna ṣaḡli
masāḥa kabyra
min almakān.’

French version – Canada:

Étagères pour manteaux, chaussures et chapeaux
Chaussures, manteau, chapeau, c'est parti!

Une entrée ordonnée et dégagée relève du fantasme chez vous? Vous rêvez de partir le matin sans perdre du temps à chercher? Nos range-chaussures et portemanteaux de styles variés gardent vos vêtements d'extérieur bien rangés sans occuper beaucoup d'espace.

The international text does not have visible cultural references and should be suitable for all the cultures, assuming that shoes, coats and hats are widely used. Moreover, the second phrase shows the function of the product as a suitable object for a hectic life style. This was translated similarly in both versions and was supposed to be sufficient. However, the elements used in the example do not necessarily suit all the cultures. Although the majority of the Arab countries have similar dressing habits, hats and coats for example are not a traditional custom neither for men nor for women in Saudi Arabia and other Gulf countries.

Therefore, adopting a neutral and simple approach might not be sound for these societies as the cultural references are not associated with their culture and habits. These elements could have been adapted for the target culture using the Saudi coat *bichte*, the *shemagh*, the *agale*, etc. It is important to shed light on the fact that the website provides the same version for all the Arab countries. Furthermore, *get-ready-in-the-morning-without-thinking part of your everyday* was translated in an incomprehensible way in Arabic, compared to the French version of Canada. In Arabic, the translation says: “being ready in the morning without thinking about a part of it” where *it* can relate to *the entrance* or *the morning*, which is not a clear structure.

On the contrary, the French version presents a creative structure that is shorter, more persuasive and readable through introducing a question *Une entrée ordonnée et dégagée relève du fantasme chez vous?* which means “Is an organized and tidy entrance a fantasy in your home?”. The Arabic version can be explained by the influence of its source, whether it was the English text or another version, although there are other factors which are not the subject of this article. Finally, *coat and hat racks* were translated as “coat and hat shelves” in Arabic, using words that do not make much sense as coats and hats are normally not stored on shelves.

Madiha Kassawat

7.3 The problematic untranslatable

The last example [T4] is a description of a cosmetic product. The product name is associated to the American culture, which complicates the localization further. This case requires a detailed “brief” on the treatment of this kind of issues. Nonetheless, this article focuses on the final localized product at the adaptation level.

The international English version:

GALifornia powder blush
sunny golden pink blush
GALifornia dreamin!

Benefit’s NEW GALifornia golden pink blush is part sun, pure fun! It blends bright pink with shimmering gold, for a sunkissed glow that complements all skintones. The soft, blendable formula captures the warmth of California sunshine, while the signature scent features notes of pink grapefruit & vanilla.

The Arabic version – Saudi Arabia:

GALifornia مستحضر

أحمر خدود ذهبي مسم

من منّا لا تحبّ إطلالة الفتاة الكاليفورنية!

يعطي أحمر الخدود الوردي الذهبي GALifornia الجديد من بنفت روح المرح والإشراقة المشمسة! إنه يجمع بين اللون الوردي المشرق والذهبي المتألّج، ليحتضن توهج شمس كاليفورنيا الدافئة في علبة. تتميز رائحة مستحضر GALifornia بنفحات فاكهة الجريب فروت الوردية والفانيليا. يأتي هذا المستحضر مع فرشاة خاصة مستديرة الشكل لتطبيق ناعم ومتناسق.

‘mustahḍar GALifornia

aḥmar ḥudūd ḍahaby musmr

man minnā lā tuḥib itlālat̄ alfatāt̄ alkalyfurnyya!

yu‘ty aḥmar ālḥudūd alwardy alḍahby GALifornia alḡadyd min benefit rūḥ

almarah̄ wal isrāqa

almušmisa! innahu yaḡma‘ bayn allawn alwardy almušriq wa ḍḍahaby

almutal‘le’ lyahṭadina

tawhhuḡa šams kālyfurnyā aldāfe’a fy ‘lba. tatamayyaz rā’eḡat mustahḍar

Galifornia binafaḡāt

fākihaṭ alḡryb frūt alwardyya w alfanylā. ya’ty haḡā almustahḍar m‘ furšā

ḡaṣa mustadyraṭ

alškl litaṭbyqin nā‘em wa mutanāsiq.’

The French version – France:

GALifornia | blush poudre soleil rose doré
 Le soleil californien dans un bel écrin.
 Les GALifornia girls, tout le monde les adore !
 Le NOUVEAU blush rose doré GALifornia de Benefit, c'est une dose d'éclat
 ensoleillé, adoptez-le ! Il mélange le rose vif et l'or chatoyant, capturant la
 lumière du soleil californien dans un poudrier. Le parfum envoûtant de
 GALifornia contient des notes de pamplemousse rose et de vanille. Inclut un
 pinceau blush à bout arrondi sur mesure pour une application diffuse et tout en
 douceur.

In this example, the product name is associated to California and has a play on words, where the first letters of *girl* and *California* are merged. It is obvious that the name choice has marketing purposes as it is used in English in all the analyzed versions, but the description should have provided more clarification. The French translation had the liberty to transform the part *GALifornia dreamin'!* It adds specific references to *California*: the sun. The Arabic translation though seems to be a re-writing as it provides a different expressive orientation to the message (Guidère 2000: 129). Nonetheless, it remains as ambiguous as the reference text: “Who among us doesn’t like the style of the Californian girl!” What the Californian girl refers to in terms of beauty is not clear for Arab women. Although foreign names are used for marketing purposes, California does not represent a particular reference in the Arab culture. The name would need more explanation or another name should have been used to suit the target locale.

In addition, the product type *powder blush* was replaced by the word *product* simply in the Arabic tagline and description, but was included in the subtitle. The pink color was adapted to brown to suit the image associated with the sun in desert-like environments, which is also the Arab women’s skin color in general. As for the description, the part *a sunkissed glow* was adapted in both versions for different reasons. In Arabic, *kissing* was replaced by *hugging* to avoid sexual connotations. In the French version, the expression was adapted into “catching the sunlight”, perhaps due to the lack of a similar expression in the French language.

8 Conclusion

It can be noticed from the examples discussed above that several translation strategies were used, including different interpretations generated from the same strategy. Moreover, the treatment of the texts does not always take the cultural peculiarity of the audience into account. A website version is usually influenced

Madiha Kassawat

by the general expectations of the consumers. The examples illustrated translation and creative variations which represent the fruit of cultural diversity, whether among the translators-localizers or the recipients who have their different cultural packages even when they share the same language. However, this creativity was not always present.

The localization process and that of internationalization should have an effect on the adaptation. As for the effect of internationalization, there was a noticeable simplicity and neutrality at the language and culture levels in several localized versions, where the English and localized versions had semantic and syntactic similarities. In addition, the international text is not always culturally filtered. Having said that, it might not be as helpful as required even for an international diverse audience. This is clear particularly with the product names and metaphors. Such an approach is sometimes necessary for marketing, but it introduces obstacles during the translation and adaptation of the text.

Moreover, the translator-localizer has the liberty to transform the text, adapt it and make it comprehensible for the consumers. Leaving a leeway of change to the translator should put internationalization in question with regard to its goal of helping in the product localization, i.e. the internationalized text seems to have two contradictory functions: reducing the adaptation time and encouraging cultural adaptation. This contradiction can also be associated to the international-to-local approach of the process (Jiménez-Crespo 2010). While this study does not look into the internationalized function per se, it is important to explore its effect on the quality of the translated text and to pay more attention to this area in research.

With regard to the employed methodology in this pilot study, it provided a general idea of the existing adaptation practices. However, a more detailed analysis is needed, particularly for each locale and industry. A detailed study can reveal the strengths and pain points in a localized version more accurately. The use of the notion of the “reference text” was necessary as the researcher cannot know the source language used to localize a website. Knowing the source language or version would have helped in providing a deeper interpretation of the spotted practices. Although this obstacle limited the analysis to a certain extent, it helped the researcher avoid the comparison with and the influence of a source text. Moreover, the study focused on the adaptation aspect, which is target-related in the first place.

To conclude, this paper attempted to explore how localized versions are adapted through comparing the variations at the country and culture levels, and highlighting the cultural richness these variations can bring if adaptation is used in localizing the textual content. The method used has provided both a multi-lingual

and multi-cultural approaches which go beyond words and encompass the product personalization in order to represent the local culture. This pilot study can contribute in more culture- and language-oriented research in website localization, which often focuses on technical aspects.

Analysed texts

[T1, T2] Available on <http://www.samsung.com>, [accessed on 22 April 2019]

[T3] Available on <https://www.ikea.com>, [accessed on 18 April 2019]

[T4] Available on <https://www.benefitcosmetics.com>, [accessed on 19 April 2019]

References

- Anastasiou, Dimitra & Reinhard Schäler. 2010. Translating vital information: Localisation, internationalisation, and globalisation. *Synthèses Journal* 3. 11–25.
- Austermühl, Frank. 2006. Training translators to localize. In Anthony Pym, Alexander Perekrestenko & Bram Starink (eds.), *Translation technology and its teaching (with much mention of localization)*, 69–81. Tarragona: Intercultural Studies Group.
- Baker, Mona. 2011. *In other words*. London: Routledge.
- Boivineau, Roger. 1972. L'A. B. C. de l'adaptation publicitaire. *Meta* 17(1). 5–28.
- Bouffard, Paula & Philippe Caignon. 2006. Vers une géolinguistique de l'espace virtuel francophone. *Meta* 51(4). 806–823.
- Canım Alkan, Sinem. 2017. Position of the translator as an agent in website localization: The case of Turkey. *Journal of Language and Linguistic Studies* 13(2). 510–525.
- Cronin, Michael. 2006. *Translation and identity*. New York: Taylor & Francis.
- Cruz-García, Laura. 2018. Advertising across cultures, where translation is nothing... or everything. *The Journal of Specialised Translation* 30. 66–83.
- Esselink, Bert. 1998. *A practical guide to software localization*. Amsterdam: John Benjamins.
- Esselink, Bert. 2000. *A practical guide to localization*. Amsterdam: John Benjamins.
- Esselink, Bert. 2003. The evolution of localization. *The Guide from Multilingual Computing & Technology: Localization* 14(5). 4–7.
- GALA. 2019. What is localization? <https://www.gala-global.org/industry/intro-language-industry/what-localization> (6 May, 2020).

Madiha Kassawat

- Gouadec, Daniel. 2003. Le bagage spécifique du localiseur/localisateur: Le vrai « nouveau profil » requis. *Meta* 48(4). 526–545.
- Gudykunst, William B. 2004. *Bridging differences: Effective intergroup communication 4th edition*. Thousand Oaks: Sage Publications.
- Guidère, Mathieu. 2000. *Publicité et traduction*. Paris: L'Harmattan.
- Guidère, Mathieu. 2008. *La communication multilingue*. Bruxelles: De Boeck.
- Gully, Adrian. 1996. The discourse of Arabic advertising: Preliminary investigations. *Journal of Arabic and Islamic Studies* 1. 1–49.
- Hofstede, Geert, Jert Jan Hofstede & Michael Minkov. 2010. *Cultures and organizations – software of the mind – intercultural cooperation and its importance for survival*. New York: McGraw Hill.
- Jiménez-Crespo, Miguel. 2010. Web internationalisation strategies and translation quality: Researching the case of “International” Spanish. *Localisation Focus* 9(1). 13–25.
- Jiménez-Crespo, Miguel. 2013. *Translation and web localization*. London: Routledge.
- Katan, David. 2016. Translation at the cross-roads: Time for the transcreational turn? *Perspectives* 24(3). 365–381.
- Ladmiral, Jean-René. 2014. *Sourcier ou cibliste*. Paris: Les Belles Lettres.
- Le Dizez, Jean-Yves. 2004. Traductologie et traduction pragmatique. *Tribune, Translittérature* 26. 59–64.
- Maitland, Sarah. 2017. *What is cultural translation?* London: Bloomsbury.
- Munday, Jeremy. 2008. *Introducing translation studies theories and applications*. London: Routledge.
- Nord, Christiane. 2005. *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis*. Amsterdam: Rodopi.
- Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London: Routledge.
- Pedersen, Daniel. 2014. Exploring the concept of transcreation – transcreation as “more than translation”? *Cultus: The Journal of Intercultural Mediation and Communication* 7. 57–71.
- Plassard, Freddie. 2007. *Lire pour traduire*. Paris: Presses Sorbonne Nouvelle.
- Pym, Anthony. 2000. *Localization and the changing role of linguistics*. <https://usuaris.tinet.cat/apym/on-line/translation/TunisPaper.pdf> (23 November, 2020).
- Pym, Anthony. 2004. *The moving text: Localization, translation, and distribution*. Amsterdam: John Benjamins.

- Pym, Anthony. 2005. Localization: On its nature, virtues and dangers. *SYNAPS* 17. 17–25.
- Pym, Anthony. 2006. Globalization and the politics of translation studies. *Meta* 51(4). 744–757.
- Pym, Anthony. 2011. Website localization. In Kirsten Malmkjær & Kevin Windle (eds.), *The Oxford handbook of translation studies*, 410–423. Oxford: Oxford University Press.
- Quah, Chiew Kin. 2006. *Translation and technology*. London: Palgrave Macmillan.
- Ranzato, Irene. 2016. *Translating culture-specific references on television: The case of dubbing*. New York: Taylor & Francis.
- Rémon, Joséphine. 2005. Interculturel et internet: Le site web, objet culturel ? In Luc Collès, Christine Develotte, Geneviève Geron & Françoise Tauzer-Sabatelli (eds.), *Didactique du FLE et de l'interculturel: Littérature, biographie langagière et médias*, 267–270. Paris: L'Harmattan.
- Schäler, Reinhard. 2007. Reverse localisation. *Localisation Focus* 6(1). 39–48.
- Stenger, Thomas & Stéphane Bourliataux-Lajoinie. 2014. *E-marketing & e-commerce, concepts – outils – pratiques*. Paris: Dunod.
- Tatilon, Claude. 1990. Le texte publicitaire: Traduction ou adaptation ? *Meta* 35(1). 243–246.
- Yunker, John. 2003. *Beyond borders: Web globalization strategies*. Indianapolis: New Riders.

Chapter 5

Movement or debate? How #MeToo is framed differently in English, Spanish and German Twitter discourse

Mario Bisiada

Universitat Pompeu Fabra

This article examines 1,353 tweets on #MeToo in English, Spanish and German from July and August 2019, revealing how #MeToo is most commonly referred to as a “movement” in English and Spanish but as a “debate” in German, a difference that echoes German-language press habits. Based on an analysis of semantic prosody, the study demonstrates that words indicating longevity such as “era” and “times” collocate with #MeToo in English and Spanish, but not in German. This points to a framing of #MeToo as influential and long-term in English and Spanish and as exaggerated and short-term in German. Reflecting this difference, #MeToo is talked about in more negative terms in German tweets compared to English and Spanish, as shown by a qualitative analysis of evaluative author stance. The study adds to existing knowledge of the power of hashtags for feminist social media activism by highlighting the importance of (cross-)linguistic corpus-assisted discourse studies of hashtags on social media, which helps understand the ways in which anti-feminist discourse taps into the channelling of emotions through hashtags to undermine cross-national women’s movements.

1 Introduction

Since it began trending in October 2017, the #MeToo hashtag has circulated in 85 countries (Gill & Orgad 2018: 1317) and has had a large-scale global impact on societies (Zarkov & Davis 2018), both in terms of positive effects (Fileborn & Loney-Howes 2019) and backlashes (Boyle & Rathnayake 2019). Two thirds



Mario Bisiada

of Canadians say the #MeToo campaign has had an impact on them personally (Angus Reid Institute 2018). Consequently, there is a sizeable amount of literature on the hashtag's power for social movements and on its political effects, and some cross-cultural work has comparatively analysed attitudes towards #MeToo in the US and Norway (Kunst et al. 2018) and investigated differences in media framing of women as "silence breakers" in the US, Japan, India and Australia (Starkey et al. 2019). Linguistic studies of hashtags have mainly concentrated on functional aspects (see Zappavigna 2018: 7), but cross-linguistic studies of hashtag use are still rare. Given that hashtags are usually transnationally used, studies of their effect on societies from a cross-linguistic and cross-cultural perspective could aid in the understanding of international hashtag activism, especially as regards the increasing global cooperation of feminist movements (Huelga Feminista 2018; Oppenheim 2018; Garibotti & Hopp 2019: 186).

This paper analyses the framing of and evaluative stance towards the #MeToo hashtag from a cross-linguistic point of view to see how discourses around what seems to be the same concept diverge across languages and which effect this may have on language users' perception of the issue. Based on a corpus of English, Spanish and German tweets from July and August 2019, I address the following three research questions:

1. How is the #MeToo hashtag represented in English, Spanish and German discourse through words that frequently accompany it?
2. Which types of evaluative stance towards #MeToo can be identified and how do they differ cross-linguistically?
3. Are there cross-linguistically similar discourse patterns around the #MeToo hashtag and what effect does this have on it as a safe space for hashtag feminism?

While hashtags are generally agreed to empower women and feminist discourse on social media, some research has also questioned whether hashtag discussions really provide safe spaces. The core purpose of this paper is to demonstrate how linguistic analysis of hashtag discourse can identify reasons why hashtag networks are not always the safe spaces that hashtag feminism makes them out to be because they also allow or even attract misogynists to link into those networks. In the following section, I provide an overview of those arguments and discuss the importance of the study of hashtags as linguistic items.

2 Communicative effects of feminist hashtags

A major aim of hashtag activism has been to lay claim on public spaces (Lünenborg & Maier 2013: 63; Bowles Eagle 2015). Considering them a “collectivising feminist response to rape culture”, hashtags such as #safetytipsforladies reveal “the feminist delight in exposing misogynist, victim blaming ideas through humor” (Rentschler 2015: 354). Hashtag activism on social media plays an important role to establish counter-narratives against the dominating forces in public discourse. The hashtag #MuslimWomensDay, for instance, gave Muslim women a voice to tell their stories and thus allowed them to challenge dominant media frames representing them as silent victims (Pennington 2018: 200). Other hashtags such as #YesAllWomen and #YesAllWhiteWomen created “feminist counter-publics” to “rewrite dominant public narratives about violence against women” (Jackson & Banaszczyk 2016: 392).

Hashtags are powerful communicative and political tools because they allow victims of abuse to express and share their experiences and because they seem to affect people on an emotional level inaccessible to newspaper reports (Keller et al. 2018; Mendes et al. 2018). Both of those capabilities make them more effective than traditional media and both can be traced to the combination of linguistic metafunctions that hashtags convey (see Zappavigna 2015). The ability to evoke the appropriate emotions at particular moments in time and to control and channel these emotions is important in gaining followers for political movements (Ahmed 2004). This power can of course be used for both liberal and repressive campaigns, as the hijacking of #MeToo by far-right supporters such as the #120dB campaign shows (Farris 2017; Sorce 2018; Wielens 2019).

Twitter enables girls and women to connect to each other and to share experiences of gender violence in hitherto unknown ways (Keller et al. 2018), bypassing the traditional news cycle and “the mainstream media’s problematic framing of sexual violence and black women” (Williams 2015: 342). Participants in a study on the #BeenRapedNeverReported hashtag (Keller et al. 2018), for instance, were reluctant to speak to the researchers and regarded Twitter as a safer way to share experiences of gender violence. The tweets “all carried the common theme that it remained professionally, emotionally, and even physically costly to report sexual violence to authorities, disrupting the prevalent myth that unreported assaults are illegitimate” (Keller et al. 2018: 27). The hashtag thus produced the benefit of giving its users a sense of community, affective solidarity and support (Keller et al. 2018: 28–29).

In this way, hashtags have the political effect of creating new followers for the international feminist movement (Schachtner & Winter 2005; Mendes et al.

Mario Bisiada

2018: 238). In her study of #WhyIStayed, [Clark \(2016: 789\)](#) argues that the political effect hashtags have is caused by its ability to turn individual stories into a narrative, drawing on the feminist movement's historical emphasis on discourse, language and storytelling, while also enabling women to engage in "dark parody of mainstream media discourse" ([Clark 2016: 796](#)).

[Clark-Parsons \(2019\)](#) investigates activists' strategies to achieve visibility in a corpus of around 3,000 tweets using the #MeToo hashtag. She finds four major categories ([Clark-Parsons 2019: 7](#)):

1. participants' understandings of the political potential of #MeToo
2. their takes on the tactic's political limitations along with their attempts to redress these shortcomings
3. their concerns regarding whose voices are included in the campaign
4. their efforts to support the campaign and protect its survivor-participants

She concludes that participants in #MeToo "reclaimed their agency and pushed back against discourses that normalise harassment and assault" ([Clark-Parsons 2019: 16](#)). The narratives collected under #MeToo have had the effect of a social transformation by making the personal political and scaling up from individual to collective visibility ([Clark-Parsons 2019: 16](#)). Her study shows that sizeable participation is required to avoid that transgressions reported under a given hashtag become personalised or framed as individual errors, something that traditional media still tends toward ([Kornemann 2018: 383](#)). If it reaches sufficient participation, "the networking functions of the hashtag bridge the personal and the political, recasting, in the case of #MeToo, sexual violence as a systemic, rather than private, issue and calling for structural changes in response" ([Clark-Parsons 2019: 16](#)).

The advantages of hashtags over traditional media may well cause unease in the latter. Research has argued that media focus on celebrity involvement "distract[s] us from systemic, structural sexism across all industries" ([Banet-Weiser 2018: 17](#)). The hashtag #Aufschrei ('outcry'), highly influential in Germany (see [Maireder & Schlögl 2014](#)), has set the agenda in the public debate on sexism in spite of a largely antipathetic media focus ([Kornemann 2018: 381](#)). Traditional media have preferred to treat the #Aufschrei movement in a symbolic way rather than engage with its content, thus reducing the general problem of gender discrimination to isolated occurrences ([Kornemann 2018: 386](#)). Such an approach

“can end up working against the calls for social change promised at [the movement’s] beginning, producing more and more visibility – and increasingly narrowing the discourses of that visibility in the process” (Banet-Weiser 2018: 17).

That individualising tendencies in reporting de-politicise feminist movements for their readership is also argued by De Benedictis et al. (2019: 733) in their study of reporting on #MeToo in 613 UK newspaper articles from October 2017 to March 2018. Although they find that the press in the UK has played an important role in promoting #MeToo overall, they identify a focus on the stories of “the ‘ideal victims’, namely, celebrity female subjects (who are predominately White and wealthy)”, concluding that,

by failing to inform the public about or to debate potential solutions, the press can be understood to have helped defuse any potential that #MeToo might contain as a mobilising social force. Rather, the press seems to have framed #MeToo largely in terms of neoliberal and popular feminisms, which disavow structural analysis and critique and largely place responsibility on individual women. (De Benedictis et al. 2019: 734)

Once a hashtag has firmly entered the public debate through mainstream media, it also becomes open to attacks by opposing forces. Based on a corpus of 700 tweets from between January 2013 and March 2014, Drüeke & Zobl (2016) analyse author stances towards #Aufschrei, establishing the four categories of “supportive”, “neutral”, “dismissive” and “impossible to specify” (Drüeke & Zobl 2016: 43). They find that, from the second week onward, the supportive tweets increased and thus emphasised the importance of the hashtag (Drüeke & Zobl 2016: 44–45). The more widespread the hashtag became, however, the more it attracted dismissive reactions, and the personal experience postings were joined by blatant anti-feminist statements (Drüeke & Zobl 2016: 44–45). They conclude that Twitter cannot be considered a safe space, as “anti-feminist and sexist comments are equally visible and might signify new experiences of violence for women” (Drüeke & Zobl 2016: 51).

The channeling emotional effect of hashtags may, then, also work the adverse way once abused by trolling or hate speech. The categorising function of the hashtag does not offer a filter to only encounter posts true to the original intention of the hashtag, which may have undesired consequences for social media users who find themselves in an emotionally fragile state induced by following the hashtag narratives. The power of hashtags is enhanced by “the transnational nature and technological affordances of social media, whereby interest groups with similar agendas can more easily find one another” (Ging 2017: 652–653).

Mario Bisiada

The study of hashtags unavoidably touches on the study of anti-feminist discourse on social media, which is transnationally networked and may share a set of cross-linguistic frames to attack feminist movements. That discourse is just the publicly visible face of a clandestine, globally networked subculture of men who describe themselves as “incels” (‘involuntary celibates’, see [Valens 2018](#)) and engage in celebrating violence against women and fostering a misogynist discourse on forums such as 4chan and 8chan (see [Jaki et al. 2019](#)). These networks are a grave international threat: the terrorist attack in Christchurch (New Zealand) inspired men to commit similar attacks in Poway (US), El Paso (US), Baerum (Norway) and most recently Halle (Germany). In all cases, the shooters were regular users of 4chan or 8chan and explained their ideology in manifestos uploaded to those networks. In those manifestos, the terrorists express anti-semitic and white supremacist fears of white genocide through a decrease in fertility rates of white people, something for which they blame what they perceive as mass-immigration, but ultimately also feminism ([Di Stefano 2019](#); [Kahlke Lorentzen & Shakir 2019](#)).

These shootings are generally attributed to racist and anti-semitic views, while misogynist motives, though demonstrably present, tend to be ignored. Mass shooters are not necessarily all far right; the perpetrator of the 2019 Dayton shooting, probably inspired by the El Paso shooter, declared himself left-wing. What united him with the other shooters was that he showed signs of misogyny ([Svokos 2019](#)). In fact, [Follman & Exstrum \(2019\)](#) show that in 22 analysed mass shootings since 2011, a third of perpetrators had a history of stalking and harassment and half of them specifically targeted women.

To understand how transnational, cross-linguistic misogynist and anti-feminist discourse works to globally attract particular groups of men, more linguistic research into its surface form on social media and hashtag discourse can be instructive. Hashtags are by necessity cross-linguistic phenomena; however, little cross-linguistic research on hashtags exists to date. Translation studies and corpus-assisted cross-linguistic discourse studies are fields that can contribute such analyses, and this chapter suggests one possible way of doing so through corpus study of semantic prosody and author stance. In the following section I outline how this study seeks to contribute such an analysis.

3 Methodology

This study is based on 1,353 tweets containing the #MeToo hashtag (505 in English, 405 in Spanish and 443 in German), gathered between July and August

2019. The data were collected using the Twitter Search API. This API allows researchers to search for particular queries over a specific time period, within seven days of the tweet being posted. The API therefore gives us a snapshot view of the posts containing, in this case, the #MeToo hashtag at the present moment. It is important to note that the API is programmed by Twitter to not yield all the tweets containing this hashtag, but just a selection of what it considers relevant at the moment of searching (see also Zappavigna 2018: 7; Boyle & Rathnayake 2019). The data analysed here should thus be understood as a snapshot comparison of activity around the #MeToo hashtag in the English, Spanish and German language communities¹ during July and August 2019.

One important aspect in the analysis of collocations across languages is the difference in how lexical units are usually formed. Spanish uses slightly different word formation rules to English and German; while it is “somewhat resistant to orthographic noun compounds” (Lang 1980: 81), typical in Germanic languages, the more natural and “highly productive” (Lang 1980: 85) construction in Spanish would be what Lang (1980: 85) calls “prepositional link syntagms”, where words are joined with the preposition *de* (in our case *x de(l) #MeToo*) to form a lexical unit (Lang 1980: 85). Not taking this difference into account may well lead to observing a higher likelihood of collocations in English and German compared to Spanish. Therefore, I have decided to count such prepositional link syntagms, which also occur in the Germanic languages, albeit at a lesser rate, among the collocations. Here is an example of one such prepositional link syntagm:

En la era del #metoo este rollo de mujer objeto del desfile de Victoria's Secret se ha quedado obsoleto, y a la gente joven les parece casposo. No veo la mala noticia por ningún lado.

[‘In the #MeToo era, this whole thing of objectifying women that is the Victoria's Secret fashion show has become obsolete and to young people it seems inappropriate. I don't see the bad news at all.’]

While the Twitter Search API gives access to the username of the author of the tweet, I have anonymised all usernames mentioned in the examples to “@user”, except institutions or public figures, defined as those with verified accounts (Twitter 2019). Images in the tweet and retweets of other tweets are converted to abbreviated links by the Search API, in addition to links the user may have posted.

¹In this paper I refer to language communities instead of countries for two reasons, which I think apply generally to cross-linguistic studies of social media data. First, the internet is a transnational space, so discussions happen within language communities, which often transcend countries. Secondly, the country stated in the profile (if at all) does not allow us to make inferences on the native language of the user.

Mario Bisiada

While the analysis of author stance has taken into account the entire tweet including images and articles linked to, I have generally removed all links from the tweets printed as examples in this paper in order to remove visual clutter. The tweets are not otherwise edited for orthography and all translations are my own. I have made the data underlying this research open and encourage readers wishing to consult the full tweets to do so.

As stated above, the objectives of this study are to analyse the semantic prosody of #MeToo in Twitter discourse and to investigate the author stance in the tweets where the #MeToo hashtag is integrated into the sentences. Semantic prosody is here understood as pragmatic colouring applied to words by its collocates (Louw 1993: 158–159; Stewart 2010; Vessey 2013: 13–14), a concept that has proved useful also in cross-linguistic research (Lewandowska-Tomaszczyk 1996). I am here specifically interested in what Stewart (2010: 61) terms the study of semantic prosody as a feature of word + co-text, by which “some lexical items are associated with prosodies whose meaning is in marked contrast with the basic meaning of the node/core item in question” (Stewart 2010: 61). We can extend this notion to the case where a new hashtag such as #MeToo is integrated into a compound and thus given a particular colouring through the semantic prosody of the accompanying term. Crucially, I adopt the view that “through his or her language, and more specifically, through the use of collocations and the effect of semantic prosodies, an acculturated speaker often (re)produces the values and judgments of his or her discourse community” (Vessey 2013: 13). In the present context, I argue that the collocates appearing alongside #MeToo allow us to discern the stance towards the issue prevalent in each respective language community.

For the analysis, I have first separated those tweets where the hashtag is part of the sentence like any other word from those where it is put at the end, outside the sentence. In the former case, it is said to serve as a lexical item in “integrated” position, while in the latter, the hashtag is in “culminative” position (Zappavigna 2018: 31–32; see also Scott 2015: 14). Hashtags in the integrated position “take on functional roles in the clause. In culminative position, while the hashtag can construe any number of functional roles on its own [...], it is typically not integrated into the clause. Instead it is appended at the end of the post.” (Zappavigna 2018: 32)

To analyse semantic prosody of the #MeToo hashtag, then, the first step is to separate occurrences according to this criterion, as well as to mark duplicate tweets, which mainly include newspaper headlines shared by many people, and unrelated uses of the hashtag, mainly to say “me too” without referring to the movement. Each hashtag integrated into the sentence is then analysed to see if it appears as a collocation and, if so, which collocate accompanies it.

To address the second research question, the analysis of author stance, I have conducted an analysis of tone using the method proposed by [Orgad & de Benedictis \(2015\)](#) and [Drüeke & Zobl \(2016\)](#). Thus, tweets were considered “positive” if they showed “commendation/appraisal/valuing/appreciation or recognition” of #MeToo, and “negative” “if they demonstrated or included substantial criticism/derision/cynicism or dismissal” ([Orgad & de Benedictis 2015](#): 424–425) of #MeToo. In addition, I have established the category “critical” for tweets that criticise #MeToo in a constructive way, expressing discontent with some aspect but being in favour overall. Tweets that cannot clearly be attributed to any of these categories have been labelled “unclear”.

For the final objective, the identification of cross-linguistic framings, I have compared the findings from the first research question and compared the semantic prosodies identified for each language in order to identify patterns.

4 Semantic prosody of #MeToo

Table 5.1 shows the absolute and relative frequencies of #MeToo hashtags that are in the integrated and in the culminative position, as well as unrelated and duplicate occurrences. In English and German, 41% of the tweets have the hashtag in integrated position while about half of all occurrences are culminative. In Spanish, 60% of the hashtag occurrences are integrated while only about a third are culminative. This may indicate a preference on the part of Spanish users to integrate the hashtag into the sentence, though a larger corpus sample is necessary to support this observation. Based on this data, I have proceeded with analysing the semantic prosody of the #MeToo hashtag.

Table 5.1: Absolute (n) and relative (p) frequencies of occurrences of the #MeToo hashtag in the corpora of tweets

Category	English		Spanish		German	
	n	p	n	p	n	p
integrated	210	41%	237	58%	184	41%
external	242	47%	126	31%	202	46%
unrelated	18	4%	10	3%	13	3%
duplicate	41	8%	32	8%	46	10%
Total	511	100%	405	100%	445	100%

Mario Bisiada

To analyse the semantic prosody of the #MeToo hashtag, I have conducted a frequency analysis of its collocations where it does not occur as a single word. There are five occurrences of a verb form *metoo'd*, a verbalisation that is interesting from a formal linguistic point of view and a commonly observed phenomenon in English, but need not concern us any further at this point. I have counted these occurrences among the single-word occurrences. The absolute and relative frequencies of #MeToo as a single word are as follows:

English 83 single-word occurrences (40% of the integrated occurrences)

Spanish 143 single-word occurrences (60% of the integrated occurrences)

German 114 single-word occurrences (62% of the integrated occurrences)

This means that in English, a slight majority of the hashtags in integrated position are parts of collocations, while in Spanish and German, single-word occurrences are in the majority. While this may indicate a slight preference among English users to form collocations with the #MeToo hashtag when compared to Spanish and German users, this quantitative observation again needs to be backed up by a larger data sample. Table 5.2 shows all the collocates of the #MeToo hashtag found in the corpora, with the number of occurrences in parentheses. I now discuss the three corpora in turn.

4.1 The English corpus

In the English corpus, the most common collocate with 38 occurrences is *movement*, exemplified below.

The theme of gender oppression runs throughout the collection, as befits current debates in the West and beyond over sexual violence and predatory behaviour in the wake of the #metoo movement.

This collocate is supported by the fact that #MeToo is now labelled a movement on Wikipedia and that there is an official movement page offering

a comprehensive database consisting of local and national organisations dedicated to providing services and safe spaces for survivors of sexual violence, healing stories, as well as articles and a glossary of terms to help give voice to your experiences (*me too*. 2018).

Table 5.2: Collocates of #MeToo in the English, Spanish and German corpora, sorted by frequency of occurrence (single-occurrence items omitted)

English (n=127)	Spanish (n=94)	German (n=70)
1. movement (38)	movimiento (44)	Debatte (16)
2. moment (15)	tiempos (9)	Hysterie (8)
3. campaign (8)	era (6)	Bewegung (6)
4. era (7)	campaña (4)	
5. merch, case (3)		Aktivistin(nen) ('activist(s)' (f.)) (3)
6. accusers, complaints, rhetoric, story, survivor, victims (2)	momento ('moment') (2)	Hexenjagd ('witch hunt'), Kampagne ('campaign'), Inquisition, Zeiten ('times'), Befindlichen ('affected'), Folgen ('consequences') (2)

While it is a neutral term in that it does not in itself show positive or negative stance, its use does mean a recognition of the hashtag as influential and with potentially large-scale effects on society, which gives the compound *#MeToo movement* an approving stance.

The collocation *moment* occurs 15 times, mainly in tweets citing some entity's #MeToo moment, as in the first example below, but it can also denote just a momentary instance of something bigger, as in the second example:

.@ruthmaclean reports on how 'Nigeria's #MeToo moment' turned against rape accuser #globaldev

A definite #MeToo Moment. I'm retweeting because she shouldn't be the one quitting. These people should be reprimanded and forced to pay restitution. The Rookie production should be behind her.

As a collocation, *#MeToo moment* is used as a label under which a series of articles as well as “updates and analysis on the #MeToo movement” are collected in the New York Times (Bennett 2019). In fact, whether #MeToo is a movement or a moment has been discussed in a series of press articles (Akhtar 2017) which argue that “unless actions replace hashtags and value signaling, we'll see old power

Mario Bisiada

structures and patterns of behaviour remain as entrenched and unequal as ever – along with a healthy new dose of mistrust and resentment with which women will contend” (Senecal 2018). Tarana Burke, the founder of the movement, has said in a TED talk in 2018 that the fact she is giving this talk shows that #MeToo “is bigger than a moment. It’s the confirmation that we are in a movement. And the most powerful movements have always been built around what’s possible, not just claiming what is right now” (Burke 2018).

In a Canadian survey, a majority of participants consider #MeToo a movement, with 53% of participants saying that “the #metoo movement will lead to some change, but it will take years, if not decades for real change” and 31% saying that “these discussions have sparked a major and permanent shift”, while only 14% believe that “people might be paying attention now, but it will blow over, and nothing will really change” (Angus Reid Institute 2018). It is this last view of #MeToo that the *moment* collocation arguably expresses, either out of sheer rejection of the movement or because its user accepts the systematicity of the transgressions, but does not have any hope that things will change.

With eight occurrences, we find the collocate *campaign*, followed by *era* with seven occurrences. The collocate *campaign* in itself is a neutral term, similar to *movement* in that it is a collective action, though not as widely influential as a movement. It was mainly used by the press in articles immediately after the initial wave of #MeToo postings and, as the examples show, might now be used by people who are critical of or reject the movement.

@therealaftonw if #AftonWilliamson was lighter, blond haired, light eyed, pointy nosed then could she then be a part of the #MeToo campaign? Investigate the matter please and get back to me

The story of fake rape allegations against #TonyMochama as found by our courts militate against the #Metoo campaign turning the entire mass hysteria on its head. Kindly guys its never worth it.... why cook a false narrative and drive it for sympathy and in the end tarnish ppl

The word *era* also denotes a time span and is thus comparable to *moment*, though it has the added semantic shade of referring to a before and after of a certain watershed moment. As the examples show, it goes along with a certain desperation at a given event and seems to be used mainly to achieve a semantic prosody of indignation based on #MeToo, without necessarily taking a stance towards the hashtag itself.

@user @brianefallon So drunk staff being inappropriate makes it OK for sobet teens to mime choking a young women in the #MeToo era with rampant campus rapes?? I don't like Whataboutism #STUPIDITY

@user @user @brianbeutler In the very unlikely event he got the nomination, Bernie's rape fantasy or teacher bashing essays will be widely disseminated by the GOP. I cannot, for the life of me, see how this isn't disqualifying in the #metoo era.

Many of the collocations used three times or less are rather self-explanatory rejections of the #MeToo movement and will not be commented on.

4.2 The Spanish corpus

Looking at the Spanish data, we find the collocate *movimiento* ('movement') with 44 occurrences as by far the most frequent collocation, just as it is for the English data. The Spanish-language Wikipedia entry for #MeToo also describes it as a movement, and various Spanish language glossaries list the hashtag as a movement. The Spanish newspaper *El País* has featured a special entitled *Revolución MeToo* (*El País* 2018), where #MeToo is either referred to without any collocates or as a movement. Here is an example illustrating the collocation:

@user Pues que ahora, gracias al movimiento #metoo se atreven a denunciar y a declarar en contra de esos antaño intocables y todopoderosos hombres que las han tratado como a cachos de carne de mercado desde su infancia.

[‘Well that now, thanks to the #metoo movement, they dare to speak out and make a statement against these formerly untouchable and all-powerful men who have treated them like pieces of meat at a market since they were kids.’]

The second most common collocate in the Spanish data is *tiempos* ('times') with nine occurrences, followed by *era* ('era') with six occurrences, taking into account that this includes prepositional link syntagms (see the Methodology section). The word *campaña* ('campaign') occurs four times. Below are some examples.

En la era #MeToo, hay que convertir los espacios públicos en lugares donde las mujeres puedan existir sin ser miradas, juzgadas o comentadas

[‘In the #MeToo era, public spaces must be converted to places where women can exist without being looked at, judged or commented on’]

A sumarnos a la campaña #METOO por menos mujeres con violencia
Más mujeres felices

Mario Bisiada

#uniendofuerzas

[‘Let’s join the #METOO campaign for fewer women with violence [sic] — more happy women — #unitingforces’]

The Spanish data thus largely mirror what has been observed in the English data.

4.3 The German corpus

The German data differ from the English and Spanish data discussed thus far in that *Bewegung*, the German equivalent for “movement”, only occurs in six tweets, most of which attack the movement, as shown below.

@Junge_Freiheit wie erwartet stellt sich in vielen Fällen heraus, dass diese #metoo Bewegung eine inquisitorische Hexenjagd auf Unschuldige war/ist, was diese widerliche Bewegung 1.000 mal schlimmer macht als das was sie anprangert.

[‘as expected, in many cases it turns out that the #metoo movement was/is an inquisitorial witch hunt for innocents, which makes this disgusting movement 1,000 times worse than that which it condemns.’]

@westfalenblatt Also DER #Leuchtturm Fall der #MeToo Bewegung steht auf so wackeligen Beweisen, dass der Fall eingestellt werden muss. Was ist denn dann mit den ganzen anderen Fällen, die nur in der #Empörungswelle mitschwammen? Wer entschädigt jetzt diese Opfer, wo blieb die Unschuldsvermutung?

[‘So THE landmark case of the #MeToo movement is based on such shaky evidence that the case has to be closed. What then happens in all the other cases that just joined the flow of indignation? Who will indemnify these victims, what happened to the presumption of innocence?’]

Interestingly, at the time of writing of this article, the German-language Wikipedia page does not define #MeToo as a movement, but just as a “hashtag”.

The most frequent collocate in the German data is *Debatte* (‘debate’), which occurs 16 times. As we saw above, this frame does not occur at all in the English and Spanish data. Below are some examples.

Schauspielerin Emilia Schüle findet die #Metoo-Debatte wichtig—und dass sich in der Gesellschaft noch viel ändern muss.

[‘The actress Emilia Schüle considers the #Metoo debate important—and that much still has to change in society.’]

Die #MeToo-Debatte um Alltagssexismus hat einer Studie zufolge die Situation für Frauen am Arbeitsplatz in den USA leicht verbessert.

[‘The #MeToo debate on daily sexism has lightly improved the situation of women in the workplace, according to a study.’]

@tazgezwitscher @user @user ohne, dass es Proteste gibt. Danke jedenfalls für den Einblick in die Blase des US-Feminismus. Vielleicht ist das das Problem bei der #metoo Debatte in Deutschland, dass man auf Twitter unter #metoo viel US-Feminismus mitbekommt, was nicht der Realität in Deutschland entspricht?

[‘without there being protests. Thanks anyway for the insight into the bubble of US feminism. Maybe that’s the problem with the #metoo debate in Germany, that you get a lot of US feminism under #metoo on Twitter, which doesn’t conform to the reality in Germany?’]

These are moderate comments on the topic, which shows that the word *debate* is the common collocate used with #MeToo in German. A query in the DWDS newspaper corpus, consisting of the most important German-language newspapers, confirms this impression: in texts from 2017 and 2018, where #MeToo as a single word occurs 384 times, *#MeToo-Debatte* occurs 178 times, while *#MeToo-Bewegung* occurs 49 times (DWDS 2019). It seems, then, that in German-language public discourse, the most common way to frame the #MeToo hashtag is not as a movement, but as a debate, and that the press is actively involved in promoting this frame through the use of the term *#MeToo-Debatte*. It seems, then, that, while international discourse on #MeToo awards it the status of a movement and all the connotations of this as described above, German discourses avoid such a labelling for some reason.

This might be explained by a tendency among German media to envisage a hierarchical cline between traditional and new media, which affects their role in society and against which they seek to strengthen their own position (Kornemann 2018: 382). Hashtag movements such as #MeToo, which enter public discourse as folksonomic creations having circumvented traditional news media gate-keepers, still leave the traditional media somewhat taken aback and struggling to react appropriately. This might be especially true of feminist movements, which meet strong hegemonic barriers through the patriarchal establishment and recurrent discrimination of women that still pervade many press agencies (Banet-Weiser 2018). A possible explanation for the establishment of the more reluctant “debate” frame may thus be that labelling #MeToo a movement would distance traditional media from further involvement, while framing it as a debate means the traditional media retains an active role in its development.

The second most common collocate in German is *Hysterie* (‘hysteria’), with 8 occurrences, a purely negative frame which is not found in the English and Spanish data either. Its occurrence may have been pushed by a lot of activity around

Mario Bisiada

an influential blog article, which may have distorted the data somewhat, but this frame goes hand in hand with the intent to reclaim the term *Opfer* ('victim') and apply it to those accused under the #MeToo hashtag, observable in a range of examples.

Anklage gegen Opfer der #MeToo Hysterie wird fallengelassen.

['Case against victims of #MeToo hysteria is dropped.']

Was von der #metoo Hysterie blieb. Jetzt läuft die #Klimahysterie. #Fridaysforfuture #Great #Grüne #Kulturbereicherung

['What remains of the #metoo hysteria. Now it's the #ClimateHysteria. #Fridaysfor-future #Great #Greens #CulturalEnrichment']

Given the frequency in English and Spanish of terms such as *era*, *tiempos* or *moment*, which frame #MeToo temporally, as something that is long-term and influential, it is notable that the only collocate in the German corpus reflecting such a frame is *Gegenwart* ('present'), and two tweets containing the phrase *in Zeiten von #MeToo* ('in times of #MeToo'), as shown below. Again, a query in the DWDS newspaper corpus confirms the absence of this frame from German discourse, as, for instance, *#MeToo-Ära* occurs just three times (DWDS 2019).

Lashana Lynch: Wie auch James Bond in die MeToo-Gegenwart gezerrt wird #James-Bond #MeToo

['How even James Bond is dragged into the MeToo present']

@ulfposh Alter weißer Mann findet alten weißen Mann gut. Solidarität unter Privilegierten in Zeiten von #MeToo, #rechtsterrorismus und #Frauenquote uvm nur logisch

['Old white man likes old white man. Solidarity among the privileged in times of #MeToo, #farrightterrorism and #women'squota etc no surprise']

To sum up, we can observe a notable difference between the English and Spanish language community on the one hand and the German language on the other just when it comes to the way #MeToo is referred to. While English and Spanish discourse generally award it the status of a movement and also use other, more appreciative collocations, German discourse is more reluctant and just labels it a debate, which has fewer empowering features than the term *movement*. This may affect the degree to which #MeToo can be seen as a safe space for women participating in internet discussions surrounding the hashtag then.

It will be interesting to see whether this difference in how #MeToo is perceived on a general discourse level is also observable in individual author positions in the tweets. The second research objective is to conduct an analysis of author

stance towards #MeToo in the English, Spanish and German tweets (see Table 5.3, visualised in Figure 5.1). The tweets in the English and Spanish corpora show a higher frequency of positive stance (47% and 42%) when compared to the tweets in the German corpus (27%), and show a lower frequency of negative stance (27% and 29%) than the German tweets (44%). Keeping in mind that the sample size is not huge and provides just a snapshot of activity, the data seem to indicate that the general stance towards #MeToo is more negative in German Twitter discourse during the recorded time span than in English and Spanish.

Table 5.3: Analysis of positive, critical, negative and unclear author stance in the three languages

	positive		critical		negative		unclear		total	
	n	p	n	p	n	p	n	p	n	p
English	98	47%	8	4%	46	22%	58	27%	210	100%
Spanish	100	42%	17	7%	52	22%	68	29%	237	100%
German	50	27%	17	9%	81	44%	36	20%	183	100%

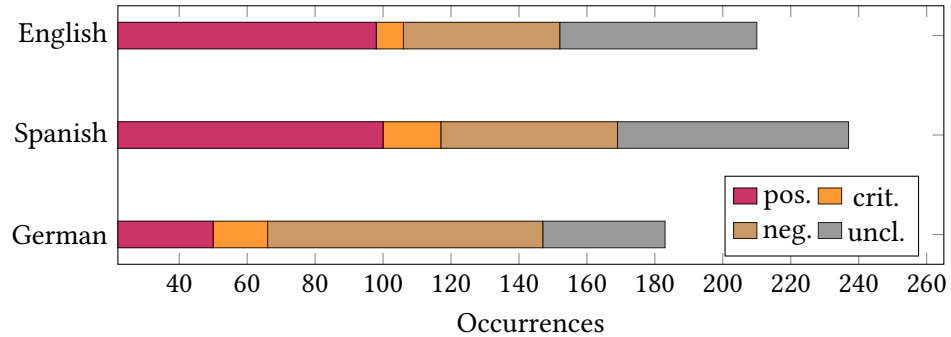


Figure 5.1: Graph of positive, critical, negative and unclear author stance in the three languages

The analysis thus suggests that in German-language Twitter discourse, contrary to English and Spanish-language discourse, the reluctance to call #MeToo a movement, or a lasting societal shift, but mainly a “debate”, is reflected in the general population’s stance towards the hashtag and translates into a more negative attitude towards it, again putting into doubt whether the hashtag in German can be considered a safe space. Of course the use of the word *debate* in itself, as the above examples suggest, does not necessarily imply negative author

Mario Bisiada

stance. It could be argued, however, that the prevalence of the “debate” frame has promoted a type of individualising tendency which, as pointed out by research discussed above, de-politicises the #MeToo movement. After all, a debate is understood as an open-ended discussion depending on individual views between at least two legitimate sides which have equal justification of existence. Given, however, that #MeToo was sparked by a series of revelations of sexual assault and developed into a movement to end gender violence in society, it is hard to see any justification for labelling it a “debate”. The often misogynistic and derisive statements we observe in this analysis, however, show that a less appreciative labelling of the hashtag accompanied by a largely negative discourse around it can endanger the perception of the hashtag as a safe space for feminist activism.

5 Cross-linguistic framings of #MeToo

Having indicated differences between the languages under analysis, I now turn to the third research question, the analysis of cross-linguistic similarities. One representation that is obvious from the collocates identified in the previous analysis is that of #MeToo as a collection of lies, which is perhaps the most lexically creative attack on the movement. Beyond that, this section identifies two framings that language users in this corpus apply to #MeToo across languages: the *organised pressure group* frame and the *exaggerated scope* frame. I also discuss the hijacking of the movement by far right groups.

The first frame that can be identified is that of #MeToo as an organised pressure group. A range of comments in the corpus suggest that commentators across languages treat #MeToo as a centrally controlled organisation:

@user @user @user @user The current #MeToo leadership are mostly the men hating lesbians. They are desperately trying to create a wedge between men and women. Therefore anything men do to honour and please women must be attacked and denigrated.

@user It specifically painted “hippie chicks” as childish, dirty, drugged out, homicidal morons whose violent deaths are to be celebrated. That had to upset more than one #MeToo architect over at Alyssa Milano’s agency, CAA. After all, Tarantino is repped by WMA, their competitor.

Lo que no entiendo es cómo, en el mismo mundo del movimiento #MeToo y del poderoso lobby feminista -uno de los grupos de presión más poderosos de la actualidad-, todavía pueda existir el reggaeton

[‘What I don’t understand is how, in this world of the #MeToo movement and the powerful feminist lobby -one of the most powerful pressure groups of our time-, there can still be Reggaeton.’]

Wieder ein Opfer der totalitären #MeToo-Inquisition, dessen Rehabilitierung am Ende mit einem Dreizeiler erledigt wird und der mit seinem sozialen und beruflichen Tod allein bleibt.

[‘Another victim of the totalitarian #MeToo inquisition whose rehabilitation in the end is given short shrift and who remains alone with his social and professional death.’]

Words like *architect*, *leadership* and *lobby* show that these authors perceive #MeToo not as the decentralised popular movement against everyday gender violence that it is, but as an organised campaign attributed to left-wing lobbies, claimed to be funded by George Soros and led by a few activist women with the political goal to eliminate men who disagree, as evidenced further in the examples below.

@user @irlembberlin @bpol_b Makes sense. Men are the key to the streets & have been propaganda targeted by #WhiteFeatherMedia & political #gaslighting; by #MeToo mass-emasculatation & dehumanisation of men – ready to *trigger* & mobilise a global army...

Was bleibt von #metoo? – die Motivation, politisch unliebsame Männer gesellschaftlich zu vernichten.

[‘What remains of #metoo? – the motivation to socially eliminate politically disagreeable men.’]

Al tío Neil las del #MeToo le van a inventar 5 violaciones por decir esto... y a ver qué le inventan los otros lobbys zurdos gringos

[‘The #MeToo women will plant 5 rapes on this guy Neil for saying that. And let’s see what the other damn left-wing lobbies will make up’]

Vean quien está detrás de #MeToo? [link to article *¿Las feministas de Soros detrás del #MeToo?*, published in *Atiempo.mx*]

[‘Do you see who’s behind #MeToo?’]

It is here that the anti-feminist reaction to #MeToo chimes in with far-right parties’ and supporters’ general allegation that their freedom of speech is curbed (Lang 2017; Salazar 2018: 140–141) while at the same time being extremely sensitive to criticisms of themselves. The effect of #MeToo that seems to disturb many men in general is an insecurity about the disruption of internalised patterns of behaviour that they deem normal and that are now challenged on a global scale, as expressed by this tweet:

Mario Bisiada

@user @politicaelle Sorry, but I'm not buying that shit. It all starts with innocently offering a bit of assistance with carry-on luggage and before you know it you're elbows deep in the #MeToo movement and God is angry at our collective impertinence. Just not worth it.

The #MeToo movement is not only derided, but also hijacked by far right interest groups, as has been reported by other scholars (Boyle & Rathnayake 2019; Wiens 2019). This hijacking usually consists in remarking on a “curious silence from #MeToo” (drawing on the “organised pressure group” frame identified above) on a case of sexual violence where the accused has a migratory background:

@user Und von den sog. “Feministinnen” und #metoo-Aktivistinnen wird bestenfalls ohrenbetäubendes Schweigen kommen. #mussmanwissen

[‘And the so-called “feminists” and #metoo activists will at best produce deafening silence. #havetoknowit’]

FeministInnenverbände die zu den Massenvergewaltigungen in Deutschland schweigen, brauchen auch nicht mehr mit #MeToo zu kommen, wenn sie mal von älteren Herren angesprochen werden.

[‘Feminist organisations that say nothing about the mass rapes in Germany might as well shut up about #MeToo when older men occasionally start talking to them.’]

@A3Noticias Parece que la importación de delincuentes no es una idea especialmente brillante. Curioso silencio de las #MeToo y las #YoSíTeCreo, por cierto.

[‘It seems that the importation of criminals is not an especially brilliant idea. Curious silence from the #MeToo and the #YoSíTeCreo women, actually’]

@user @user Machista? Que la cultura musulmana sólo es machista? Te violan SÓLO por ser MUJER OCCIDENTAL y NO MUSULMANA y tú hablas de machismo? Encaja el #Metoo en el Taharrush, en la violación de la niña española dejando marchar a la musulmana, los 17€ entre risas (españolaspumas) vamos!

[‘Chauvinist? Muslim culture is just chauvinist? They rape you JUST for being a WESTERN WOMAN and NOT MUSLIM and you talk about chauvinism? Try to fit #Metoo to the Taharrush, to the rape of the Spanish girl while the muslim girl was let go, laughing at the 17€ (Spanish whores), come on!’]

The hijacking of #MeToo to spread islamophobia is a cross-national phenomenon (Farris 2017; Mast 2018), and has happened in Germany under the hashtag #120dB, which is also observed in the corpus of this study. The #120dB campaign emerged with a video of German and Austrian women condemning “neglected” acts of sexual violence committed by migrants and refugees and is “an exemplary case

to illustrate how anti-immigration groups tap into women's voices in order to produce solidarities on the basis of discrimination and sociocultural exclusion" (Sorce 2018: 1124). Such groups selectively report cases of violence against women only when the alleged perpetrators can be stereotypically assigned to a possible migrant profile. The messages appear without overt evaluation, but the #120dB and other hashtags, attributing the violence to Merkel's welcoming policies, for instance, make the political background obvious.

Related to the "organised pressure group" frame is that of exaggerating the scope of the movement (see Franks 2019: 86 on the largely intangible consequences of #MeToo so far). That debates sparked by feminist hashtags become framed as exaggerations has also been observed by Kornemann (2018: 382) on the German #Aufschrei. One label that occurs here, perhaps not surprisingly, is that of #MeToo as a witch hunt, which has been used by prominent figures such as Catherine Deneuve and Michael Haneke (Mumford 2018; Clark-Parsons 2019: 3) and is observed in each language under analysis. Another cross-linguistically observable pattern in the corpora is the exaggerated importance given to dropped court cases such as the one against Kevin Spacey, described as a majorly important case rather than one of several, the backlash against Amber Heart in her case against Johnny Depp, or the frustration in many men about the cancellation of a Victoria's Secret fashion show.

Lastly, screw the #metoo movement for getting involved and the women who believe falsely accused men should be fired/arrested. They can't stop targeting Depp as guilty when turns out he was innocent all along and Amber Heart was the real abuser. #real-monsteramberheart.(5/5)

Wie ist das jetzt eigentlich mit #KevinSpacey ? Entschuldigt sich irgendjemand von diesem radikalfeministischen, hysterischen #metoo - Lynchmob? Wohl eher nicht, oder? Naja... War auch nicht anders zu erwarten von denen. Ist das gleiche, wie mit Nazis ? ['What's happening now with #KevinSpacey? Will anyone from this radically feminist, hysterical #metoo lynch mob apologise? Probably not, right? Oh well...Didn't expect anything different from them. It's the same as with the Nazis?']

Kevin Spacey reaparece. Cuanto daño ha hecho el puritanismo y la caza de brujas que desató el #MeToo!!! #KevinSpacey #Libertad #Freedom
['Kevin Spacey reappears. How much damage this puritanism and witch hunt that #MeToo unleashed has done!!!']

Recordemos esa maravilla que era el desfile de #VictoriaSecret y que le jodan a estos totalitarios del feminismo el #MeToo y @el_pais
['Let's commemorate the marvel that was the #VictoriaSecret fashion show and fuck those totalitarians of feminism, #MeToo and @el_pais']

Mario Bisiada

As #MeToo is probably one of the most widely known and influential feminist hashtags, it is perhaps hardly surprising that it has attracted the kinds of attacks and anti-feminist discourse discussed in this section. Given that, and even after its “coming of age” as a hashtag movement, it is all the more inspiring to see that even a snapshot analysis like this one shows the power of #MeToo to increase awareness of gender violence, to unite and give warmth and hope to women across language communities:

It's weirdly healing, always upsetting, and never surprising to bond with a woman over your experiences of sexual assault. #metoo my have shocked a lot of men but I can't imagine it shocked many women.

So happy to see this man has been arrested! 3 years ago he kept following and harassing my friend until she found refuge at a streetside dhaaba. She had no photos, no way to report him. This is the power of social media. This is why #MeToo exists and why it's needed.

@user Aber genau das gehört zur Bewertung vom “Leben”. Eine Vergewaltigung ist kein Unfall, der Vergewaltiger hat sich dazu entschlossen. Und lange wurde seine Tat rechtfertigt und ich ausgegrenzt, im Berufsumfeld. Zusätzlich zum eigentlichen Trauma. Das gehört alles dazu bei #metoo.

[‘But exactly that belongs to the assessment of “life”. A rape is not an accident, the rapist decided to do it. And his deed was justified for a long time and I was excluded in my professional life. In addition to the original trauma. All that is part of #metoo.’]

Después de esto, es imposible que alguien diga que el #metoo no sirve para nada. Las denuncias salen porque otras las empezaron y porque nos fortalecemos A DENUNCIAR en una plataforma que NOS CREE. Ojalá dejen de cuestionar pruebas cuando hay casos como estos. Fuente: @metooperu

[‘After this, it's impossible that anyone would say that #metoo has no effects. The reportings happen because others started them and because we gathered the strength TO SPEAK OUT on a platform that BELIEVES US. Hopefully they will stop questioning evidence in cases like this one. Source: @metooperu’]

6 Conclusion

This study has provided a snapshot analysis of English, Spanish and German discourse surrounding the #MeToo hashtag on Twitter in the months of July and August 2019. I have found that the semantic prosody of the #MeToo hashtag, i.e. the use of collocates to colour its meaning, is comparable among English and

Spanish users, who mainly label #MeToo a movement, while also using temporal collocates such as *moment*, *era* and *times*. German users, in contrast, do not frequently use the term *movement*, but seem to prefer the collocation #MeToo-*Debatte*, referring to #MeToo as a “debate”, a tendency that has been shown to echo common use in German newspapers. *Debate* is not used at all as a collocate in the English and Spanish data, while *movement* and other collocates framing #MeToo as influential rarely occur in German.

Whether this difference influences the public’s attitude towards #MeToo, something that the data analysed here tentatively indicate, should be investigated in greater depth in future studies. Cross-linguistic studies of this nature can be a great source of information to help understand the differing perceptions of and attitudes towards feminist hashtag activism, which is *per se* international and thus calls for transnational and cross-cultural analysis.

As the study draws on data gathered during a period of a few weeks and had to be based on a small enough sample size to facilitate qualitative analysis, its findings cannot be generalised, which is a general issue of hashtag-based sampling (Zappavigna 2018: 7). As such, it calls for further research into the semantic prosody of the #MeToo hashtag. A follow-up project might search specifically for collocations involving *movement* and *debate* and provide a diachronic overview of their evolution as well as a stance analysis, possibly also indicating diachronic shifts. A more nuanced understanding of how hashtag activism is picked up and framed by traditional media is required. As this study has indicated, the way a hashtag is framed may have consequences for the perception of a movement in a given language community.

Finally, this study has identified some cross-linguistic patterns of discourse surrounding the #MeToo hashtag, mainly intending to undermine its potential. The data analysed here both reflects known phenomena such as the hijacking of feminist movements to promote far right ideology and islamophobia and the exaggeration of its effects to stoke antipathy towards it, but also patterns that have not received much scholarly attention, such as the framing of #MeToo as an organised pressure group headed by a few individuals and with politically left-wing aims or the unbalanced attention given to few particular cases to undermine the movement.

As discussed above, hashtag campaigns play a significant role in all of these issues in social media debates, both in order to drive forward a movement and to attack it through counter hashtags. The linguistic study of hashtags and their framing, be it through semantic prosody or through general author stance, is therefore an important path towards an understanding of how hashtags, which

Mario Bisiada

have become key parts of everyday language use, affect the way we perceive and communicate feminist movements on social media and in society.

References

- Ahmed, Sara. 2004. *The cultural politics of emotion*. Edinburgh: Edinburgh University Press.
- Akhtar, Monica. 2017. #MeToo: A Movement or a Moment? <https://www.washingtonpost.com/news/the-intersect/wp/2017/11/09/metoo-a-movement-or-a-moment/> (25 November, 2019).
- Angus Reid Institute. 2018. #MeToo: Moment or Movement? <http://angusreid.org/me-too> (25 November, 2019).
- Banet-Weiser, Sarah. 2018. *Empowered: Popular feminism and popular misogyny*. Durham: Duke University Press.
- Bennett, Jessica. 2019. *The #MeToo Moment*. <https://www.nytimes.com/series/metoo-moment> (25 November, 2019).
- Bowles Eagle, Ryan. 2015. Loitering, lingering, hashtagging: Women reclaiming public space via #BoardtheBus, #StopStreetHarassment, and the #Everyday-Sexism project. *Feminist Media Studies* 15(2). 350–353. DOI: [10.1080/14680777.2015.1008748](https://doi.org/10.1080/14680777.2015.1008748).
- Boyle, Karen & Chamil Rathnayake. 2019. #HimToo and the networking of misogyny in the age of #metoo. *Feminist Media Studies* Advance Online Access. 1–19. DOI: [10.1080/14680777.2019.1661868](https://doi.org/10.1080/14680777.2019.1661868).
- Burke, Tarana. 2018. *Me Too Is a Movement, not a Moment*. https://www.ted.com/talks/tarana_burke_me_too_is_a_movement_not_a_moment/transcript (25 November, 2019).
- Clark, Rosemary. 2016. “Hope in a Hashtag”: The discursive activism of #Why-IStayed. *Feminist Media Studies* 16(5). 788–804. DOI: [10.1080/14680777.2016.1138235](https://doi.org/10.1080/14680777.2016.1138235).
- Clark-Parsons, Rosemary. 2019. I SEE YOU, I BELIEVE YOU, I STAND WITH YOU: #Metoo and the performance of networked feminist visibility. *Feminist Media Studies* Advance Online Access. 1–19. DOI: [10.1080/14680777.2019.1628797](https://doi.org/10.1080/14680777.2019.1628797).
- De Benedictis, Sara, Shani Orgad & Catherine Rottenberg. 2019. #Metoo, popular feminism and the news: A content analysis of UK newspaper coverage. *European Journal of Cultural Studies* 22(5-6). 718–738. DOI: [10.1177/1367549419856831](https://doi.org/10.1177/1367549419856831).
- Di Stefano, Mark. 2019. *A New Poll Has Found a Third of Young British People Have Anti-Feminist Views*. <https://www.buzzfeed.com/markdistefano/new-poll-third-young-british-males-anti-feminism> (25 November, 2019).

- Drücke, Ricarda & Elke Zobl. 2016. Online feminist protest against sexism: The German-language hashtag #aufschrei. *Feminist Media Studies* 16(1). 35–54. DOI: [10.1080/14680777.2015.1093071](https://doi.org/10.1080/14680777.2015.1093071).
- DWDS. 2019. *DWDS word trajectory for MeToo, MeToo-Debatte, MeToo-Bewegung, MeToo-Ära, created by the Digital Dictionary of the German Language*. <https://www.dwds.de/r/plot?view=1&corpus=zeitungen&norm=date%2Bclass&smooth=spline&genres=0&grand=1&slice=1&prune=0&window=3&wbase=0&logavg=0&logscale=0&xrange=2013%3A2018&q1=MeToo&q2=MeToo-Debatte&q3=MeToo-Bewegung&q4=MeToo-%C3%84ra> (25 November, 2019).
- El País. 2018. *Revolución MeToo: Un año del grito de las mujeres contra el acoso y la violencia sexual: qué ha cambiado*. https://elpais.com/tag/movimiento_metoo/ (25 November, 2019).
- Farris, Sara R. 2017. *In the name of women's rights: The rise of femonationalism*. Durham: Duke University Press.
- Fileborn, Bianca & Rachel Loney-Howes (eds.). 2019. *#Metoo and the politics of social change*. London: Palgrave Macmillan. DOI: [10.1007/978-3-030-15213-0](https://doi.org/10.1007/978-3-030-15213-0).
- Follman, Mark & Olivia Exstrum. 2019. *Armed and Misogynist: How Toxic Masculinity Fuels Mass Shootings*. <https://www.motherjones.com/crime-justice/2019/06/domestic-violence-misogyny-incels-mass-shootings/> (25 November, 2019).
- Franks, Mary Anne. 2019. A thousand and One stories: Myth and the #metoo movement. In Bianca Fileborn & Rachel Loney-Howes (eds.), *#Metoo and the politics of social change*, 85–95. London: Palgrave Macmillan.
- Garibotti, María Cecilia & Cecilia Marcela Hopp. 2019. Substitution activism: The impact of #metoo in argentina. In Bianca Fileborn & Rachel Loney-Howes (eds.), *#Metoo and the politics of social change*, 185–199. London: Palgrave Macmillan.
- Gill, Rosalind & Shani Orgad. 2018. The shifting terrain of sex and power: From the “Sexualization of Culture” to #metoo. *Sexualities* 21(8). 1313–1324. DOI: [10.1177/1363460718794647](https://doi.org/10.1177/1363460718794647).
- Ging, Debbie. 2017. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities* 22(4). 638–657. DOI: [10.1177/1097184X17706401](https://doi.org/10.1177/1097184X17706401).
- Huelga Feminista. 2018. *International*. <http://hacialahuelgafeminista.org/international/> (25 November, 2019).
- Jackson, Sarah J. & Sonia Banaszczyk. 2016. Digital standpoints: Debating gendered violence and racial exclusions in the feminist counterpublic. *Journal of Communication Enquiry* 40(4). 391–407. DOI: [10.1177/0196859916667731](https://doi.org/10.1177/0196859916667731).

Mario Bisiada

- Jaki, Sylvia, Tom De Smedt, Maja Gwózdź, Rudresh Panchal, Alexander Rossa & Guy De Pauw. 2019. Online hatred of women in the Incels. Me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict* Advance Online Access. 240–268. DOI: [10.1075/jlac.00026.jak](https://doi.org/10.1075/jlac.00026.jak).
- Kahlke Lorentzen, Maia & Kevin Shakir. 2019. *Den hvide internetterrorist er også anti-feminist*. <https://cybernauterne.dk/blog/den-hvide-internetterrorist-er-ogsaa-anti-feminist/> (25 November, 2019).
- Keller, Jessalynn, Kaitlynn Mendes & Jessica Ringrose. 2018. Speaking “Unspeakable Things”: Documenting digital feminist responses to rape culture. *Journal of Gender Studies* 27(1). 22–36. DOI: [10.1080/09589236.2016.1211511](https://doi.org/10.1080/09589236.2016.1211511).
- Kornemann, Laureen. 2018. Die Sexismus-Debatte in der deutschen Öffentlichkeit – Brüderle vs. #aufschrei. In Margreth Lünenborg & Saskia Sell (eds.), *Politischer Journalismus im Fokus der Journalistik*, 369–390. Wiesbaden: Springer. DOI: [10.1007/978-3-658-18339-4_15](https://doi.org/10.1007/978-3-658-18339-4_15).
- Kunst, Jonas R., April Bailey, Claire Prendergast & Aleksander Gundersena. 2018. Sexism, rape myths and feminist identification explain gender differences in attitudes toward the #metoo social media campaign in two countries. *Media Psychology* Advance Online Access. 1–26. DOI: [10.1080/15213269.2018.1532300](https://doi.org/10.1080/15213269.2018.1532300).
- Lang, Juliane. 2017. Feindbild Feminismus: Familien- und Geschlechterpolitik in der AfD. In Stephan Grigat (ed.), *AfD & FPÖ: Antisemitismus, völkischer Nationalismus und Geschlechterbilder*, 61–78. Baden-Baden: Nomos.
- Lang, Mervyn Francis. 1980. *Spanish word formation: Productive derivational morphology in the modern lexis*. Abingdon: Routledge.
- Lewandowska-Tomaszczyk, Barbara. 1996. Cross-linguistic and language-specific aspects of semantic prosody. *Language Sciences* 18(1). 153–178.
- Louw, Bill. 1993. Irony in the text or insincerity in the writer?: The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 157–176. Amsterdam: John Benjamins.
- Lünenborg, Margreth & Tanja Maier. 2013. *Gender Media Studies: Eine Einführung*. Konstanz: UVK Verlagsgesellschaft mbH.
- Maireder, Axel & Stephan Schlögl. 2014. 24 hours of an #outcry: The networked publics of a socio-political debate. *European Journal of Communication* 29(6). 687–702. DOI: [10.1177/0267323114545710](https://doi.org/10.1177/0267323114545710).
- Mast, Nina. 2018. *Far-Right Activists and “Alt-Right” Trolls are Using the #MeToo Movement to Bolster their Xenophobia*. https://www.mediamatters.org/breitbart-news/far-right-activists-and-alt-right-trolls-are-using-metoo-movement-bolster-their?redirect_source=/blog/2018/02/05/far-right

- activists - and - alt - right - trolls - are - using - metoo - movement - bolster - their - xenophobia/219260 (25 November, 2019).
- me too. 2018. *You are not alone*. <https://metoomvmt.org> (25 November, 2019).
- Mendes, Kaitlynn, Jessica Ringrose & Jessalynn Keller. 2018. #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies* 25(2). 236–246. DOI: [10.1177 / 1350506818765318](https://doi.org/10.1177/1350506818765318).
- Mumford, Gwilym. 2018. *Michael Haneke: #MeToo has led to a witch hunt “coloured by a hatred of men”*. <https://www.theguardian.com/film/2018/feb/12/michael-haneke-metoo-witch-hunt-coloured-hatred-men> (25 November, 2019).
- Oppenheim, Maya. 2018. *International Women's Day: Hundreds of Trains Cancelled as Spanish Women Walk out in First “Feminist Strike”*. <https://www.independent.co.uk/news/world/europe/international-womens-day-spain-feminist-strike-protests-work-walkout-discrimination-gender-pay-gap-a8245696.html> (25 November, 2019).
- Orgad, Shani & Sara de Benedictis. 2015. The “Stay-at-Home” mother, postfeminism and neoliberalism: Content analysis of UK news coverage. *European Journal of Communication* 30(4). 418–436. DOI: [10.1177/0267323115586724](https://doi.org/10.1177/0267323115586724).
- Pennington, Rosemary. 2018. Making space in social media: #Muslimwomens-day in twitter. *Journal of Communication Inquiry* 42(3). 199–217. DOI: [10.1177 / 0196859918768797](https://doi.org/10.1177/0196859918768797).
- Rentschler, Carrie. 2015. #Safetytipsforladies: Feminist Twitter takedowns of victim blaming. *Feminist Media Studies* 15(2). 353–356. DOI: [10.1080 / 14680777. 2015.1008749](https://doi.org/10.1080/14680777.2015.1008749).
- Salazar, Philippe-Joseph. 2018. The Alt-Right as a community of discourse. *Javnost – The Public* 25(1–2). 135–143. DOI: [10.1080/13183222.2018.1423947](https://doi.org/10.1080/13183222.2018.1423947).
- Schachtner, Christina & Gabriele Winter (eds.). 2005. *Virtuelle Räume – neue Öffentlichkeiten: Frauennetze im Internet*. Frankfurt: Campus Verlag.
- Scott, Kate. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics* 81. 8–20. DOI: [10.1016/j.pragma.2015.03.015](https://doi.org/10.1016/j.pragma.2015.03.015).
- Senecal, Lisa. 2018. *Is #MeToo a Movement or a Moment?* <https://www.thedailybeast.com/is-metoo-a-movement-or-a-moment> (25 November, 2019).
- Sorce, Giuliana. 2018. Sounding the alarm for right-wing #metoo: “120 Dezibel” in germany. *Feminist Media Studies* 18(6). 1123–1126. DOI: [10.1080 / 14680777. 2018.1532146](https://doi.org/10.1080/14680777.2018.1532146).
- Starkey, Jesse C., Amy Koerber, Miglena Sternadori & Bethany Pitchford. 2019. #Metoo goes global: Media framing of silence breakers in four national

Mario Bisiada

- settings. *Journal of Communication Inquiry* 43(4). 437–461. DOI: [10.1177/0196859919865254](https://doi.org/10.1177/0196859919865254).
- Stewart, Dominic. 2010. *Semantic prosody: A critical evaluation*. London: Routledge.
- Svokos, Alexandra. 2019. *Alleged Dayton Gunman Connor Betts Showed Signs of Misogyny, Mirroring a Grim Pattern for Shooters*. <https://abcnews.go.com/US/alleged-dayton-gunman-connor-betts-showed-signs-misogyny/story?id=64826324> (25 November, 2019).
- Twitter. 2019. *About Verified Accounts*. <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts> (25 November, 2019).
- Valens, Ana. 2018. *Why Incels are Nothing to Laugh at*. <https://www.dailydot.com/irl/what-is-an-incel/> (25 November, 2019).
- Vessey, Rachelle. 2013. Challenges in cross-linguistic corpus-assisted discourse studies. *Corpora* 13(1). 1–26. DOI: [10.3366/cor.2013.0032](https://doi.org/10.3366/cor.2013.0032).
- Wielens, Alia. 2019. “Wo bleibt euer Aufschrei?” Rassistische Umdeutungen von #aufschrei und #metoo durch Identitäre Frauen. *Femina Politica – Zeitschrift für feministische Politikwissenschaft* 1. 111–120. DOI: [10.3224/feminapolitica.v28i1.10](https://doi.org/10.3224/feminapolitica.v28i1.10).
- Williams, Sherri. 2015. Digital defense: Black feminists resist violence with hashtag activism. *Feminist Media Studies* 15(2). 341–344. DOI: [10.1080/14680777.2015.1008744](https://doi.org/10.1080/14680777.2015.1008744).
- Zappavigna, Michele. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25(3). 274–291. DOI: [10.1080/10350330.2014.996948](https://doi.org/10.1080/10350330.2014.996948).
- Zappavigna, Michele. 2018. *Searchable talk: Hashtags and social media metadiscourse*. London: Bloomsbury.
- Zarkov, Dubravka & Kathy Davis. 2018. Ambiguities and dilemmas around #metoo: #Forhowlong and #where to? *European Journal of Women’s Studies* 25(1). 3–9. DOI: [10.1177/1350506817749436](https://doi.org/10.1177/1350506817749436).

Chapter 6

Investigating patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German

Felix Hoberg

Leipzig University

This paper investigates the patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German. As being part of an overall evaluation of the Skype Translator on a dialogue-oriented level, a case study on 21 German-speaking participants was conducted. Despite not having any proficiency in Catalan, these participants had to text-chat with Catalan native speakers via Skype, while the Skype Translator was activated. The sessions were observed by an eye tracking system. The collected data thus represents a naturalistic starting point to evaluate how users structure computer-mediated communication situations when real-time machine translation is involved while having to rely on that output.

1 Introduction

Automatic language processing, auto speech recognition and machine translation (MT) are considered valuable innovations by the language industry. However, progress in this field is still viewed skeptically, which in turn calls for continuous evaluation of the aforementioned systems (i. e. [Ramlow 2009](#); [Bowker & Ciro 2019](#)). There are indeed different metrics and standards which allow for categorically evaluate the machine-translated output either manually or automatically (see §2.3).



Felix Hoberg

Especially when it comes to dialogic interactions between humans and MT, research has so far tackled either the interactive or the technological aspect, but seldom both of them at once. Microsoft's Skype Translator will thus serve as a central element in this case study, as it offers real-time machine translation in 10 languages in voice and video chats and 60 languages in text chats.

The general aim of this entire project is to highlight how MT evaluation can be applied on a dialogue-oriented level to services like the Skype Translator where all messages are displayed in a two-column-design with outgoing messages right-aligned and all MT output and incoming messages left-aligned. This study hence does not intend to offer an evaluative application of MT quality metrics on the Skype Translator's output but to outline the users' perception and behaviour when it comes to using the machine-translated output in a real-time conversation. Thus, the present article combines research in the fields of communication research (i. e. [Beißwenger 2007](#)) and machine translation (i. e. [Fišer & Beißwenger 2017](#)).

To examine the users' behaviour, an exploratory eye-tracking-based case study was carried out. In that study, Skype Translator-mediated text chats between German and Catalan native speakers were captured in order to investigate the eye movement patterns on characteristic areas of interest of the Skype Translator, namely the entry mask and each single text chat message box (see [Fig. 6.1](#), p. 128).

This paper's guiding research question thus is how the participants are perceiving the incoming and outgoing text messages. Based on the assumption that the MT output into German will need more attention than the other messages and that Catalan messages will be nonetheless taken into account (as the – possibly error-prone – new information is presented in both languages), it has to be investigated how participants handle this bilingual input. Special attention will be drawn upon saccadic eye movements.

For that reason, [§2](#) introduces the theoretical background in terms of research on dialogue and computer-mediated conversation in the context of computer-mediated communication and previous findings on eye movements in reading tasks. [§3](#) gives insights on the overall project conception, before explaining in detail to which extent the collected data is used for this analysis. Then, [§4](#) presents the results of the saccadic eye tracking data and situates them along the theoretical background (see [§5](#)), before the conclusion in [§6](#) sums up the analysis, going back to the overall project.

2 Background

2.1 Research on dialogue and conversation

Since the early 1990s, various concepts in communication research have been modelled and restructured to fit modern computer-mediated communication (cf. [Fišer & Beißwenger 2017: 7](#)). Apart from taking a look at global concepts such as text, sender, recipient or conversation, the interest in research has now passed on to questions which reflect the transitional processes web-based communication has undergone over the last two decades: How do we interact online? How does online interaction change our ways of communicating? Can we still speak of sender and recipient after all? How do we cope with this extensive amount of data and the rising machine learning technologies? (cf. [Beißwenger 2007](#)).

These questions also implicitly refer to the phenomena of turn-taking and speaker switch or the rising use of the term *hypertext* to describe digital textual behaviour (cf. [Storrer 2001](#)), central elements which have already been extensively studied regarding analogue, face-to-face and monolingual web-based communication, but so far have not been adopted to bilingual, machine-translated, web-based conversations such as presented in this paper. This gap might be attributed to the fact that online communication follows different rules than offline communication.

There are two obvious differences between oral, face-to-face and chat communication. The latter appears in written or typed form and lacks almost all non- and paraverbal elements like gesture, intonation or eye contact etc. which usually help to structure the communication act (cf. [Beißwenger 2007: 172](#)). In contrast, an online chat message passes through more sections between sender and addressee than an oral, face-to-face talk. From the sender's mind, it goes from typing on the keyboard to the computers' short-time memory and from there to the server the software in use is connected to. From that server it goes to the addressee's software and is subsequently processed by the computer to be displayed on screen before the addressee can spend cognitive resources on it (cf. [Beißwenger 2017: 146](#)). In the case of the Skype Translator, one has additionally to take into account the time it takes to send, machine-translate and receive the original message. In case of high latency, this time gap can have a severe impact on communication, because while the receiving person is still answering one incoming message, the other may already have sent another text. This can result in an asynchronous communication.

Thus, the use of computer-mediated communication technology, and in this case more precisely the Skype Translator, leads to a change in the communication process of sending and receiving messages. A text chat message has to be

Felix Hoberg

completely written before it can be sent¹ and it has to be received and completely read before it can be reacted to. At the same time, as opposed to oral communication, the communication partners are not necessarily in the same location, nor near to each other at all (cf. *ibid.*: 146). *Storrer* (cf. 2001: 3) points out another important feature: even though online chatting appears mostly in written form, it follows the rules of oral production. The relationship of officially standardized language and its informal, but also widely accepted online communication use, which follows its own rules has been object of many research projects ever since, as for example in *Verheijen* (2017) in the context for Dutch. This relationship might helpfully be investigated by an eye tracking study.

Consequently, the indicators explained below in §2.4.1 can be taken as initial points of reference on how the participants process the information on screen when text-chatting with people, whose language they do not speak.

2.2 The Skype translator

As has already been stated in the introduction, Skype features a real-time translation engine called Skype Translator for text chats between 60 different languages and for voice and video chats between eleven languages². Both the written and the video or voice real-time translation engine are based on machine learning and Microsoft's proprietary neural machine translation system, meaning that the output is supposed to enhance in terms of quality by every time the feature (and any other product of Microsoft) is used. Additionally, some of the supported languages come with language detection, text-to-speech, speech-to-text, transliteration, a dictionary and the possibility of customizing the output according to individual terminology.³

2.3 Machine Translation Evaluation

There are several manual or automatic methods to evaluate the translation quality in general. With the expanding use of machine translation, evaluation methods are being adopted to the new environments (see e.g. *multidimensional quality metrics*⁴, *LISA QA* or *SAE J2450*⁵). Not only under the cloak of post-editing (cf.

¹Real-time text chat, where the text is transmitted immediately so that every user can observe the production process, will not be considered here.

²see <https://www.skype.com/en/features/skype-translator/>, last access on 4 November 2020.

³see <https://www.microsoft.com/en-us/translator/business/languages/>, last access on 4 November 2020.

⁴see <http://www.qt21.eu/quality-metrics/>, last access on 4 November 2020.

⁵see <https://blog.taus.net/the-8-most-used-standards-and-metrics-for-translation-quality-evaluation>, last access on 4 November 2020.

Vardaro et al. 2019: 2), but also with respect to raw MT output, automatic MT evaluation metrics are being modelled and investigated. (cf. Doherty & O'Brien 2014)

Most metrics and standards are designed to provide results that are comparable in quality to human translations, but are based on rather subjective ground since even the most automatic metrics often compare MT output to human reference translations. Another closely-related problem is the vast amount of different aspects to account for when evaluating MT systems (name entities, lexical issues, syntactic issues etc.) (Han 2018: 2f.). In contrast, „eye tracking could remove much of the subjectivity involved in human evaluation of machine translation quality as the processes it measures are largely unconscious.“ (cf. Doherty et al. 2010: 12) Furthermore, „[e]ye tracking has been used successfully as a technique for measuring cognitive load in reading, psycholinguistics, writing, language acquisition etc. for some time now.“ (ibid.) From another point of view, „[i]nclusion of users in evaluation of MT systems can provide benefits in both directions: such as positive influences on system development and its usability“ (Doherty & O'Brien 2014: 4) to thereby improve the system's performance, output and efficiency.

2.4 Eye-tracking and machine translation evaluation

Making sense of the process that leads to a final translated product has been object of translation studies for decades. There are multiple tools and methods to acquire information on the current cognitive processes of (mostly student) translators when asked to translate something: think-aloud protocols, corpus studies, product evaluations, comprehensibility tests, stimulated recall interviews.

On the contrary, „[r]ecords of eye movements, however, can do this very unobtrusively“ Schaeffer et al. (2017: 23), since it has been pointed out that „[c]ertain characteristics of readers' eye movements have been shown to be sensitive to the underlying cognitive processes involved in lexically identifying words“ (ibid.). Additionally, as has already been stated in §2.3, MT evaluation always has to keep an eye on usability and employability of the respective system and MT output. In consequence, using eye-tracking methods in translation process research leads to a better understanding of the effectiveness, efficiency and satisfaction of the task that is completed by a specific user Doherty & O'Brien (cf. 2014: 6).

Therefore, instead of being closely guided by the quality metrics for MT evaluation, which all aim to possibly reach error-free (almost human) quality, the investigation of Skype Translator-mediated conversations focuses on the usefulness and usability of the MT output in general and the way of users making sense

Felix Hoberg

of what they are reading. [Doherty & O'Brien \(cf. 2014: 4\)](#) e.g. state that „there are relatively few studies on the usability of raw machine translated output“. There is few research done yet on real-time chat communication – and even less on bi- or multilingual machine-translated communication. A study using eye tracking methods firstly explores the perception of software like this.

2.4.1 Eye-tracking, saccadic eye movements and the Skype Translator

Based on the communication research background above, it has to be made clear that this article focuses on Skype's text chat function, that is, on written communication. Similar issues concerning voice and video chat will not be discussed here, since Catalan is not supported in those modes. That being stated, the focus passes on to written text and its perception by its readers (or users), which is continuously being investigated in eye tracking studies. Apart from fixations, saccadic eye movements can be taken as an early measure of cognitive load and mental processing. As has already been investigated, saccades vary among different kinds of reading tasks ([Rayner 1998: 373](#)). [Jacobson & Dodwell 1979](#) for example studied left-to-right and vice versa directed saccades on (pseudo-)words, showing „that the probabilities of word components (letters, bigrams, etc.) can affect the speed with which words must be synthesized from their components before recognition occurs“ (ibid., p.313). [Schaeffer et al. \(2017: 24\)](#) hypothesise that proofreading a text requires more cognitive load than reading for comprehension. They found out that saccades made during proofreading were shorter than during reading for comprehension. With respect to the Skype Translator, name entities, numbers or words of similar characters in all the involved languages may represent a comparable challenge.

More precisely, respective studies also require fine-grained equipment to capture those high-velocity movements. In this context, [Leube et al. 2017](#) point out the diverging quality of capturing saccades with mobile eye tracking systems with a sampling rate of 60 or 120 Hz and a stationary system with 1000 Hz. This is important, since saccade duration mostly tends to range between 10 to 100ms (cf. [Duchowski 2017: 40](#)). Saccades represent movements of multiple characteristics that include blinks, regressions, corrections and glissades. All of theses have to be kept in mind and will be investigated in upcoming studies.

The present article focuses exclusively on saccade amplitude and duration as both are well described in scientific literature and thus widely used. They are defined as follows: „The saccadic amplitude (...) is the distance travelled by a saccade from its onset to the offset. The unit is typically given in visual degrees (°) or pixels (...)“ ([Holmqvist 2011: 312](#)).

During reading, for instance, saccadic amplitude is known to adapt to combined physical, physiological, and cognitive factors. Reading saccades are limited in length by the visual spanwidth which is around 7-8 letters (2°) in the average reading situation. (Holmqvist 2011: 312)

Shorter saccades in terms of amplitude are made if a text is complex and thus difficult to read, which in turn can be taken as indicative for increased cognitive load. (Schaeffer et al. 2017: 24) Similarly, reduced saccade amplitude occurs when a participant inspects something carefully.

Saccadic duration ('transition time'; not the same as transitions between AOIs) is defined as the time the saccade takes to move between two fixations or instances of smooth pursuit. (Holmqvist 2011: 321)

A longer saccadic duration can be taken as indicative for processing more difficult tasks (ibid.). „Thus, as text gets more difficult, fixations get longer, saccades get shorter, and more regressions are made“ (Rayner 2009: 1460).

This article is therefore based on the assumption that, given a bilingual, machine-translated reading and text-chatting task, the saccade amplitude and duration varies depending on the different languages (Catalan vs. German) and text types (MT vs. original). It is then interesting to take a look at how the difficulty of reading MT output and foreign language differs in real-time text chat communication. The last claim on investigating saccades is the general question of how useful this indicator is in general when looking at reading behaviour in text chat communication.

3 Research design

3.1 Participants and task

For this study, 25 students with no proficiency in Catalan were recruited. The legal consent on the anonymous processing of their data was obtained explicitly before the study started and all participants were debriefed after having taken part. They also were rewarded with 10 euro each. Of those 25 participants, four had to be excluded due to insufficient data quality. Of the remaining cohort, 20 were students of the Leipzig University and one was a student of the Leipzig University of Applied Sciences (HTWK). As the call for participation was sent to almost all departments of these two universities, the participants vary in terms of programs they are enrolled in.

Felix Hoberg

Three Catalan native speakers – two female and one male, ages 26, 24 and 26, respectively – were recruited as text chat counterparts for this study. All three came from different cities in the Catalan Countries: Valencia, Girona and Barcelona. All three were proficient in German since they took part in an exchange program during their studies and/or lived in Germany for a while.

Considering the amount of time each session took, the restricted access to the eye tracking system and the individual availability of every single participant, it was impossible to have the German participants text-chatting with always the same single Catalan native speaker. Recruiting only one Catalan native speaker would definitely have contributed positively to the comparability of the study, though, but this option was rejected facing the problem of recruiting students which were supposed to meet the above mentioned multiple conditions.

The task the German participants had to fulfill was split into three steps. First, they were asked to answer a questionnaire on their communication behaviour and their foreign language proficiencies. Second came the text chat session with a Catalan native speaker via Skype, with the Skype Translator activated. This part was captured by an eye tracking system. In order to get comparable data, the participants were given an introductory instruction: To have a central theme the participants could chat about, they were told to imagine they were about to spend a year abroad in Catalonia trying to get some information in advance on where to live and how to start there. Therefore, they were contacting the Catalan native speaker. On the one hand, this task allowed the participants to text-chat freely in a naturalistic manner according to their individual communication behaviour. On the other hand, this constraining task was intended to produce comparable linguistic data, which can be analyzed in possibly upcoming corpus studies. Lastly, to get an impression of the participant's individual experience during the Skype session, they were asked to fill out another questionnaire afterwards concerning the output quality of the Skype Translator.

The introductory questionnaire provides additional data regarding the composition of the cohort. The students participants mean age was 23.7 (SD = 4.0, range = 20–32 years). When it comes to (foreign) language proficiency with the Common European Framework of Reference for Languages (CEFR) as criterion, all of them indicated German as their first language with respect to use in ordinary and work life. 17 participants had English as a foreign language. As for roman languages, French and Spanish were reported nine times each, Italian and Portuguese one time each. Possible influences of roman language proficiencies on the participants' behaviour have to be taken into consideration in a full-range analysis, but will not be discussed in this article.

Looking at the user behaviour regarding Skype, only 17 participants reported using the software, but 13 of them even less than once per month. With regards to the duration per session, four participants used Skype no longer than 15 minutes, five no longer than 30 minutes, four up to one hour and four even beyond one hour.

The next part of the questionnaire was devoted to the use of alternative software, which includes all of the Skype's functions or just some of them, such as voice chat, followed by a detailed inquiry on alternatives for the individual Skype functions voice chat, video chat and text chat. Of 17 participants using alternatives, 16 used WhatsApp for voice chats, 15 for video chats and 16 again for text chats. Some participants stated that they were using other alternatives such as Telegram or Discord, too. Only three of them declared Skype as their preferred and most used software for video chats. As for voice or text chat, Skype was mentioned zero times as preferred and most used software. Instead, WhatsApp was indicated to be used most times. Last, the questionnaire took into account the participants' experience of living abroad. 13 of them have reported some experience living abroad with a mean of 30.53 months ($SD = 36.36$, range = 1–108 months).

In summary, this questionnaire draws a picture of the participants' high familiarity with communication software and their proficiency in at least one, but often even two or more languages apart from their mother tongue. The latter observation is also supported by the high range of experience in living abroad. Taking a closer look at Skype, the software is not the primary mean of communication but other wide-spread, mobile applications such as WhatsApp. This leaves room for two opposed suggestions: Either the participants rely on their foreign language skills and foreign culture experiences and thus do not need a machine-translated communication feature like the Skype Translator or – vice versa – as the questionnaire insinuates, the participants are hardly aware of this feature and thus have not made use of it. Both suggestions can be used to strengthen the claim for investigating the users' behaviour when communicating via machine-translated output.

3.2 Data collection

The *EyeLink Portable Duo* eye tracking system was used to conduct the study. The sessions were recorded in the *head-free-to-move* setup at a sampling rate of 1000Hz and binocular tracing. The overall setup included an eye tracking camera on a tripod, which was placed directly between the screen and keyboard – around 60–70cm from the participants' head –, a display computer with Skype and the

Felix Hoberg

screen captioning software packages installed, and a host computer to handle the eye tracking system. The software in use also allowed to capture messages (buttons pressed etc.).

The core element of this study was the latest version of Skype to that date (8.x), which already presented the Skype Translator as a built-in feature. The only requirement was to start a new conversation and add the Skype Translator service by clicking on the respective button in the user's profile one wanted to chat with. The service displayed messages in a two column structure: original messages of the user appear right-aligned, the MT output of the user, and the counterpart's incoming messages and the respective MT output appear left-aligned (see Fig. 6.1). During all sessions, Skype was displayed on maximum on screen to ensure equal quality for every participant and recording session. Nevertheless, Skype does not allow to use bigger font size in order to identify single words as AOI as is often recommended for reading studies (cf. O'Brien 2009: 261). That is why the data preparation (see Subsection 3.3) is restricted to the chat message level. The proprietary EyeLink Data Viewer-Software (SR Research Ltd. 2019) was used to process the raw eye tracking data. R version 3.4.3, (R Development Core Team 2019) and RStudio were consecutively used to analyse the processed data.

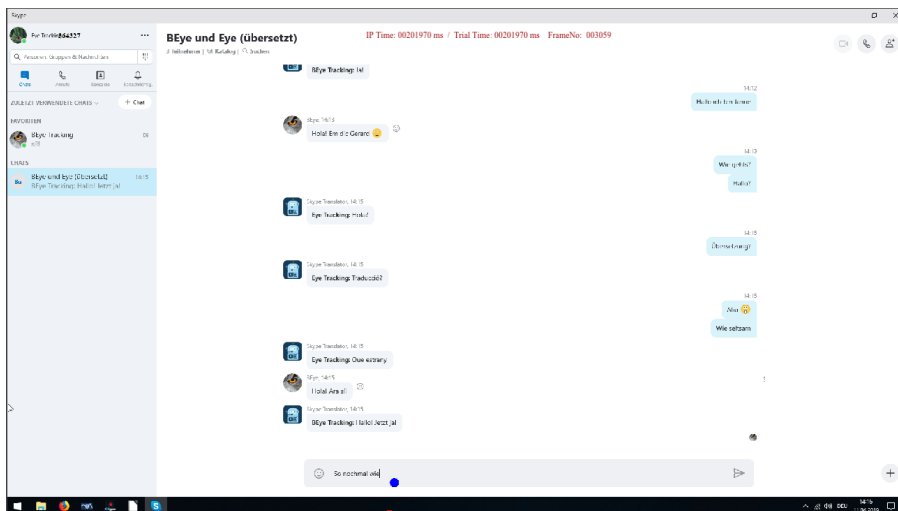


Figure 6.1: Example of text boxes in Skype. Left-aligned (grey): incoming messages and all MT output. Right-aligned (light blue): original messages of the participant.

3.3 Data preparation

There are two kinds of analyzable data that come from this study. On the one hand, there is the bilingual, authentic linguistic material produced by the participants, the Catalan native speakers and the machine translation of Skype which can be subdivided into four categories: the German and the Catalan original and the machine translated output, respectively. This kind will be spared for further research and publications.

On the other hand, there are the screen captions of the eye tracking sessions. These had to be annotated with dynamic areas of interest as the single text panels in Skype move when a new message is displayed on screen. To allow for a detailed analysis of those four linguistic categories mentioned above, every text box of each session is marked by its own consecutively numbered area of interest (see Fig. 6.1). Following the language codes proposed by ISO-639-2⁶, the following abbreviations were used to label those areas of interest: **GerO** – *German original*, **GerMT** – *Machine Translation into German*, **CatO** – *Catalan Original* and **CatMT** – *Machine Translation into Catalan*. The (static) entry mask was labelled **Entry**. Moreover, these five categories allowed for a detailed analysis of the eye tracking data as it was thus possible to create subsets sorted by participants, by label, by participant and label or other indicators.

The aforementioned 21 eye tracking sessions resulted in video material of a total duration of 375 minutes, or 18 minutes on average per trial. Taking the interest area count as measure, the mean count of German text messages is 21 (SD = 9.60, range = 6–48), of machine translated messages into Catalan 20 (SD = 9.79, range = 6–48), of Catalan text messages 27 (SD = 10.85, range = 11–49) and of machine translated messages into German 26 (SD = 10.66, range = 11–49). A diverging number of original and MT messages can be explained by the Skype Translator's MT output that was for no obvious reason automatically merged into one text box even if two original messages were written.

As one can see in Fig. 6.2 and 6.3, most attention is paid to the lower third of the screen, right above the left area of the entry mask which is where new incoming messages and the MT output are displayed. The screenshots of one participant depicted here stand for every other test person as the fixation heat maps (Fig. 6.3) and saccadic eye movement patterns (Fig. 6.2) look similar. Moreover, there are some remarkably large saccades that even reach above the recognizable screen size (cf. Leube et al. 2017: 6). One explanation might be that the participants were

⁶see <https://www.bib-bvb.de/web/kkb-online/rda-sprachencode-nach-iso-639>, last access on 4 November 2020.

Felix Hoberg

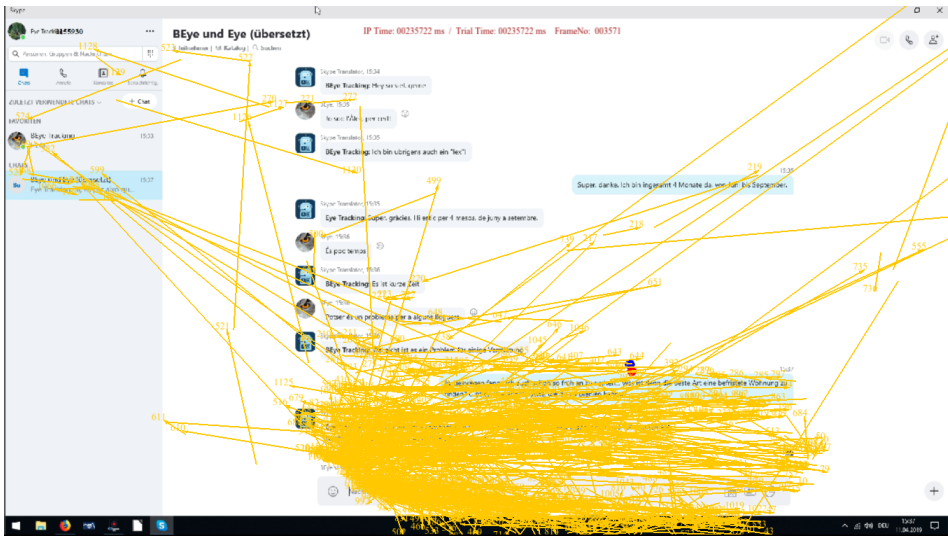


Figure 6.2: Saccadic patterns

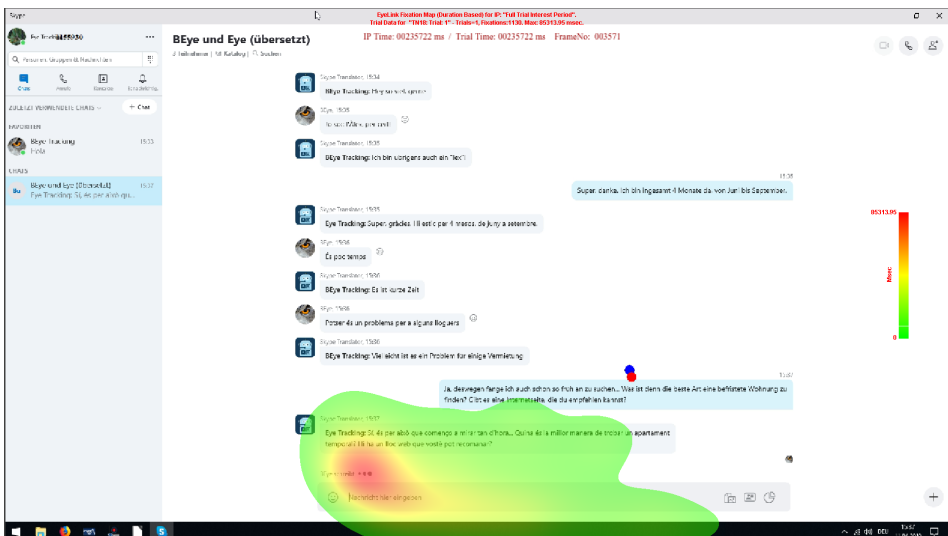


Figure 6.3: Fixation heatmap

distracted or thinking and therefore did not keep their eyes in the covered range of the eye tracking camera.

4 Results

As [Holmqvist 2011](#): 321 points out, saccade amplitude and duration are closely related. Referencing [Carpenter 1988](#), both parameters are correlated linearly, which can be investigated using the correlation test with Spearman’s Rho, as in this present case study both amplitude and duration are not normally distributed. Correlation coefficient Spearman’s Rho is comprised between -1 and 1: -1 indicates a strong negative correlation, 0 means that there is no association between the two variables, 1 indicates a strong positive correlation⁷. The overall data set and the subsets by AOI tag turn out to be positively correlated: Global: ($S = 2.9267e+11$, $p < 0.01$, $\rho = 0.65$), GerO: ($S = 368854171$, $p < 0.01$, $\rho = 0.76$), CatMT: ($S = 3028797579$, $p < 0.01$, $\rho = 0.64$), CatO: ($S = 784453871$, $p < 0.01$, $\rho = 0.7$), GerMT: ($S = 1.0672e+10$, $p < 0.01$, $\rho = 0.69$) and Entry: ($S = 2144558389$, $p < 0.01$, $\rho = 0.46$).

4.1 Saccade amplitude

Table 6.1: Mean and SD of the saccade amplitude per AOI Tag

AOI tag	mean	SD
GerO	1.94	1.51
CatMT	1.82	1.47
CatO	1.90	1.42
GerMT	1.79	1.30
Entry	1.86	1.56
Global	1.84	1.43

Only saccades that start and end in one of the respective AOIs were taken into consideration. Furthermore, amplitude outliers greater than 2.5 times the standard deviation from the mean were excluded ($SD = 1.43$, range = 0.2–8.28). The remaining data set consisted of 1977 saccades with the label of GerO, 3627

⁷see <http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r>, last access: 4 November 2020.

Felix Hoberg

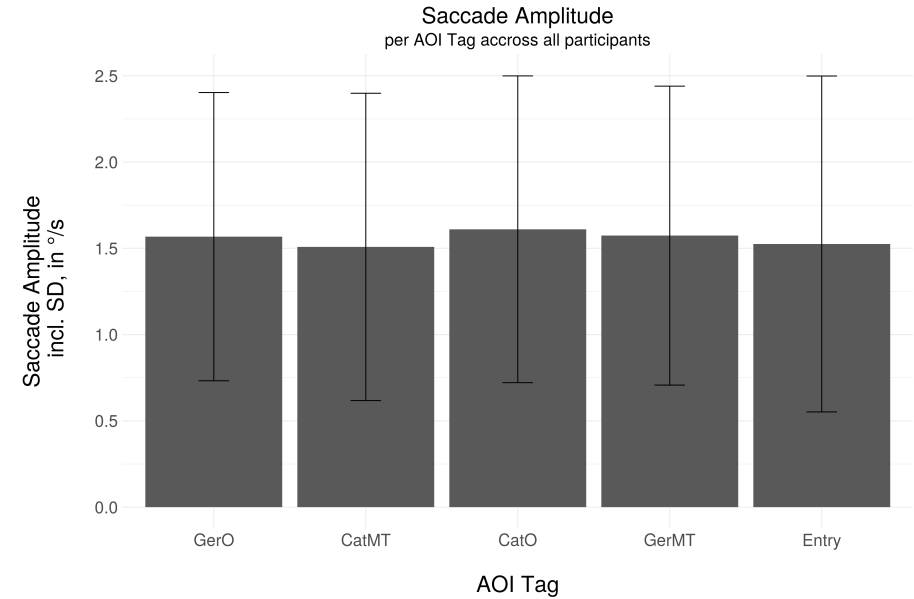


Figure 6.4: Mean saccade amplitude per AOI Tag in °/s

Table 6.2: Results of the Dunn-Test: Pairwise comparison of AOI tags for saccade amplitude

AOI tag pair	z-score	p-value adjusted
CatMT – CatO	-4.946410	(0.0000)*
CatMT – Entry	0.869816	(0.2136)
CatO – Entry	5.615444	(0.0000)*
CatMT – GerMT	-3.244902	(0.0010)*
CatO – GerMT	2.524951	(0.0083)*
Entry – GerMT	-4.090050	(0.0000)*
CatMT – GerO	-4.789648	(0.0000)*
CatO – GerO	-0.151941	(0.4396)
Entry – GerO	-5.427294	(0.0000)*
GerMT – GerO	-2.511196	(0.0075)*

of CatMT, 2453 of CatO, 5852 of GerMT and 3235 of Entry (see Table 6.1). That makes 17144 saccades in total.

Normal distribution of the saccade amplitude data was investigated using the Anderson-Darling-Test that can handle larger data sets than the commonly used Shapiro-Wilk-Test. As the AD-Test indicated a non-normal distribution of the overall data set ($A = 1253.1$, $p < 0.01$) and the subsets by AOI Tag (GerO ($A = 162.84$, $p < 0.01$), CatMT: ($A = 305.34$, $p < 0.01$), CatO ($A = 168.12$, $p < 0.01$), GerMT ($A = 393.26$, $p < 0.01$), Entry: ($A = 226.42$, $p < 0.01$)) and a logarithmic transformation did not change the data set's distribution towards normality, Kruskal-Wallis-Tests were performed to investigate the differences in the saccade amplitudes between participants and between AOI tags.

A Kruskal-Wallis-Test proves that there are significant differences in the saccade amplitudes between some of the participants (Chi-Square(20) = 286.86, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) showed, that 110 of 210 possible pairs (52.38 %) differ significantly. This is no surprise, as the amplitude varies from participant to participant, an observation that has already been stressed by cf. Holmqvist 2011: 312.

A second Kruskal-Wallis-Test shows that there are significant differences of the saccade amplitude between the AOI tags (Chi-Square(4) = 56.56, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) reveals that 8 of 10 possible pairs (80 %) differ significantly (see Table 6.2, Asterisks indicate the significance level: * $\alpha < 0.05$).

4.2 Saccade duration

Only saccades that start and end in one of the respective AOI were taken into consideration. Furthermore, outliers greater than 250ms were excluded (SD = 37.40, range = 14–249). The remaining data set consisted of 2099 saccades with the label of GerO, 3700 of CatMT, 2491 of CatO, 5899 of GerMT and 2883 of Entry (see Table 6.3). That makes 17072 saccades in total.

Normal distribution of the saccade duration data was investigated using the Anderson-Darling-Test that can handle larger data sets than the commonly used Shapiro-Wilk-Test. As the AD-Test indicated a non-normal distribution of the overall data set ($A = 3442.2$, $p < 0.01$) and the subsets by AOI Tag (GerO ($A = 374.99$, $p < 0.01$), - CatMT: ($A = 749.75$, $p < 0.01$) - CatO ($A = 510.65$, $p < 0.01$), - GerMT ($A = 1246.5$, $p < 0.01$), Entry: ($A = 537.5$, $p < 0.01$)) and a logarithmic transformation did not change the data set's distribution towards normality, Kruskal-Wallis-Tests were performed consequently to investigate the differences in the saccade duration between participants and between AOI tags.

Felix Hoberg

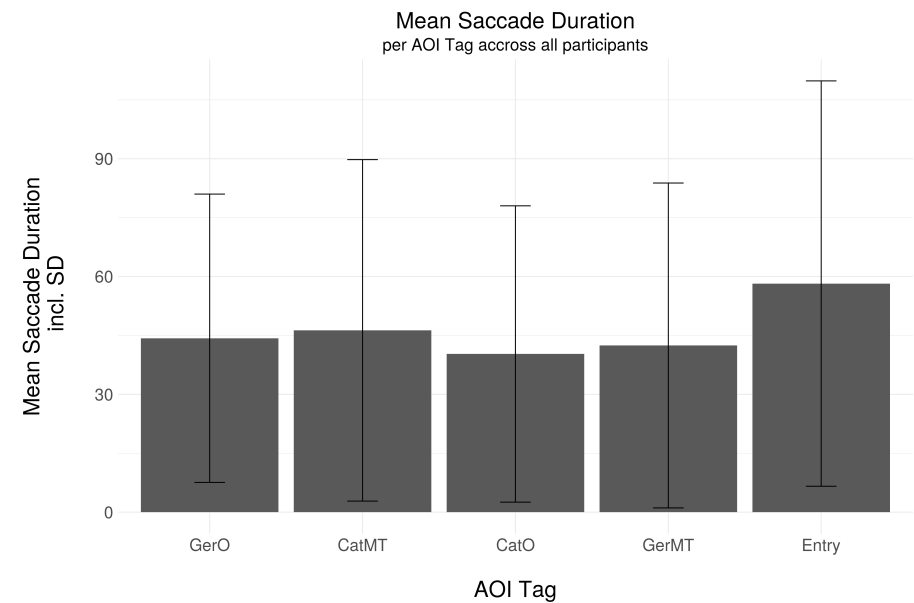


Figure 6.5: Mean saccade duration across all participants in ms

Table 6.3: Mean and SD Saccade duration per AOI Tag

AOI tag	mean	SD
GerO	31.19	31.64
CatMT	34.25	39.59
CatO	33.44	37.71
GerMT	30.60	33.97
Entry	38.12	43.76
Global	33.15	37.41

A Kruskal-Wallis-Test reveals that there are significant differences in the saccade duration between some of the participants (Chi-Square(20) = 184.64, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) showed, that 84 of 210 possible pairs (40.0 %) differ significantly. As in the case of saccade amplitude, these differences are a natural, by-participant phenomenon.

A second Kruskal-Wallis-Test shows that there are significant differences in the saccade duration between the AOI tags (Chi-Square(4) = 49.43, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) indicates, that 6 of 10 possible pairs (60 %) differ significantly (see Table 6.4, Asterisks indicate the significance level: * $\alpha < 0.05$).

Table 6.4: Results of the Dunn-Test: Pairwise comparison of AOI tags for saccade duration

AOI tag pair	z-score	p-value adjusted
CatMT – CatO	-1.543763	(0.0767)
CatMT – Entry	-4.700784	(0.0000)*
CatO – Entry	-2.806306	(0.0050)*
CatMT – GerMT	0.928215	(0.1963)
CatO – GerMT	2.489077	(0.0107)*
Entry – GerMT	5.995560	(0.0000)*
CatMT – GerO	-3.588229	(0.0004)*
CatO – GerO	-1.958909	(0.0358)
Entry – GerO	0.652672	(0.2570)
GerMT – GerO	-4.623838	(0.0000)*

5 Discussion

A look at both the saccade counts for amplitude and duration shows that there are nearly three times as much saccades on the MT output into German as on the German original. In other words, least saccades were made on outgoing messages of the participants compared to incoming texts regardless of language or MT. Taking the count as an indicator for reading depth, the MT into German is read by far more deeply than the German original messages.

As one can see in Tables 6.1, 6.3 and Figures 6.4, 6.5, mean saccade amplitude and duration by AOI tag is comparable to previous studies on reading tasks (cf. Rayner 1998: 373, Gangl et al. 2018, Nikolova et al. 2018), but the amplitude on the

Felix Hoberg

MT output into German is shorter than on the German original and the smallest number in general. The shorter the amplitude, the more closely the participants have read the respective AOI and vice versa. Given that the mean amplitude on both German and Catalan original messages and the entry mask is above average, it can be assumed that these AOI are read less attentively. Then again, the mean saccade duration on both German message types is below average, meaning that shorter saccades are made within these two regions compared to the Catalan messages that are above average. As for shorter saccade amplitude, a smaller duration value represents an increased reading depth and vice versa. This explanation adds up for both types of Catalan messages. Since the participants are not proficient in this language, it seems plausible that they are read only superficially. But when it comes to the lower duration value on German original messages, it is still questionable why these should be closely read.

Taking a closer look at the mean and SD of saccade amplitude and duration, the values reveal high dispersion in the data set. More precisely, both mean and SD values are close to each other. That can be additionally attributed to some reasons less desirable than the above mentioned: First of all, research based on naturalistic studies has to deal with by-participant variance. That is why statistical analyses in this field of research almost always have to deal by nature with the variance that lies within the data set. Second, even the most accurate experimental set-up might miss one crucial variable which therefore deviates the results and has an impact on the interpretation. Third, a high error-rate can also be considered a reason for high dispersion in the data set. Saccades are the fastest movements the human body is capable of. Observing saccades requires therefore precise and accurate equipment. But even then, saccadic movements might go beyond the technical limits of this equipment, making them almost impossible to capture. Lastly, false-positive and false-negative results can also deviate the interpretation. The eye-tracker might detect saccades where none have been or vice versa.

Another strong reason might be the fact that the MT output is just error-prone and therefore requires deeper processing. The observations are supported visually by Fig. 6.2, that depict saccadic eye movements during one session. Most saccades fall into the bottom third, left-aligned area of all machine-translated and Catalan messages. This area covers the entry mask and the latest displayed messages. Only a few saccades are made above the lower third of the screen (on older messages). In addition to that, one single clear gaze path along the session info on the left upper corner of Skype can be identified (see Fig. 6.2). This can be taken as a hint that the participants seldom jump back to older messages but stick more or less to the most recent output of the text chat utterances on screen.

The pairwise calculated tests show that the German original messages cause significantly increased saccade amplitudes compared to every other AOI tag except the Catalan original (see Table 6.2). In contrast, significantly increased amplitudes in comparison to every other AOI tag except the German original can be observed on MT messages into German. The fact that the tests showed no significant results for the pairwise comparison of German original and Catalan original may lead to the conclusion that the participants are somehow noticing the incoming message in a language they are not proficient in. Given that German and Catalan share the same character system, participants may be switching between original and machine-translated utterance in search for words they can recognize. Those can be names, numbers, words that share the same root in both languages or even words that can be deduced from another (roman) language the participants are proficient in. As both pairs, German original vs. MT into Catalan and Catalan original vs. MT into German, turn out to differ significantly, this might be taken as a first indication for this hypothesis but has to be explored in upcoming studies.

These observations may also be seen as indicators for the participants reading the German MT output more carefully due to typical MT errors in terms of syntax, semantics or orthography, which then results in shorter saccade amplitudes. The longer saccade amplitude on the German and Catalan original messages leads consecutively to the opposite assumption: the participants' reading behavior is less deep since they are already familiar to the German original. When it comes to the reason why longer saccades are made when reading the Catalan original, the participants might spend less care on reading a language they are not proficient in. One definitely would have to link the saccadic observations to their respective fixations to check for complete plausibility of this hypothesis.

When it comes to saccade duration, German original messages turn out to be significantly different compared to MT into German and into Catalan (see Table 6.4). Coming back to the observations of the relation between cognitive load and saccadic eye movements described by cf. Holmqvist 2011: 313f. reading (one's own) German original messages requires less cognitive capacity than processing incoming MT output into German, which in fact is new information to the participants and therefore definitely takes longer to process. In contrast, there is no significant difference between German and Catalan original messages. Given that the participants have written the German original messages themselves and are already familiar to the information, the non-existence of any significant result might indicate that reading the Catalan original is as (little) cognitively demanding as is reading the German original. In a further step, this might be taken as a hint that the Catalan original is only read superficially. The same goes for

Felix Hoberg

the entry mask, where participants write, revise and send their messages. On the contrary, MT into German differs only significantly from the Catalan original and the Entry mask.

6 Conclusion

The present study tackled the question if saccadic eye movement patterns differ according to the type of text chat messages in reading tasks. Indeed, significant variance can be identified not only between participants – as is typical for saccade amplitude and duration – but also between incoming and outgoing or machine-translated and original messages. But it has to be stressed that the results represent nothing more than a first exploratory overview. It will be necessary to take a closer look at the interplay of saccade amplitude or duration and AOI size, blink patterns and the overall scan path. Further on, it has to be assumed that the limits of investigating saccadic eye movements in the context of technologies like the Skype Translator exist on word level, as it is only marginally feasible to annotate single words with dynamic AOI on text chat message level. The text box and font size is just too small, so that the impact of eye tracking indicators on (for example) orthographic information can only be deduced globally. Certainly, the evaluation of other commonly operationalised indicators in reading studies (fixation duration, dwell time, regressions, fixation count etc.) has then to be considered and to be linked to the findings on saccadic eye movements, too. It becomes clear that the exclusive investigation of saccadic eye movements in CMC studies therefore seems not to be enough for extracting valid results. Furthermore, the impression on the subjective quality of the communication as stated in the questionnaires at the beginning and end of each session has to be situated along the analysis.

The data set presented here is only half of the data that were collected during the overall project. The other half consists of eye-tracking data of monolingual text chats of seven German native participants via Skype. Both sets have to be linked to draw conclusions on the differences between monolingual and machine-translated text chat communication.

Nevertheless, it seems a quite promising endeavour to observe how participants react to real time machine translation in text chats when they are not proficient in one of the involved languages. It helps to assess the requirements to better understand this type of communication technologies. In a more global context, it eventually contributes to understand how all this shapes the way of communicating on the internet.

In the end, all similar endeavours have to be aware of the fast developing technology they are based on. The total count of languages featured by the Skype Translator has steadily increased since this project started. What is more, the layout of the Skype Translator changed as well, now only displaying messages in the operating system's language, leaving aside the two column comparable design this present study investigates. This fast changing environment can be taken as another argument to continuously investigate the human-machine-interaction in everyday life. This kind of technology has already penetrated every single aspect of our lives which is why it would be highly negligent to not evaluate the human behaviour when dealing with it.

Abbreviations

HTWK – Leipzig University of Applied Sciences

Labels of the areas of interest

GerO – German Original

GerMT – Machine Translation into German

CatO – Catalan Original

CatMT – Machine Translation into Catalan

Acknowledgements

I am grateful to our institute's student assistant Tim Feldmüller who took care of preparing most collected research data.

References

- Beißwenger, Michael. 2017. *Empirische Erforschung internetbasierter Kommunikation*. Berlin: De Gruyter.
- Beißwenger, Michael. 2007. *Sprachhandlungskoordination in der Chat-Kommunikation* (Linguistik, Impulse & Tendenzen 26). Berlin: W. de Gruyter.
- Bowker, Lynne & Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. OCLC: on1075580986. Bingley, UK: Emerald Publishing.

Felix Hoberg

- Carpenter, R. H. S. 1988. *Movements of the eyes*. 2nd rev. & enlarged ed. London, England: Pion Limited.
- Doherty, Stephen & Sharon O'Brien. 2014. Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking. *International Journal of Human-Computer Interaction* 30(1). 40–51. DOI: [10.1080/10447318.2013.802199](https://doi.org/10.1080/10447318.2013.802199).
- Doherty, Stephen, Sharon O'Brien & Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 24(1). 1–13. DOI: [10.1007/s10590-010-9070-9](https://doi.org/10.1007/s10590-010-9070-9). <http://link.springer.com/10.1007/s10590-010-9070-9> (22 July, 2020).
- Duchowski, Andrew T. 2017. *Eye tracking methodology*. Cham: Springer International Publishing. DOI: [10.1007/978-3-319-57883-5](https://doi.org/10.1007/978-3-319-57883-5). <http://link.springer.com/10.1007/978-3-319-57883-5> (17 October, 2019).
- Fišer, Darja & Michael Beißwenger (eds.). 2017. *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*. 1st edn. (Book series translation studies and applied Linguistics). Ljubljana: Ljubljana University Press. <https://e-knjige.ff.uni-lj.si/> (27 October, 2017).
- Gangl, Melanie, Kristina Moll, Manon W. Jones, Chiara Banfi, Gerd Schulte-Körne & Karin Landerl. 2018. Lexical reading in dysfluent readers of German. *Scientific Studies of Reading* 22(1). 24–40. DOI: [10.1080/10888438.2017.1339709](https://doi.org/10.1080/10888438.2017.1339709).
- Han, Lifeng. 2018. *Machine Translation Evaluation Resources and Methods: A Survey*. arXiv: 1605.04515. <http://arxiv.org/abs/1605.04515> (2 September, 2020).
- Holmqvist, Kenneth (ed.). 2011. *Eye tracking: A comprehensive guide to methods and measures*. OCLC: ocn741340045. Oxford ; New York: Oxford University Press.
- Jacobson, J. Zachary & P. C. Dodwell. 1979. Saccadic eye movements during reading. *Brain and Language* 8(3). 303–314. DOI: [10.1016/0093-934X\(79\)90058-0](https://doi.org/10.1016/0093-934X(79)90058-0).
- Leube, Alexander, Katharina Rifai & Siegfried Wahl. 2017. Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal of Eye Movement Research* 10(3). 1–11.
- Nikolova, Mirela, Stephanie Jainta, Hazel I. Blythe & Simon P. Liversedge. 2018. Binocular advantages for parafoveal processing in reading. *Vision Research* 145. 56–63. DOI: [10.1016/j.visres.2018.02.005](https://doi.org/10.1016/j.visres.2018.02.005). <https://linkinghub.elsevier.com/retrieve/pii/S0042698918300233> (30 April, 2020).
- O'Brien, Sharon. 2009. Eye tracking in translation process research: Methodological challenges and solutions. *Methodology, technology and innovation in translation process research* 38. 251–266.

- R Development Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Ramlow, Markus. 2009. *Die maschinelle Simulierbarkeit des Humanübersetzens: Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik* (TransÜD 27). OCLC: 553597343. Berlin: Frank & Timme.
- Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124(3). 372–422. DOI: [10.1037/0033-2909.124.3.372](https://doi.org/10.1037/0033-2909.124.3.372).
- Rayner, Keith. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62(8). The 35th Sir Frederick Bartlett Lecture: 1457–1506. DOI: [10.1080/17470210902816461](https://doi.org/10.1080/17470210902816461). (19 August, 2020).
- Schaeffer, Moritz, Kevin B. Paterson, Victoria A. McGowan, Sarah J. White & Kirsten Malmkjær. 2017. Reading for translation. In Arnt Lykke Jakobsen & Bartolomé Mesa-Lao (eds.), *Translation in transition: Between cognition, computing and technology*, vol. 133 (Benjamins Translation Library), 17–53. Amsterdam: John Benjamins Publishing Company. DOI: [10.1075/btl.133](https://doi.org/10.1075/btl.133). <http://www.jbe-platform.com/content/books/9789027265371> (14 February, 2019).
- SR Research Ltd. 2019. *EyeLink Data Viewer*. Mississauga, Ontario, Canada.
- Storrer, Angelika. 2001. Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In Michael Beißwenger (ed.), *Chat-kommunikation. Sprache, Interaktion, sozialität & identität in synchroner computervermittelter kommunikation. Perspektiven auf ein interdisziplinäres forschungsfeld*. 3–24. Stuttgart: ibidem.
- Vardaro, Jennifer, Moritz Schaeffer & Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* 6(3). 41. DOI: [10.3390/informatics6030041](https://doi.org/10.3390/informatics6030041). <https://www.mdpi.com/2227-9709/6/3/41> (20 March, 2020).
- Verheijen, Lieke. 2017. WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication. In Darja Fišer & Michael Beißwenger (eds.), *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*, 1st edn. (Book series translation studies and applied Linguistics), 72–101. Ljubljana: Ljubljana University Press. <https://e-knjige.ff.uni-lj.si/> (27 October, 2017).

Chapter 7

What can Euclidean distance do for translation evaluations?

Éric André Poirier

Université du Québec à Trois-Rivières

We describe an empirical method to screen informational translation shifts in parallel segment pairs extracted from a bilingual or multilingual translation corpus using two linguistic features that are independent of the languages matched by the translation. The method applies to most known languages and in one or the other of the two translation directions (direct or inverse). The features measured for each segment in source and target languages are character length and lexical word count (or information volume). Information volume is compiled through an algorithm coded in Python using spaCy v2.1.3 core linguistic models. The values of source and target segment features and the translation precision ratio of each segment pairs are averaged over the text to which they belong and all segment values are standardized in relation to their textual average. The deviation between standardized values for each segment in a pair, as measured by the weighted Euclidean distance, allows for the screening and identification of target segments that are atypical or heteromorphic in comparison with their source segment. Our hypothesis is that those heteromorphic segment pairs, as opposed to isomorphic ones, are more likely to contain informational translation shifts. The objective and reproducible method described herein allows for semi-automatic identification of problematic translations and uncovering of textual and linguistic facts revealing translation processes, contingencies, and determinism.

1 Introduction

We describe below the theoretical framework and the methodological steps of the method that we have applied in a systematic and exploratory way to parallel bilingual corpora in different languages and in different translation directions with



Éric André Poirier

English, French, and Spanish. The method may be applied manually on small texts and for pedagogical purposes in the analysis, evaluation, and comparison of translations and translation processes, or it can be implemented for manual identification of informational translation shifts in automatically screened segment pairs in large corpora. Automatic POS tagging of all segment pairs was done with spaCy v2.1.3 (a commercial grade natural language processing environment, ([Explosion_AI 2016–2020](#))) core linguistic models in an algorithm coded in Python version 3.7.3 with language models `en_core_web_sm` (version 2.1.0) for English and `es_core_news_sm` for Spanish (version 2.1.0). POS tagging is required to calculate the information volume of each segment. Languages that are covered with our method are determined by the availability of a specific linguistic module in the spaCy environment designed for Python programming.

For illustrative purposes, we present the results obtained with the method applied to the United States President Barack Obama’s speech to the Cubans on March 22, 2016, for which an official translation is provided in Spanish. The bitext used for the analysis was compiled with the original English version¹ and its official Spanish translation², both of which are posted on the obamawhitehouse.archive.gov Web site, which includes official speeches delivered by President Obama. The speech has 2,420 words in English, 2,468 in Spanish, and the raw bitext was segmented in 255 segment pairs, as described below in §3.

2 Theoretical framework

Before explaining our method, we describe the typology of informational translation shifts for the manual annotation and analysis that is required to measure the efficiency of the method. This framework also describes key concepts in the evaluation of the efficiency and utility of the method we present regarding the screening of segment pairs which are most likely to contain translation shifts.

2.1 Free and fixed translation shifts

First, let us define what we mean by translation shift and propose a typology of the types of informational shifts found in the segment pairs of parallel translation corpora. The term shift is used in its broad sense to mean “a change in position

¹<https://obamawhitehouse.archives.gov/the-press-office/2016/03/22/remarks-president-obama-people-cuba>

²<https://obamawhitehouse.archives.gov/the-press-office/2016/03/22/discurso-del-presidente-obama-al-pueblo-cubano>

or direction”³. Translation shifts generally refer to specific changes attributable to translation, explained thus: “The transformation which is occasioned by the translation process can be specified in terms of changes with respect to the source texts, changes which are termed ‘shifts.’” (Bakker et al. 2011: 269)

In this sense, translation shifts do not include systematic or systemic differences between languages. Although no empirical criteria have been provided to differentiate between translation shifts and differences between languages, it is generally accepted that these two transformations in the translation process must be distinguished. To account for these two very different types of shifts, we have adopted the terminology of Wecksteen-Quinio et al. (2015). The authors distinguished fixed shifts that are attributable to differences between languages from free shifts that are attributable to the translation operation itself and result from a choice freely exercised by the translator, from bias on the part of that person, or simply from translation errors. While fixed shifts are mandatory, free shifts are by definition free or the result of a deliberate choice. Strictly speaking, they are members of a group of at least two expressions that adequately translate the expression or the same elements of the source segment. In theory, fixed shifts describe conventional translation processes, while free shifts describe creative, original, or to some extent novel translation processes. Instead of relying exclusively on our own judgment on the acceptability of Spanish translations, we designed a process that supports the empirical definition of free shifts based on the *tertium comparationis* provided by machine translation. For a source expression, if a literal translation in the target text co-occurs with an acceptable literal translation of the same expression in DeepL⁴, the shift in the official translation is fixed. When the target text contains a non-literal translation, if the same source expression is translated literally in DeepL, the shift is considered free. A good example among others (see §4) is the translation of the segment number 187 “that is a measure of our progress as a democracy” that was translated as “Esa es la medida de nuestro progreso”, which is not literal and which co-occurs with a literal translation in DeepL “que es una medida de nuestro progreso como democracia”. The comparison with DeepL highlights the omission of the content word *democracia* in the official translation. Translation shifts screened with our method are limited to informational translation shifts and can either result in the addition of one or more content words or the omission of one or more content words (see §2.3 below).

³Source: Online Cambridge Dictionary at <https://dictionary.cambridge.org/>.

⁴<https://www.deepl.com/translator>

Éric André Poirier

2.2 Informational translation shifts

The term “informational shift” refers to a particular type of translation shift. In the identification of all translation shifts (semantic, lexical, syntactic, stylistic, terminological, socio-linguistic, etc.) that are required for the knowledge and maintenance of a coherent set of translation processes (which constitute the basic elements of translation learning and teaching), informational translation shifts represent a critical group of translation shifts. In fact, they are requisite to the proper identification and definition of all other types of shifts since informational shifts affect the information content of the messages to be translated, which is required to be invariant in the translation of pragmatic texts, and on which the analysis and evaluation of other translation shifts depend.

We hypothesize that informational translation shifts are most likely present when a comparison of source and target segments show an important discrepancy or “distance” in two correlative linguistic features: the string length in characters and the lexical word count. Lexical words are numerous; they carry a lexicalized or stable meaning and form an open class of elements. This is in contrast with grammatical words that are few, do not carry a lexicalized meaning, and form a closed set of elements. By counting lexical words in source and target segments (in two different languages), the method we describe here allows for the quantifying of the translation precision in terms of information volume. This measure is defined in the next section.

2.3 Positive and negative information shifts

As discussed in §2.1, information shifts may result in the addition or the omission of information. The volume of information as measured by the lexical word count is an approximation of the quantity of basic (stable) information present in source and target segments. The translation precision ratio (TPR) is calculated by dividing the information volume of the source segment by the information volume of the target segment and may be “positive”, “negative” or “neutral”. TPR is a numeric measure of the discrepancy of information volume between target and source segments. When segment pairs contain an equal volume of information in both the source and target segments, the TPR between the two segments is “neutral” with a value of 1.0 and those segment pairs are isomorphic. When segment pairs contain at least one negative information shift, that is, the omission of information in the target segment, the information volume of the target segment is smaller than the information volume of the source segment. The TPR between the two segments is “negative” with a value lower than 1.0 and those

translation segment pairs are negative heteromorphic. When segment pairs contain at least one positive information shift, that is, the addition of information in the target segment, the information volume of the target segment is greater than the information volume of the source segment. The TPR between the two segments is “positive” with a value higher than 1.0 and those translation segment pairs are positive heteromorphic.

Since information shifts mostly occur within the segment level, numerous combinations of positive and negative shifts may exist in isomorphic, negative heteromorphic, and positive heteromorphic segment pairs. For example, an isomorphic segment pair may have one positive shift and one negative shift, each canceling out the value of the other and a heteromorphic segment pair may have multiple negative shifts and positive shifts. In this case, there may be a single positive or negative shift, as the case may be, or there may be multiple negative or positive shifts that combine within a segment pair that is either negative or positive as a whole.

2.4 Antinomic shifts

Antinomic shifts are those whose positive or negative nature is opposite to that of the whole segment to which they belong. For example, a positive heteromorphic segment pair may contain two positive shifts of one lexical word each or a single positive shift of two lexical words, in combination with a negative shift of one lexical word that does not contribute to the positive orientation of the segment pair. The positive or negative orientation of antinomic shifts is opposite to that of the orientation of all the combined shifts of a pair of segments. In neutral isomorphic segments (having a TPR of 1.0), any pair of information shifts that may occur (one positive and one negative) cancel each other out and are therefore both antinomic. For this reason, it should not be concluded that there is no informational translation shift in isomorphic segment pairs. However, as demonstrated in §5, we hypothesize that there are fewer of them in isomorphic segment pairs than in the positive or negative heteromorphic segment pairs.

2.5 False shifts and undetected shifts

Because of the shortcomings of the spaCy v2.1.3 core linguistic models and the erroneous results they sometimes produce as regards POS tagging, we created two other categories of information shifts that could only be detected through manual and meticulous analysis of the segment pairs screened by the weighted Euclidean distance (see §3.3). One difficulty in POS tagging is that most tokens

Éric André Poirier

belong to several lexical or grammatical word classes. Some parts-of-speech are also equivocal regarding their belonging to a lexical or a grammatical class. This is the case, for example, of verbal auxiliaries in English, Spanish or French, or for some particles in phrasal verbs in English – are they adverbs or prepositions? Most POS tagging algorithms struggle to provide a proper analysis of all source and target segment tokens (despite, and with the support of, language-specific rules), and for specific tokens or POS may present original aberrations that need to be corrected. For some older releases of spaCy’s POS tagger, Giesbrecht & Evert (2009) report a success rate of less than 93%, and this rate varies (downward) depending on the type of text analyzed. When manual analysis reveals errors or anomalies in POS tagging of tokens, the involved information shifts have been classified as false shifts (in the way that they are false positives) that owe their existence only to POS tagging errors. Another development that would enhance the efficiency of the empirical method described here is the improvement of POS tagging such that every token and every compound or group of tokens would be properly tagged as a lexical or a grammatical item. Like we explained in a previous paper (Poirier 2017: 8), converting even a 97% POS tagging accuracies at the segment level makes it less impressive since it can be reasonably argued that most segments (and sentences) generally have at least 10 words or more. For ten segments of 10 words, an accuracy of 97% would imply that as much as three segments out of ten (that is 30% of segments) would contain a POS tagging inaccuracy provided the three words inaccurately tagged out of 100 are distributed in three different segments. Furthermore, considering that parallel corpora involve two different languages (and two different POS tagging source of errors), this number may skyrocket to 60% of all 10 segment pairs if the two language-specific groups of 30% erroneous segments are each matched to a properly analyzed source or target segments.

When the POS tagging modules produce an erroneous analysis that results in the inexistence of an information shift (and which produces a false negative), these information shifts that go unnoticed have been classified as undetected shifts, i.e., shifts that were not detected because of wrong POS tagging. For example, an undetected shift was found in segment number 63 of our corpus (see §4.1) which contained the expression “a multi-party democracy” matched with the Spanish translation “*una democracia de múltiples partidos*”. The source segment was wrongly analyzed as having four lexical words by the English language model of spaCy⁵, giving rise to a false shift and a fourth lexical word). In

⁵In this case, this was due to the the hyphen being wrongly analyzed as an adjective, but this was not the only wrongful POS tagging issue with the hyphen since in parallel segment no 239 (see §4.3) it was analyzed as a proper noun.

this case, the target segment was analyzed correctly with three lexical words. What the module analysis made as a negative heteromorphic segment pair turns out to be a positive (antinomic) heteromorphic segment pair because *multi-party* should be analyzed as a unitary lexical word (compound). Thus, in this segment, our manual analysis found an undetected information shift that both linguistic language models have been unable to bring to light.

3 Corpus data processing methodology

The file used as input is a bitext file in HTML format provided free online by YouAlign⁶ (maximum file size for each file is limited to 1MB). The speech file size of our corpus did not exceed this limit but one could use a comma-separated value file format or other proprietary bitext creation software such as Logiterm Pro v5.8.2 for larger files and corpora. It has been verified that the alignment of all segments of the bitext is adequate and that each source segment matches its translation with one or more target segments, if applicable. Manual processing was necessary at this step on the source and target language plain text of the speech. In our corpus, annotations such as “Applause” and “Laughter” that describe the audience’s reaction to the speaker’s words have not been included and translated in the target text. It seems fair and reasonable to delete those items that were not genuinely communicated by the speaker and not translated in Spanish because they were provided by the context. For reasons that are difficult to explain (and which probably have to do with character encoding or the core and basic language models that were used even if some testings with more complete language models that were available at the time did not demonstrate noticeable improvements), some abbreviated forms with apostrophes in English needed to be modified as the last recourse solution (such as *that’s* = *that is*) because the apostrophes were recognized as lexical words, which is not accurate. In the English source text, the last greeting from the speaker is “*muchas gracias*” in Spanish which obviously does not need to be translated. This last single segment needed to be removed from the bitext since it cannot form a pair of parallel bilingual segments. For the target speech in Spanish, the segmentation results with the dash and colon had to be corrected to match the segmentation results of their corresponding punctuation marks in English.

Once these modifications were made to our corpus, a module written in Python analyses all the pairs of segments of the corpus one by one. In this analysis, two specific linguistic modules are called sequentially for the source segment and the

⁶<https://youalign.com>

Éric André Poirier

target segment to count their lexical words and measure the total information volume of each segment as well as the TPR of each segment pair. These values are appended to a variable and then exported to a CSV file which can be read in a spreadsheet. The character length of each segment could be quantified afterward in the spreadsheet with the help of a function such as LEN (cell) function in Excel. We also calculated for each source and target language texts the average value of information volume and character’s length by segment for the whole corpus.

Our English-Spanish parallel corpus of Obama’s speech to the Cubans contains 255 segments of 9.49 lexical words and 89.68 characters on average in English and 9.68 lexical words and 96.15 characters on average in its translated version in Spanish. These averages were calculated with the values of both linguistic features for the whole corpus as described in the Table 7.1.

Table 7.1: Total values of linguistic features in source and target languages

	Characters	Lexical words
English (source)	22,869	2,420
Spanish (target)	24,517	2,468
Difference	+ 7.21%	+ 1.98%

In recent years, we applied different versions of our methodology and translation precision algorithm to some corpora (see [Poirier \(2019\)](#) for an example of English-French analysis and earlier methodology). We have found that three linguistic features may be measured for each parallel segment pairs in bitexts, which are character’s length, total word count (or token count), and lexical word count. We tested the correlation of each feature in different corpora that were analyzed with our methodology. In order to measure the correlation of these features we simply applied the Pearson correlation coefficient between two variables (values of linguistic features in source and target segments) as defined with the following formula, where *cov* is the covariance, ρ_X and ρ_Y are the standard deviations of X and Y, respectively:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}$$

The Table 7.2 below presents the correlation which was calculated with different political speeches in English translated in Spanish, such as Abraham Lincoln’s Gettysburg Address (1863), Inaugural Address of John F. Kennedy (1961),

Martin Luther King’s *I have a dream* (1963), Obama’s speech to the Cubans (2016), Donald Trump’s State of the Union (2018) and Oval Office Address on Border Wall (2019).

Table 7.2: Correlation of three linguistic features in English-Spanish translations

Speech	Word length (text)	Character’s length	Total word count	Lexical word count
Gettysburg	268	0.9725	0.963	0.9691
KennedyInauguralAddr	1393	0.9888	0.984	0.9736
DreamMLKing	1673	0.9766	0.9772	0.9684
ObamaCuba	4161	0.9733	0.9663	0.9583
TrumpStateUnion2018	5188	0.9649	0.9408	0.9573
Trump_BorderWall	1119	0.9513	0.9205	0.9501
Averages		0.9712	0.9586	0.9628

Table 7.2 shows that on average, the characters’ length has the strongest correlation (0.9712)⁷, followed by the lexical word count (0.9628) and by the total word count (0.9586). Because of this high correlation of these two features between the source and target segments, the significance of the lexical word and character differences between the source and target segments is difficult to establish when the length of segment pairs may vary widely. For example, the absence of a lexical word in a target segment that is associated with a source segment of 30 lexical words is not as significant as the absence of a lexical word in a target segment that is associated with a source segment of three lexical words.

3.1 Standardized values of segment pairs

To account for the relative length of each string in segment pairs, and to make each segment pair comparable in terms of their selected features, we standardized the value of the two features of each segment by relating them to their average value for the whole source or target segments in the parallel corpus. To this end, a rule of three was used to determine the standardized values of information

⁷These data support previous works in machine translation, such as the seminal paper of Gale & Church (1993: 89), who found that there exist very high correlations between the length of a paragraph in characters and the length of its translation.

Éric André Poirier

volume and character's length for each segment in pairs. In the context of Barack Obama's English-Spanish corpus, let's take for example a source and target segments having respectively 3 and 4 lexical words and 25 and 31 characters. If we relate these numbers to their average value for the corpus (9.49 lexical words and 89.68 characters for the source segments, and 9.68 lexical words and 96.15 characters for the target segment), we get standardized values of 4.07 ($3 \times [9.49 + 9.68] / 7$) lexical words, for the source segment, and 5.53 ($4 \times [9.68 + 9.49] / 7$) lexical words, for the target segment. The same formula is used for the character's length standardized values. The standardized value of the two features measured for each pair of segments is crucial since they will make it possible to detect target segment pairs that are atypical (or unusually distant from their source segment), as measured by the weighted Euclidean distance.

3.2 Precision deviation factor

To characterize the positive or negative value of the information volume of the whole target segment compared to the whole source segment, we subtracted its TPR from the average value of this ratio for the whole text, a value which is normally close to 1.00 (target segment normally contains the same volume of information of the source segment). In Barack Obama's Speech English-Spanish corpus, this figure was 1.02. Any segment pair having a TPR lower than 1.02 would, therefore, have a negative value, and, conversely, any segment pair having a TPR higher than 1.02 would have a positive value. The precision deviation factor (PDF) used in the calculation of the Euclidean distance is simply a multiple (10 times) of this value (positive or negative⁸). Just like the positive and negative values of information shifts were an indication of a potentially wrong additional or missing information in the translation, the negative or positive value of the Euclidean distance would point to a potentially wrong additional or missing information in the target segment.

3.3 The weighted Euclidean Distance for screening segment pairs

The Euclidean distance is calculated using the standardized lexical word count and the standardized string length in characters that were calculated for the source and target segment of each parallel pair in the corpus. The exact formula of the Euclidean distance ($d(p,q)$) that we used is defined as follows:

⁸A value of zero is theoretically possible with a TPR of 0.99 but this value did not occur in our corpus. Some adjustments might be required in the following calculations to take this value into account.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Simply put, the Euclidean distance is measured by the square root of the sum of the squared deviations of the two features (information volume and string length in characters) measured and standardized for each source and target segment of all parallel pairs in our corpus. Since translation is an operation that takes into account meaning, we gave more weight to the difference in information volume than to the difference in the number of characters in the calculation of the weighted Euclidean distance. Multiplying the Euclidean distance by the positive or negative precision deviation factor gives more weight to the information volume and results in a positive or negative value of the distance between the source and target standardized segments of each parallel pair. When the value is negative, the target segment contains fewer lexical words or characters than the source segment and is likely to contain at least one or more negative shifts. Similarly, when the value is positive, it means that the target segment contains more lexical words or characters than the source segment and is therefore likely to contain at least one or more positive shifts.

This method has made it possible to calculate the weighted Euclidean distance separating each pair of segments. Of the 255 pairs of segments in the English-Spanish corpus of Barack Obama's speech in Cuba, the weighted value (by the precision deviation factor) of the Euclidean distance is between -193.81 and 313.26. Segment pairs with extreme negative or positive values of weighted Euclidean distance are highly heteromorphic and their target segment is very likely to contain informational translation shifts. The two most heteromorph segments and their particular calculations are described in the next table. The segments are preceded by their sequential number in the English-Spanish corpus. Proper and improper content words (leading to false shifts) identified with the spaCy v2.1.3 language models are underlined. The volume of information and the number of characters in each source and target segment is in square brackets. In the calculation of the precision deviation factor, TPR is averaged at 1.02 and the weighting of this difference has been multiplied by a constant of 10. The precision deviation factor for segment #12 is therefore $(0.333 - 1.02) * 10 = -6.87$. For calculating standardized values, the average value of source segments features are 9.49 lexical words and 89.68 characters, and for the target segment features these figures are 9.68 lexical words and 96.15 characters. Some slight differences may occur due to the rounding of the decimals and their precision. Table 7.3 presents the detailed

Éric André Poirier

calculations for the most negatively heteromorphic segment pair number 12 and the most positively heteromorphic segment pair number 144 in the corpus. The table shows the translation precision ratio (TPR), the precision deviation factor (PDF), the standardised information volume (SID) and string length (SSL) of the segment, the Euclidean distance (d) and the weighted Euclidean distance (wd).

Table 7.3: Most negative and positive heteromorphic segment pairs and their linguistic feature values (lexical words in bold)

Segment	12. Thank you very much. = Muchas gracias. [3, 20 = 1, 15]	144. Not everybody agrees with me on this. = No todo el mundo está de acuerdo con- migo sobre esto. [3, 37 = 6, 52]
TPR	0.33	2.00
PDF	-6.87	9.80
SID	14.38 and 4.79	6.39 and 12.78
SSL	106.19 and 79.64	77.26 and 108.57
d	28.22	31.96
wd	-193.81	313.26

Manual analysis of the shifts in the segments number 12 shows that the great negative Euclidean distance is due to a false shift that is attributable to the classification of the Spanish adverb *muchas* as a determinant (a grammatical word), compared to the English adverb *much*, which is classified as an adverb and therefore as a lexical word. In the same segment, there is a second fixed shift with the use of the adverb *very* in English which has no corresponding Spanish equivalent (probably because *muchas* is already used as an adverb). The two shifts taken together explain the shift in the information volume of 2 found between the two segments. The number of characters is in the same negative direction of the information volume shift and reveals that the target segment of the pair is shorter than the source segment.

In segment pair number 144, there are three positive shifts of one lexical word for each in favor of the target segment. First, there is a fixed shift with the correspondence of the verb *agree* (one lexical word) and the Spanish phrase *está de acuerdo* (two lexical words). Then there are two false shifts with a positive value due to the wrong POS tagging of *conmigo* as an auxiliary (lexical word) rather than a preposition or prepositional phrase, and another wrong POS tagging of *sobre* as a verb (lexical word) rather than as a preposition. These two false shifts are due to POS tagging errors in the spaCy v2.1.3 language models. The analysis

of *todo* seems to have been well done by the spaCy v2.1.3 language models since it is categorized as a grammatical word even though the type of grammatical word seems to be wrong, i.e. a determiner rather than an indefinite pronoun.

These two examples show the importance of POS tagging in the analysis of translations and the calculation of the information volume. It is to be hoped that significant progress will be made in this area. Despite scientific articles that regularly report success rates of 95% to 98% in POS tagging, it seems that these data are inaccurate, at least with spaCy v2.1.3 POS tagging modules.

4 Results

After having applied the corpus data processing methodology described above, we wanted to validate its efficiency regarding the screening of negative and positive informative translation shifts. For this purpose, we manually analyzed three samples (A, B, and C) of twenty pairs of segments screened automatically with the numeric value of the weighted Euclidean distance. Segment pairs in two of those samples (A and B) were selected for their highest (positive) and lowest (negative) weighted Euclidean distance between the source and target segments (and for being representative of the most negative and positive heteromorphic segment pairs within the analyzed English-Spanish corpus). In a third sample (C), twenty other segment pairs were selected for their very neutral (near zero) weighted Euclidean distance between the source and target segments (and for being representative of the most isomorphic segment pairs within the analyzed English-Spanish corpus). These three groups of 21 segment pairs were analyzed manually as regards the presence or the absence of information shifts described in §2. Detailed data on the manual analysis of each of the three samples of twenty segments is described in the next subsections below.

In each three annotation tables on the left column, lexical words in segment pairs are underlined to inform the reader of the results of the automatic POS tagging process. For difficult or ambiguous word-forms, part-of-speech are indicated in uppercase when needed. Tag set that is used is the same as spaCy v2.1.3 POS tag symbols that are called Universal POS tags and that comes from the Universal Dependencies Scheme⁹. The empty symbol (\emptyset) is used to describe an item having no semantic match in the target segment. The asterisk symbol (*) is used to describe erroneous tagging which explains the false shift POS annotation. Information shifts are described in their order of appearance in the target segment.

⁹<https://universaldependencies.org/u/pos/>

Éric André Poirier

4.1 Sample A annotations – most negative heteromorphic pairs

This section contains the manual analysis and annotations of sample A segments for the classification of information shifts observed in the most negative heteromorphic segment pairs. Translation pairs have been sorted from the most negative weighted Euclidean distance (-67.52) to the less negative weighted Euclidean distance (-15.67).

(12) **Thank-v** you **very-ADV** **much-ADV**. 3, 20 → **Muchas-^{*}DET/ADJ** **gracias**. 1, 15
-193.81 [-2 lexical words] (2 shifts)

1. False shift POS (-1): much-ADV → muchas-^{*}DET/ADJ
2. Fixed shift (-1): very-ADV → ∅ [confirmed with DeepL: *Muchas gracias*.]

(258) And it **will not be easy**, and **there-^{*}ADV/PRON** **will be setbacks**. 8, 52 → Y **no será fácil**, y **habrá reveses**. 5, 33
-165,00 [-3 lexical words] (3 shifts)

1. Fixed shift (-1): will-v → ∅ [future tense]
2. False shift POS (-1): there-^{*}ADV/PRON be-v → haber-v
3. Fixed shift (-1): will-v → ∅ [future tense]

(11) **Thank-v** you **so-ADV****much-ADV**. 3, 18 → **Muchas-^{*}DET/ADJ** **gracias**. 1, 15
-133.37 [-2 lexical words] (2 shifts)

1. False shift POS (-1): much-ADV → muchas-^{*}DET/ADJ
2. Free shift (-1): so-ADV → ∅ [DeepL: *Muchísimas gracias*.]

(75) **Why-ADV****now-ADV?** **Why-ADV****now-ADV?** 4, 17 → ¿**por qué ahora-ADV?** 1, 15
-125.86 [-3 lexical words] (2 shifts)

1. False shift POS (-1): why-ADV (1) → por qué (0)
2. Free shift (-2): Why now? (2) → ∅ (0) [confirmed with DeepL, translated twice]

(259) It **will take time**. 3, 18 → **Tomará tiempo**. 2, 14
-83.19 [-1 lexical word] (1 shift)

1. Fixed shift (-1): will-v → ∅ [future tense]

(95) It is **called Miami**. 3, 19 → se **llama Miami**. 2, 15

-78.43 [-1 lexical word] (1 shift)

1. Fixed shift (-1): is called (2) → se llama (1) [confirmed with DeepL]

(132) What **changes come will depend** upon the **Cuban people**. 6, 52 → Lo que **cambie dependerá del pueblo cubano**. 4, 42

-71.15 [-2 lexical words] (2 shifts)

1. Free shift (-1): changes-N come-v (2) → lo que cambie-v (1) [DeepL: *Los cambios que se produzcan...*]
2. Fixed shift (-1): will-v → Ø [future tense]

(187) that is a **measure** of our **progress** as a **democracy**. 4, 49 → Esa es la **medida** de nuestro **progreso**. 3, 37

-70.40 [-1 lexical word] (1 shift)

1. Free shift (-1): democracy-N (1) → Ø (0)

(174) I **am not saying** this is **easy**. 5, 29 → No **digo** que sea **fácil**. 4, 22

-56.31 [-1 lexical words] (1 shifts)

1. Fixed shift (-1): am saying (2) → digo (1)

(173) That **was** because of the **freedoms** that **were afforded** in the **United States** that we **were able** to **bring** about-ADP **change**. 10, 113 → Eso **fue** por las **libertades otorgadas** en los **Estado Unidos** que **pudimos traer** el **cambio**. 8, 86

-55.67 [-2 lexical words] (2 shifts)

1. Fixed shift (-1): were afforded (2) → otorgar (1) [confirmed with DeepL]
2. Fixed shift (-1): were able (2) → pudimos (1) [confirmed with DeepL]

(186) Who **would have believed** that back-ADV in **1959?** 5, 44 → ¿**Quién habría apostado** por eso en **1959?** 3, 39

-51.14 [-2 lexical words] (2 shifts)

1. Fixed shift (-1): would have (2) → habría (1) [conditional tense]
2. Fixed shift (-1): back-ADV → Ø

Éric André Poirier

(178) That's-*PROPN/POSS **how**-ADV we **made enormous gains in women's rights and gay rights**. 10, 68 → **Es como**-*V/ADV **hicimos grandes avances en los derechos de las mujeres y de los homosexuales**. 8, 85

-45.67 [-2 lexical words] (2 shifts)

1. False shift POS (-1): 's-*PROPN/POSS → ∅
2. Free shift (-1): rights-N and ...rights-N (2) → derechos-N y ... (1) [DeepL: ...los derechos de las mujeres y los derechos de los gays.]

(228) that is **why**-ADV their **heartache is so great**. 6, 40 → **Es por**-ADP **eso**-PRON que la **pena en sus corazones es tan grande**. 5, 51

-42.06 [-1 lexical word] (3 shifts)

1. Fixed shift (-1): why-ADV (1) → por-ADP eso-PRON (0)
2. Antinomic fixed shift (+1): heartache-N (1) → pena-N en sus corazones-N (2)
3. Free shift (-1): is-v (1) → ∅ (0) [DeepL: ...su dolor de corazón es tan grande.]

(76) There-ADV **is one**-NUM **simple answer**: 5, 27 → **La respuesta es sencilla**: 3, 25

-36.39 [-2 lexical words] (2 shifts)

1. Free shift (-1): there-ADV → ∅ [DeepL: ...Hay una respuesta simple.]
2. Fixed shift (-1): one-NUM → ∅ or una-DET

(188) So **here**-ADV **is my message to the Cuban government and the Cuban people**: 7, 67 → **Este es mi mensaje para***V/ADP **el gobierno y pueblo de Cuba**: 6, 53

-35.39 [-1 lexical word] (3 shifts)

1. Fixed shift (-1): here-ADV → ∅
2. Antinomic false shift POS (+1): to-ADP → para, parir *V/ADP
3. Free shift (-1): *Cuban government and the Cuban people* → *el gobierno y pueblo de Cuba* [confirmed with DeepL]

(10) **Muchas**-*PROPN/ADJ **gracias**. 2, 15 → **Muchas gracias**. 1, 15

-33.23 [-1 lexical word] (1 shift)

1. False shift POS (-1): muchas-*PROPN/ADJ → muchas *DET/ADJ [target expression used in source text]

(161) We **do** have **too**-ADV **much**-ADJ **money** in **American** politics. 7, 47 → Sí-ADV que **hay demasiado dinero** en la **política estadounidense**. 6, 58
-31.80 [-1 lexical word] (1 shift)

1. Fixed shift (-1): too-ADV much-ADJ → demasiado-ADV [confirmed with DeepL]

(56) For all-DET of our **differences**, the **Cuban** and **American** **people** share **common** values in their **own**-ADJ **lives**. 9, 97 → Con todas nuestras **diferencias**, el **pueblo estadounidense** y el **pueblo cubano** **comparten** los **mis-**mos-*DET/ADJ **valores** en sus **propias**-*DET/ADJ **vidas**. 8, 126
-31.72 [-1 lexical word] (3 shifts)

1. Antinomic free shift (+1): people (1) → pueblo y pueblo (2)
2. False shift POS (-1): common values → mismos-*DET/ADJ valores-N
3. False shift POS (-1): own-ADJ lives → propias-*DET/ADJ vidas N

(63) the **United States** is a **multi**-ADJ **party** **democracy**. 7, 45 → **Estados Unidos** es una **democracia** de **múltiples** **partidos**. 6, 55
-30.36 [-1 lexical word] (2 shifts)

1. False shift POS (-1) : -ADJ [-1] → Ø [0]
2. Undetected fixed shifts (+1) : multi-party-N (*3/1) → múltiples-ADJ partidos-N (2)

(194) **Many**-ADJ **suggested** that I **come** **here**-ADV and ask the **people** of **Cuba** to **tear something**-N **down**-ADV – but I am **appealing** to the **young** **people** of **Cuba** who **will lift something**-N up, **build something**-N **new**. 21, 180
→ **Muchos**-PRON **han sugerido** que **vengo aquí**-ADV **para**-*AUX/ADP **pedir** al **pueblo cubano** que **destruya** algo-PRON; pero yo me **dirijo** a los **jóvenes** de **Cuba** quienes **alzarán** y **construirán** algo-PRON **nuevo**. 15, 163
-29.80 [-6 lexical words] (10 shifts)

1. False shift POS (-1): many *ADJ/PRON → muchos-PRON
2. Antinomic free shift (+1): suggested → han sugerido [DeepL: *Muchos me sugirieron...*]

Éric André Poirier

3. Antinomic false shift POS (+1): and-CONJ → para, parir *AUX/ADP
4. Fixed shift (-1): something-N → algo-PRON [confirmed with DeepL: ...
que derribara algo...]
5. Fixed shift (-1): tear-v down-ADV → destruya-v
6. Fixed shift (-1): am-v → yo-PRON
7. Fixed shift (-1): young people → jóvenes
8. Fixed shift (-1): will → Ø [future tense]
9. Fixed shift (-1): something-N → algo-PRON
10. Fixed shift (-1): something-N → Ø (algo-PRON)

(207) It gives everyone-*N/PRON in this hemisphere hope. 4, 42 → Le brinda
esperanza a todos-PRON en este hemisferio. 3, 47

-29.14 [-1 lexical word] (1 shift)

1. False shift POS (-1): everyone-*N/PRON → todos-PRON

We found 46 information shifts in sample A, with 12 false shifts POS (due to various POS tagging errors), 24 fixed shifts, and 10 free shifts. For all types of shifts, 5 antinomic shifts were found. We will not go into the details of the analysis but provide to the reader a brief survey of what we can deduct from the data collected. A more detailed review of these results is of high interest for translation studies and training but deserves to be addressed in a separate publication. First, this sample contains mostly fixed information shifts due to source language constraints such as verb compositions (modals, active/passive (mandatory)), transformations for verbal constructions exclusive to one language (*is called* translated by *se llama* in pair 95, or *there be* translated by *haber* in pair 258, for example), some peculiar uses of adverbs in English that may be omitted in Spanish or are translated by a preposition, and some English-Spanish POS tagging difference regarding functional words such as pronouns (in pair 194, the pronoun something is analyzed as a noun and translated with the pronoun *algo*, for example). Second, regarding the free information shifts found in sample A, those are far fewer in number. Some can be explained with the traditional concepts of “concentration” or “concision” used in translation studies. Most of them seem to be due to some sort of Spanish grammatical “flexibility” or “freedom” which allows to the translation process very acceptable syntactical reductions of redundant information in source language such as non-repetitions of generic nouns in noun phrase coordination like in pair Cuban government and Cuban people reduced to *el gobierno y pueblo de*

Cuba. We found one characteristic omission of the notion of democracy in pair 187 that is due to the different political systems of reference between the United States and Cuba, but that illustrates very well one political issue between the two countries.

4.2 Sample B annotations – most positive heteromorphic pairs

This section contains the manual analysis and annotations of sample B segments for the classification of information shifts observed in the most positive heteromorphic segment pairs. Translation pairs are presented from the highest positive weighted Euclidean distance (313.26) to the lowest positive weighted Euclidean distance (52.35).

(144) **Not everybody-N agrees with-ADP me on this.** 3, 37 → **No todo-DET el mundo está-AUX de acuerdo-*V/N conmigo-*AUX/ADP sobre-*V/ADP esto.** 6, 52

313.26 [+3 lexical words] (3 shifts)

1. Fixed shift (+1): agree-v (1) → está-v de acuerdo-*v/N (2)
2. False shift POS (+1): with-ADP → conmigo-*AUX/ADP
3. False shift POS (+1): on-ADP → sobre, sobrar-*v/ADP

(177) **that is how-ADV we got health care for more-ADJ of our people.** 7, 54 → **Es como-*AUX/CONJ conseguimos servicios de salud para-*v/ADP una mayor cantidad de personas-*v/N del país.** 10, 84

165.63 [+3 lexical words] (4 shifts)

1. Antinomic false shift POS (-1): how-ADV → como-*AUX/CONJ
2. False shift POS (+1): for-ADP → para, parir-*v/ADP
3. Free shift (+1): more-ADJ (1) → mayor-ADJ cantidad-N (2)
4. Free shift (+1): people-N (1) → personas-*v/N del país-N (2) [DeepL: ... *para más de nuestra gente.*]

(191) **And we – like-ADP every-DET country – need the space that democracy gives us to change.** 6, 81 → **Y nosotros –al-*v/ADP+DET igual-*ADV/N que todos-DET los países– necesitamos el espacio que la democracia nos-*ADV/PRON da-AUX para-*AUX/ADP cambiar.** 10, 104

152.58 [+4 lexical words] (4 shifts)

Éric André Poirier

1. Fixed shift (+1): like-ADP (1) → -al-*v/ADP+DET igual-*ADV/N (loc adv) (*2/1)
2. False shift POS (+1): Ø → -al*v/ADP+DET
3. False shift POS (+1): us-PRON → nos, no-*ADV/PRON
4. False shift POS (+1): to-ADP → para, parir-*AUX/ADP

(46) We have welcomed both-DET immigrants who came a great distance to start new lives in the Americas. 10, 94 → Ambos-**NUM** hemos abierto **nuestras** **puertas** a **inmigrantes** que **recorrieron** **grandes** **distancias**-*AUX/N **para**-*AUX/ADP **empezar** **vidas** **nuevas** en el **continente** **americano**. 14, 139
134.12 [+4 lexical words] (4 shifts)

1. False shift POS (+1): both-DET → ambos-**NUM**
2. Free shift (+1): have welcomed (2) → hemos abierto puertas (3) [DeepL: *hemos acogido a ambos inmigrantes ...*]
3. False shift POS (+1): to-ADP → para, parir-*AUX/ADP
4. Free shift (+1): Americas (1) → continente americano (2) [DeepL : *...en las Américas.*]

(157) I welcome this open debate and dialogue. 4, 40 → **estoy** **dispuesto** a **tener** este **debate** y **diálogo** **abierto**. 6, 54
134.12 [+2 lexical words] (1 shift)

1. Free shift (+2): welcome-v (1) → estoy-v dispuesto-ADJ a tener-v (3) [DeepL: *Me complace este ...*]

(183) You can see that in the election going on back-ADV home-*ADV/N. 6, 52 → Lo **podemos** **apreciar** en las **elecciones** que **están** en **curso** **ahora**-ADV **mismo**-ADJ en mi **país**. 8, 80
123.81 [+2 lexical words] (2 shifts)

1. Free shift (+3): going-v (1) → están-v en curso-N ahora-ADV mismo-ADJ (4)
2. Antinomic free shift (-1): back-ADV home-*ADV/N (2) → en mi país (1) [DeepL : *... que se están llevando a cabo-LOC-ADV en casa.*]

(38) I want to be clear: 3, 19 → **Quiero** **dejar** una **cosa** **clara**: 4, 28
111.83 [+1 lexical word] (1 shift)

1. Free shift (+1): be clear-v (2) →dejar-v una cosa-N clara-ADJ (3) [DeepL : *Quiero ser claro*.]
- (145) Not everybody-N agrees with the American people on this. 5, 54 →No todo el mundo está de acuerdo con el pueblo estadounidense sobre-*v/ADP esto. 7, 73
- 106.34 [+2 lexical words] (2 shifts)
1. Fixed shift (+1): agree-v (1) →está-v de acuerdo-N (2)
 2. False shift POS (+1): on-ADP →sobre, sobrar-*v/ADP
- (113) It is an outdated burden on the Cuban people. 5, 45 →Es una carga anticuada que lleva a costas el pueblo cubano. 7, 60
- 101.61 [+2 lexical words] (1 shift)
1. Free shift (+2): on-ADP (0) →llevar-v a costas-N (2) [DeepL : ... *anticuada para el pueblo cubano*.]
- (122) It is up to you. 1, 16 →Eso es cosa suya. 2, 17
- 83.47 [+1 lexical word] (1 shift)
1. Free shift (+1): is-v up (1) →es-v cosa-N (2) [DeepL : *Depende de usted*.]
- (41) But before-ADP I discuss those issues, we also-ADV need to recognize how-ADV much-ADJ we share. 8, 79 →Pero antes-ADV de hablar sobre-*v/ADP esos temas, también-ADV es nuestro deber reconocer cuánto-ADJ tenemos en común. 11, 98
- 71.63 [+3 lexical words] (5 shifts)
1. Fixed shift (+1): before-ADP →antes-ADV
 2. False shift POS (+1): Ø →sobre, sobrar-*v/ADP
 3. Free shift (+1): we need-v (1) →es-v nuestro deber-N (2) [DeepL: ... *también necesitamos reconocer cuánto compartimos*.]
 4. Antinomic fixed shift (-1): how-ADV much-ADV (2) →cuánto-ADJ (1)
 5. Free shift (+1): share-v (1) →tenemos-v en común-N (2)
- (27) The blue waters beneath-ADP Air Force One once-ADV carried American battleships to this island – to liberate, but also-ADV to exert control over

Éric André Poirier

Cuba. 15, 139 → Las aguas azuladas bajo-*v/ADJ Air Force One transportaron en su día los barcos de batalla estadounidenses hasta esta isla, para-*AUX/ADP liberar pero-CONJ también-ADV para-*AUX/ADP ejercer control sobre-*v/ADP Cuba. 20, 175

67.31 [+5 lexical words] (5 shifts)

1. Fixed shift (+1): beneath-ADP → baja, bajar-*v/ADJ [confirmed with DeepL: ... *bajo* ...]
2. False shift POS (+1): to-ADP → para, parir-*AUX/ADP
3. False shift POS (+1): to-ADP → para, parir-*AUX/ADP
4. False shift POS (+1): over-ADP → sobre, sobrar-*v/ADP
5. Free shift (+1): battleships-N (1) → barcos-N de batalla-N (2) [DeepL: ... *acorazados* ...]

(190) Not because American-ADJ democracy is perfect, but precisely because we are not. 8, 76 → No porque pienso que la democracia en Estados-PROPN Unidos-PROPN sea perfecta, sino precisamente porque no lo somos. 10, 104

66.67 [+2 lexical words] (2 shifts)

1. Free shift (+1): Ø → pienso-v
2. Free shift (+1): American-ADJ (1) → Estados-PROPN Unidos-PROPN (2) [DeepL: *No porque la democracia americana sea perfecta* ...]

(155) He has a much-ADV longer-ADJ list-N. 4, 26 → Él tiene una mucho-ADV más-ADV lista-N larga-ADJ. 5, 35

63.25 [+1 lexical word] (1 shift)

1. Fixed shift (+1): much-ADV (1) → mucho-ADV más-ADV (2)

(128) Before 1959, some Americans saw Cuba as-ADP something-N to exploit, ignored poverty, enabled corruption. 10, 98 → Y desde 1959, algunos-PRON estadounidenses veían Cuba como-v un lugar del que se podían aprovechar, ignoraron la pobreza y permitieron la corrupción. 12, 142

61.40 [+2 lexical words] (2 shifts)

1. False shift POS (+1): as-ADP → como, comer-*v/ADP
2. Free shift (+1): something-N to exploit-v (2) → lugar-N del que se podían-v aprovechar-v (3) [DeepL: ... *como algo para explotar* ...]

- (129) And since-ADP 1959, we have been shadow-boxers in this battle of geopolitics and personalities. 8, 91 → Desde-ADP 1959, hemos sido como-*V/ADP boxeadores con un contrincante-*ADV/N imaginario en esta batalla de geopolítica y personalidades. 10, 118

55.43 [+2 lexical words] (2 shifts)

1. False shift POS (+1): Ø → como, comer-*V/ADP
2. Free shift (+1): shadow-boxers-N (2) → boxeadores-N con un contrincante-*ADV/N imaginario-ADJ (3) [DeepL: ... *hemos sido boxeadores en la sombra en esta batalla* ...]

- (45) Like-ADP the United States, the Cuban people can trace their heritage to both slaves and slave-owners. 10, 98 → Al igual-*ADJ/N que en Estados Unidos, el pueblo cubano puede encontrar sus orígenes tanto-*ADV/PRON en los esclavos-*PRON/N como-*V/CCONJ en los dueños de los esclavos-ADJ. 12, 135

53.21 [+2 lexical words] (5 shifts)

1. Fixed shift (+1): like-ADP (0) → Al igual-*ADJ/N (1)
2. False shift POS (+1): both-DET → tanto-*ADV/PRON
3. Antinomic false shift POS (-1): slaves-N → esclavos-*PRON/N
4. False shift POS (+1): and-CCONJ → como, comer-*V/CCONJ
5. Undetected fixed shift [+1] : slave-owners-N (*2/1) → dueños-N de los esclavos-N (2)

- (91) Hope that is rooted in the future that you-PRON can choose and that you-PRON can shape, and that you-PRON can build for your country. 11, 118 → Esperanza que tiene una base en el futuro que ustedes-*N/PRON pueden elegir; que ustedes-*N/PRON pueden moldear; que ustedes-*N/PRON pueden construir para-*V/ADP su país. 15, 139

53.15 [+4 lexical words] (4 shifts)

1. False shift POS (+1): you-PRON → ustedes, vosotros-*N/PRON
2. False shift POS (+1): you-PRON → ustedes, vosotros-*N/PRON
3. False shift POS (+1): you-PRON → ustedes, vosotros-*N/PRON
4. False shift POS (+1): for-ADP → para, parir-*V/ADP

Éric André Poirier

- (70) We have begun initiatives to cooperate on health and agriculture, education and law enforcement. 9, 96 → Hemos lanzado iniciativas para-*v/ADP cooperar en temas-*v/N de salud y agricultura, educación y autoridades del orden público. 12, 115

53.13 [+3 lexical words] (3 shifts)

1. False shift POS (+1): to-ADP → para, parir-*v/ADP
2. False shift POS (+1): Ø → temas, temer-*v/N
3. Free shift (+1): law enforcement (2) → autoridades del orden público (3) [DeepL: ... la aplicación de la ley.]

- (25) Havana is only 90 miles from Florida, but to get here-ADV we had to travel a great distance – over barriers of history and ideology; 15, 129 → La Habana se encuentra tan-*N/ADV solo-*N/ADV a 90 millas de Florida, pero para-*AUX/ADP llegar hasta aquí tuvimos que recorrer una gran distancia: derribar-v las barreras de la historia y la ideología; 18, 177

52.56 [+3 lexical words] (3 shifts)

1. Free shift (+1): only-ADV → tan-*N/ADV solo-*N/ADV [DeepL: La Habana está a sólo 90 millas ...]
2. False shift POS (+1): to-ADP → para, parir-*AUX/ADP
3. Free shift (+1): over-ADP → derribar-v) [DeepL: ... por encima de las barreras de la ...]

- (135) But having removed the shadow of history from our relationship, I must speak honestly about the things that I believe – the things that we, as Americans, believe. 13, 63 → Pero ahora-ADV que hemos quitado la sombra de la historia de nuestra relación, debo hablar honestamente sobre-*v/ADP las cosas en las que yo creo --*AUX/PUNCT las cosas en las que nosotros, como-*v/ADP estadounidenses, creemos. 17, 198

52.35 [+4 lexical words] (4 shifts)

1. Free shift (+1): Ø → ahora-ADV [DeepL: Pero habiendo eliminado la sombra ...]
2. False shift POS (+1): about-ADP → sobre, sobrar-*v/ADP
3. False shift POS (+1): – → --*AUX/PUNCT
4. False shift POS (+1): as-ADP → como, comer-*v/ADP

Sample B contains 59 information shifts in total with 30 false shift POS, 8 fixed shifts, and 21 free shifts. For all types of shifts, 4 antinomic shifts and 1 undetected shift were found. By comparison with sample A, we can see that free shifts are more than two times the number of fixed shifts, which were two times more numerous than the former in sample A. Here is a brief overview of the trend we can observe from the annotations. A lot of the numerous positive free shifts seem to be associated with some form of mandatory and translation-inherent explicitations (see [Blum-Kulka \(1986\)](#) who proposed the explicitation hypothesis, as well as the work of [Becher \(2010; 2011\)](#) who rejected the hypothesis and the more recent synthesis article by [Murtisari \(2016\)](#) on the concept of explicitation in translation studies). One first example is the periphrastic translation of *shadow-boxers* with *boxeadores con un contrincante imaginario* in pair 129 or the creative translation of *battleship* with *barcos de batallas* in pair 27. A good example of “political” explicitness that tends to reduce a statement is found in pair 190 when President Obama state that American democracy could be seen as “perfect” (even though he clearly states that this is not the case). The translation makes explicit that the statement is its own way of thinking by adding the verb phrase *pienso que*. As regards the eight positive fixed shifts, these are mostly the opposite operations that were described in the analysis of sample A negative fixed shifts, such as the addition of an adverb or the translation of a preposition by an adverb as in pair 41. These last data validate the existence of mandatory explicitations and implicitations processes that are symmetrical and dependent from syntactic and lexical structures of languages (see [Klaudy 2011](#)).

4.3 Sample C annotations – most isomorphic and less heteromorphic pairs

This section contains the manual analysis and annotations of sample C segments for the classification of information shifts observed in the most isomorphic and less heteromorphic segment pairs. Translation pairs all have near-zero weighted Euclidean distance and are presented from the highest negative weighted Euclidean distance (-0.42) to the highest positive weighted Euclidean distance (1.29).

(176) But **democracy** is the way that we **solve** them. 4, 44 → Pero la **democracia** es la **forma** de **cambiarlos**. 4, 45

-0.42 [0 difference in lexical words] (0 shift)

(104) Look at **Papito Valladeres**, a **barber**, whose-DET **success** allowed him to **improve conditions** in his **neighborhood**. 9, 105 → Miren a Papito Val-

Éric André Poirier

laderes, un barbero, cuyo-PRON éxito le permitió mejorar las condiciones en su vecindario. 9, 103

-0.36 [0 difference in lexical words] (0 shifts)

(133) We **will not impose** our **political** or **economic system** on you. 6, 59 → No **vamos a imponerles** nuestro **sistema político** ni **económico**. 6, 60

-0.31 [0 difference in lexical words] (0 shifts)

(18) We **will-v do** whatever is **necessary** to **support** our friend and ally, Belgium, in **bringing** to justice those who are **responsible**. 12, 123 → **Haremos** lo que **sea necesario para-***AUX/ADP **apoyar** a nuestra amiga y aliada, **Bélgica, para-***AUX/ADP **ajusticiar-v** a aquellos que **sean responsables**. 12, 125

-0.30 [0 difference in lexical words] (4 shifts)

1. Fixed shift (-1): will-v do-v (2) → haremos-v (1)
2. False shift POS (+1): to-ADP → para-^{*}AUX/ADP
3. False shift POS (+1): in-ADP → para-^{*}AUX/ADP
4. Free shift (-1): bringing-v to justice-N (2) → ajustar-v (1) [DeepL: ..., *para llevar a la justicia a los responsables.*]

(234) And I **have come here-ADV** – I **have traveled** this **distance** – on a **bridge** that **was built** by Cubans on both-DET **sides** of the **Florida Straits**. 13, 131 → Y **he venido aquí-ADV** – **he-***ADP **viajado** esta **distancia** – **sobre-***V/ADP un **punte construido-ADJ** por los **cubanos** a **ambos-NUM** **lados** del **Estrecho** de la **Florida**. 13, 129

-0.29 [0 difference in lexical words] (4 shifts)

1. False shift POS (-1): have-AUX → he-^{*}ADP/AUX
2. False shift POS s(+1): on-ADP → sobre, sobars-^{*}V/ADP
3. Free shift (-1): was-AUX built-v (2) → construido-ADJ (1) [DeepL: ... *en un puente que fue construido por ...*]
4. Fixed shift (+1): both-DET → ambos-NUM [DeepL: ... *a ambos lados del Estrecho de Florida.*]

(216) I **know** that **many-ADJ** of the **issues** that I **have talked** about **lack** the **drama** of the **past**. 8, 83 → Sé que **muchos-ADV** de los **problemas** de los que **he hablado** **carecen** del **drama** del **pasado**. 8, 82

-0.23 [0 difference in lexical words] (0 shift)

- (208) We took different journeys to our support for the people of South Africa in ending apartheid. 9, 93 → Tomamos diferentes pasos en nuestro apoyo al pueblo de Sudáfrica para-^{*AUX/ADP} acabar con el apartheid. 9, 94
-0.20 [0 difference in lexical words] (2 shifts)
1. Fixed shift (-1): South Africa (2) → Sudáfrica (1)
 2. False shift POS (+1): in ADP → para-^{*AUX/ADP}
- (201) But-CONJ no-DET one should deny the service that thousands of Cuban doctors have delivered for the poor and suffering. 10, 109 → Pero-CONJ nadie-PRON debe negar el servicio que miles de médicos cubanos han prestado a los pobres y a los que sufren. 10, 108
-0.17 [0 difference in lexical words] (0 shift)
- (84) Creo en el pueblo Cubano. 3, 25 → Creo en el pueblo cubano. 3, 25
0.00 [0 difference in lexical words] (0 shift)
- (86) This is not just-ADV a policy of normalizing relations with the Cuban government. 8, 77 → Esto no es solo-ADJ una política de normalizar relaciones con el gobierno Cubano; 8, 77
0.00 [0 difference in lexical words] (0 shift)
- (102) Look at Sandra Lidice Aldama, who chose to start a small business. 8, 66 → Miren a Sandra Lidice Aldama, que eligió abrir un pequeño negocio. 8, 66
0.00 [0 difference in lexical words] (0 shift)
1. Note: This is the central pair of isomorphic segments, being exactly in the middle of two other isomorphic segment pairs.
- (151) the death penalty; 2, 18 → la pena de muerte; 2, 18
0.00 [0 difference in lexical words] (0 shift)
- (66) Cuba has emphasized the role and rights of the state; 6, 53 → Cuba ha reforzado el papel y los derechos del estado; 6, 53
0.00 [0 difference in lexical words] (0 shift)

Éric André Poirier

- (197) And **given-v** your **commitment** to Cuba's-^{*PROPN/POSS} **sovereignty** and **self-determination**, I am **also-ADV** **confident** that you **need not** **fear** the **different** **voices** of the **Cuban** **people** – and their **capacity** to **speak**, and **assemble**, and **vote** for their **leaders**. 22, 229 → **Teniendo en cuenta-v/N** su **compromiso** con la **soberanía** y la **autodeterminación** de Cuba, **también-ADV** **estoy** **seguro** de que **no** **tiene** que **temer** las **diferentes** **voces** del **pueblo** **cubano** – y-^{*ADJ/CCONJ} su **capacidad** **par-v/ADP** **hablar**, y **reunirse**, y **votar** por sus **líderes**. 23, 233

0.42 [+1 lexical word] (5 shifts)

1. Free shift (+1): given-v (1) → **teniendo-v en cuenta-v/N** (2) [DeepL: Y dado su compromiso con la soberanía ...]
2. False shift POS (-1): 's-^{*PROPN/POSS} → **de-ADP**
3. False shift POS (-1): self-determination (2) → **autodeterminación** (1)
4. False shift POS (+1): – and-ADP → – y-^{*ADJ/CCONJ}
5. False shift POS (+1): to-ADP → **para-v/ADP**

- (236) And I **know** **how-ADV** they **have** **suffered** **more-ADJ** than the **pain** of **exile** – they also **know** what it is like-ADP to **be** an **outsider**, and to **struggle**, and to **work** **harder** to **make** **sure** their **children** **can** **reach** **higher** in **America**. 22, 207 → Y sé que **han** **sufrido** **más-ADV** que el **dolor** del **exilio**: **saben** lo que se **siente** al **ser** un **extraño**, al **luchar**, al **trabajar** **más-ADV** **duro-v/AUX/ADJ** **para-v/ADP** **asegurarse** de que sus **hijos** **puedan** **llegar** **más-ADV** **lejos-ADV** en los **Estados** **Unidos**. 23, 203

0.47 [1 lexical word] (5 shifts)

1. Fixed shift (-1): how-ADV (1) → **que-SCONJ** (0)
2. Fixed shift (+1): harder-ADJ (1) → **más-ADV** **duro-v/AUX/ADJ** (2)
3. False shift (+1): to-ADP → **para-v/ADP**
4. Fixed shift (+1): higher-ADJ (1) → **más-ADV** **lejos-ADV** (2)
5. Fixed shift (-1): make-v sure-ADJ → **asegurarse-v**
6. Free shift (+1): America-PROPN → **Estados-PROPN** **Unidos-PROPN** [DeepL: ...*más alto en América.*]

- (162) But, in **America**, it is **still-ADV** **possible-ADJ** for **somebody-v/N/PRON** like-ADP me – a **child** who **was** **raised** by a **single** **mom**, a **child** of **mixed** **race** who **did not** **have** a **lot** of **money** – to **pursue** and **achieve** the **highest-ADJ**

office in the land. 23, 212 → Pero en EEUU, todavía-ADV es posible que alguien-PRON como-*V yo, un niño que fue criado por una madre soltera, un niño de raza mixta que no tenía mucho-DET dinero, pueda-v ir-v atrás-ADV de y conseguir el cargo más alto del país. 24, 206

0.63 [1 lexical word] (6 shifts)

1. False shift POS (-1): somebody-*N/PRON → alguien-PRON
2. False shift POS (+1): like-ADP → como-*V/ADP
3. Fixed shift (-1): did not have (3) → no tenía (2)
4. False shift POS (-1): a lot-N → mucho-*DET/ADJ
5. Free shift (+2): pursue-v (1) → pueda-AUX ir-v atrás-ADV de (3) [DeepL: *...persiga y logre el cargo más alto de la tierra.*]
6. Fixed shift (+1): highest-ADJ → más-ADV alto-ADJ

(209) But **President Castro** and I could both-DET be there-ADV in **Johannesburg** to-ADP pay tribute to the legacy of the great Nelson Mandela. 12, 120 → Pero el presidente Castro y yo pudimos estar allí-ADV en **Johannesburgo** para-*AUX/ADP rendir homenaje al legado de gran Nelson Mandela. 13, 121

0.69 [1 lexical word] (1 shift)

1. False shift POS (+1): to-ADP → para-*AUX/ADP

(136) As **Marti said**, “-*PROPN/PUNCT **Liberty** is the right of every-DET man to be honest, to think and to speak without hypocrisy.”-PUNCT 12, 105 → Como dijo Martí: “-*PROPN/PUNCT La libertad es el derecho de todo-DET hombre a ser honesto, pensar y hablar sin hipocresía”-*N/PUNCT. 13, 106

0.74 [1 lexical word] (1 shift)

1. False shift POS (+1): ”hypocrisy.”-PUNCT → hipocresía”-*N/PUNCT

(214) From the **beginning** of my time-N in office, I have urged the people of the Americas to leave behind-ADP the ideological battles of the past. 11, 133 → Desde el inicio de mi mandato, he instado a los pueblos del continente americano a dejar atrás-ADV las batallas ideológicas del pasado. 12, 131

1.16 [1 lexical word] (3 shifts)

1. Free shift (-1): time-N in office-N → mandato-N [DeepL: *Desde el principio de mi tiempo en la oficina ...*]

Éric André Poirier

2. Free shift (+1): Americas (1) → continente americano (2) [DeepL: *He instado a los pueblos de América ...*]

3. Fixed shift (+1): leave-v behind-ADP (1) → dejar-v atrás-ADV (2)

(210) And in **examining** his **life** and his **words**, I **am** **sure** we both-DET **realize** we **have more-ADJ work** to do to **promote equality** in our **own-ADJ countries** – to-ADP **reduce discrimination** based on race in our **own-ADJ countries**. 20, 195 → Y al **examinar** su **vida** y sus **palabras**, **estoy seguro** de que **ambos-NUM nos-^{*}ADV/PRON damos-v cuenta-N** de que **tenemos mucho trabajo** por **hacer** --^{*}PROPN/PUNCT **para-^{*}AUX/ADP reducir** la **discriminación basada** en la raza en **ambos países**. 21, 187

1.18 [1 lexical word] (6 shifts)

1. Fixed shift (+1): both-DET → ambos-NUM

2. False shift POS (+1): we-PRON → nos-^{*}ADV/PRON

3. Free shift (-4): to promote equality in our own countries → ∅ [DeepL: *... para promover la igualdad en nuestros propios países.*]

4. Fixed shift (+1): realize-v (1) → damos-v cuenta-N (2)

5. False shift POS (+1): --PUNCT → --^{*}PROPN/PUNCT

6. False shift POS (+1): to-ADP → para-^{*}AUX/ADP

(239) “--^{*}PROPN/PUNCT You **recognized** me, but I **did-v not recognize** you,” -PUNCT **Gloria said** after-ADP she **embraced** her **sibling**. 9, 93 → “--^{*}PROPN/PUNCT Tú me **reconociste**, pero yo-PRON **no te reconocí**” -^{*}PROPN/PUNCT, le **dijo** Gloria a su **hermana después-ADV** de **abrazarla**. 10, 94

1.29 [+1 lexical words] (3 shifts)

1. Fixed shift (-1): did V → yo-PRON

2. Fixed shift (+1): after-ADP (0) → después-ADV (1)

3. False shift (+1): “-PUNCT (0) → ^{*}PROPN/PUNCT (1)

In this sample, antinomic shifts may be positive or negative since the segment pairs all have zero or near zero weighted Euclidean distance while almost half of them are negatively close to zero or positively close to zero. Sample C has 40 information shifts in total, with 18 false shifts POS, 14 fixed shifts, 8 free shifts, and 19 antinomic shifts. This sample contains the highest number of antinomic information shifts among the three samples annotated. Because the TPR of most

segments is neutral ($=1.0$), the number of antinomic shifts is doubled as it is the case for segments 18, 234, and 208 which account for 10 antinomic shifts. The other 9 antinomic shifts appear in lengthy segment pairs having a small positive TPR value. Their number could be reduced if the segmentation of the text could have a finer or smaller granularity to the level at least of propositions. This is a development that would enhance the efficiency of the empirical screening method of information shifts described here.

We can observe for sample C annotations that segments having zero weighted Euclidean distance contain no information shift at all. The number of these segments is small (5), but it's worth noting the efficiency of the method for screening pairs having no information shift at all. We can also note that some particular lengthy segment pairs have a lot of information shifts while all the other short segment pairs have zero information shifts. The average length of the 11 segment pairs having at least one information shift is 144 characters while the average length of the 10 segment pairs having zero information shift is less than half of this amount with 63,8 characters. In the case of mostly isomorphic segment pairs, the short length in characters seems to be predictive of the absence of information shifts. This correlation hypothesis needs to be further tested and set for different corpora.

5 Conclusion

We described in detail an empirical method for screening segment pairs in parallel corpora for informational translation shifts. Our manual analysis of the three samples A B and C of parallel pairs screened with our method confirm our hypothesis that heteromorphic segment pairs, as opposed to isomorphic ones, contain higher numbers of informational translation shifts. These tendencies can be observed with the number of information shifts that were detected in the most negative (46) and positive (59) heteromorphic segment pairs, compared to the number of information shifts present (40) in more isomorphic pairs (among which 5 pairs having a weighted Euclidean distance of exactly zero contained no information shift). If we discard the false information shifts which are erroneous, the observation is perhaps strengthened with a lower volume of information shifts (22) in sample C by comparison with 34 in sample A and 29 in sample B. Regarding our hypothesis for information shift screening, another criteria that need to be taken into account is that we found that the length of mostly isomorphic pairs seems to be predictive of the presence or absence of information shift. The discovery of this correlation for sample C pairs needs to be further tested with other corpora and against heteromorphic segment pairs.

Éric André Poirier

What was also found unexpectedly in the annotations of the two most heteromorphic samples is that negative heteromorphic segment pairs tend to contain much more (mandatory) fixed shifts than free shifts while it is the exact opposite for positive heteromorphic segment pairs. This could be in line with the explicitation hypothesis in translation which can be viewed as a tendency to add content in the target segments in translation (thus creating an information asymmetry) by giving more details and explanations than what is given in the source text (to make sure for instance that the content is well understood or clear for the intended audience). Further studies and progress on the empirical methods developed herein are needed to shed light on this result.

A better knowledge of the origin, the cause and the impact of fixed information shifts are essential for a better knowledge of language constraints in translation (in contrastive phraseology, translation difficulties, and their idiosyncratic solutions) while the study of free information shifts should shed lights on cognitive issues in translation operations (errors, individual and cultural biases). The manual examination and categorization of 145 informational shifts have shown that fixed and free shifts are relevant categories for the study of these phenomena. In order to reduce false shifts (false positives) and undetected shifts (false negatives), new POS tagging models and methods in English and in other major languages would need to be developed. The situation was found to be worse for the Spanish language, where many significant errors in parts of speech tagging were found, especially for many simple tokens such as *sober* and *para* used as prepositions that were wrongly tagged as verbs.

In the methodology we propose, we also demonstrated the usefulness of machine translation in the comparison of translation solutions by leveraging the standardization of style and expressions that seem to be favored because of their consumption of the enormous amount of corpus data. In fact, we have shown that machine translation may be used to distinguish automatically, most of the times, fixed shifts, which are confirmed when machine translation also produce the same information shift, from free shifts, which are confirmed when the information shift in human translation is not present in an otherwise grammatically and semantically correct machine translation.

Finally, in the context of increased interest towards more formal and objective methods in human and machine translation assessment and evaluation, we hope that the methodology described in this paper could lay the foundation for language-independent translation assessment procedures and models. For example, the weighted Euclidean distance could be used in association with other automatic translation quality control methods that rely on reviewing translations of

specific lexical items in source segment against conventional translations found in bilingual dictionaries or other reference material or documentations.

Abbreviations

TPR Translation precision ratio
PDF Precision deviation factor

References

- Bakker, Matthijs, Cees Koster & Kitty Van Leuven-Zwart. 2011. Shifts. In Mona Baker & Gabriela Saldanha (eds.), *Routledge encyclopedia of translation studies*, 2nd edn., 269–274. London: Routledge.
- Becher, Viktor. 2010. Abandoning the notion of “translation-inherent” explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1). 1–28.
- Becher, Viktor. 2011. *Explicitation and implicitation in translation: A corpus-based study of English-German and German-English translations of business texts*. Universität Hamburg. (Doctoral dissertation).
- Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House & Shoshana Blum-Kulka (eds.), *Intercultural communication: Discourse and cognition in translation and second language acquisition*, 17–35. Tübingen: Narr.
- Explosion_AI. 2016–2020. *spaCy v2.1.3, Industrial Strength Natural Language Processing in Python*. <https://spacy.io> (5 August, 2020).
- Gale, William A. & Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1). 75–102.
- Giesbrecht, Eugenie & Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In Iñaki Alegria, Igor Leturia & Serge Sharoff (eds.), *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, 27–35. San Sebastian: Elhuyar Fundazioa.
- Klaudy, Kinga. 2011. Explicitation. In Mona Baker & Gabriela Saldanha (eds.), *Routledge encyclopedia of translation studies*, 2nd edn., 104–108. London: Routledge.
- Murtisari, Elisabet Titik. 2016. Explicitation in translation studies: The journey of an elusive concept. *Translation & Interpreting* 8(2). 64–81.

Éric André Poirier

- Poirier, Éric André. 2017. A comparison of three metrics for detecting crosslinguistic variations in information volume and multiword expressions between parallel bitexts. In Ruslan Mitkov (ed.), *Proceedings of EUROPHRAS 2017*, 1–10. Geneva: Editions Tradulex.
- Poirier, Éric André. 2019. Repérage des décalages informationnels de traduction au moyen du criblage automatique des segments hétéromorphes d'un corpus parallèle. *TTR: Traduction, terminologie, rédaction* 32(2). 279–308.
- Wecksteen-Quinio, Corinne, Mickaël Mariaule Corinne & Cindy Lefebvre-Scodeller. 2015. *La traduction anglais-français: Manuel de traductologie pratique*. Louvain-la-Neuve: De Boeck.

Chapter 8

Between audiovisual translation and localization: The case of *Detroit: Become Human*

Laura Mejías-Climent

Universitat Jaume I

The concept of Audiovisual Translation (AVT) is changing continuously in the current technological landscape, in which localization has emerged as a key process to adapt different types of modern multimedia products. Whether AVT encompasses localization, or vice versa, or they can be conceived as different fields instead, remains unclear. It is not my intention to find a unique solution to this unanswered question, but rather to shed some light on the convergences of AVT and localization from the specific perspective of dubbing, in a product that, in turn, poses some questions to the genre it belongs: the graphic adventure *Detroit: Become Human* (Quantic Dream, 2018). This chapter aims to highlight some of the differences and convergences between AVT and localization analyzing the dubbing synchronies applied in a video game belonging to a genre closer to traditional movies, compared to other adventure games, due to the strong presence of cinematic scenes and the lower level of interaction. The results will indicate that, even though clear differences can be found between the localization of a game and translation of non-interactive products, the convergences in terms of dubbing synchronies, particularly in the cinematographic scenes, are quite evident.

1 The current landscape of audiovisual translation

1.1 Introduction

Within the landscape of translation studies, and as a professional practice too, audiovisual translation (AVT) represents an umbrella term encompassing different



Laura Mejías-Climent

translation modes. These modes depend on the nature of the original product and the translated one (Hurtado Albir 2011), as well as the technical methods used to transfer the linguistic message from an original audiovisual text to the target one (Chaume 2004).

Traditionally, AVT is divided into two broad groups (Chaume 2013): revoicing and captioning. The first refers to those modes in which an additional soundtrack is included in the original product: dubbing, voice-overs, simultaneous interpreting of movies, free commentary, fandubs and audiodescription. The latter, captioning, encompasses those modes based on text inserted or next to the screen in which the original product is shown: subtitling, surtitling, respelling, subtitling for the deaf and hard of hearing and fansubbing. Both lists keep broadening together with technological advancements, products and market preferences.

AVT has developed at a rapid pace in the last decades, not only as a professional practice, but also as a research field. “Technological developments have brought about new audiovisual transfer modes, or new combinations of the latter” (Chaume 2018a: 41). With the exponential increase in the audiovisual production during the last decade, the concept of AVT has faced the challenge of dealing with many different types of audiovisual products and translation modes, and with different passive and active forms of consumption. All of this has caused the emergence not only of additional AVT modes, but also of other terms that coexist with that of AVT, sometimes referring to the same concept, sometimes evidencing our technological changing reality more specifically. In all of them, translation is the underlying concept that accounts for the access to any audiovisual product to a different audience.

In the current and changing technological landscape of audiovisual production, some of the “characteristics of AVT that are expanding the borders of the concept of translation” are the following, as Chaume (2018b: 88) points out: not only interlingual and intralingual transfer of content takes place in AVT modes such as subtitling (interlingual) or subtitling for the deaf (intralingual), but also intersemiotic transfers, in the case of audio guides for museums, sign language, and audiodescription. Transadaptation (Neves 2005; Gambier 2003) “could encompass all AVT modes known to date” (Chaume 2018b: 92) and, for Pruys (2009), it consists of two variations of the same topic. Transcreation implies a high level of creativity to “tilt the balance towards the target audience” (Bernal-Merino 2015: 90) and can be understood as another form of semiotic adaptation dealing with many different types of audiovisual texts. Transmedia narratives, rewritings (Bernal-Merino 2015) and media adaptations such as remakes are common practice nowadays. Finally, localization has emerged either as a synonym of a broader

concept of AVT, or as a different professional field dealing with the adaptation of software, websites and video games to a target culture (see §2).

Regarding the concept of *localization*, the borders between AVT and localization are no longer clear, if they ever were, and the use of different AVT modes can be found in any multimodal product nowadays. Whether AVT encompasses localization, or vice versa, or they can be conceived as different fields instead, remains unclear. It is not my intention to find a unique solution to this unanswered question, but rather to shed some light on the convergences of AVT and localization, from the specific perspective of dubbing, in a multimedia product that, in turn, poses some questions to the genre it belongs: the graphic adventure *Detroit: Become Human* (DBH), developed by Quantic Dream.

1.2 The aim of this chapter

In order to trace some similarities, but also to point out some of the differences, that can be observed in the final localized version of a video game and that of any traditional movie, the results of a case study will be discussed in the following pages. The aim is to offer a concise analysis of the similarities between the dubbing of traditional movies and that of graphic adventures, focusing on the case of the video game *Detroit: Become Human* and, more specifically, in the types of synchronies used in the cinematic scenes of this video game. A total of 20 hours of gameplay of this interactive audiovisual product was analyzed, focusing on the different game situations and dubbing synchronies used in each of them, especially, in the cinematic scenes.

To contextualize the analysis, first, the definitions of AVT and localization, and their convergence, will be discussed, as introduced above. Secondly, the characteristics of the AVT mode of dubbing in movies will be reviewed. Thirdly, the same will be done for the features of dubbing in video games, followed by a discussion of the game situations in which the analyzed video game is articulated. Finally, the results of the empirical and descriptive analysis will be presented to determine the convergences between the dubbing of a traditional movie, a fully-interactive video game (Mejías-Climent 2017; 2019) and the case of the graphic adventure *Detroit: Become Human*.

2 Audiovisual translation within localization or vice versa?

As introduced above, AVT encompasses nowadays a wide range of translation modes (Chaume 2004: 31; Hurtado Albir 2011: 54) that challenge the traditional

Laura Mejías-Climent

concept of translation in the strictest sense of “linguistic transfer”. The ever-growing variety of multimodal products that keep breaking into the market require translation practices to adapt, in order to accommodate to these continuous changes in the technological configuration of the products and the way they are consumed. In fact, nowadays the term *localization* itself can encompass

both consolidated as well as new groundbreaking interlingual, intralingual and intersemiotic “audiovisual translation” practices, namely dubbing, subtitling, surtitling, respeaking, audiosubtitling, voice-over and partial dubbing, simultaneous interpreting in film festivals, free-commentary, subtitling for the deaf and the hard of hearing, audio description for the blind and visually-impaired, fansubbing and fandubbing. (Chaume 2018b: 84)

Nonetheless, the term localization emerged in the 1980s when software developers identified the need to adapt their products to expand their markets to other cultures (O’Hagan & Mangiron 2013: 87). With the rapid growth of the game industry, especially since the 1990s, the term localization has settled among professionals and it is generally understood as a more complex process of adaptation, beyond a mere linguistic translation (Bernal-Merino 2006). The author highlights the fact that, however, this term does not refer to anything new that the concept of translation did not include already. Since its use is widely spread among the industry, it seems necessary to accept it within translation studies but always preceded by “linguistic” to differentiate it from the whole adaptation and industrial process described by professionals such as Esselink (2000) or Maxwell-Chandler (2005).

Other scholars and professionals such as Muñoz Sánchez (2017), Granell et al. (2015) or Mangiron & O’Hagan (2006) emphasize the idea that localization includes the adaptation of the interactive product on many different levels (including the linguistic one) to make it meet the needs of the target market completely.

The debate is open about the link between the concept of localization and the field of AVT. As Vázquez Rodríguez (2018: 9–23) points out, a group of professionals defend the position of localization as a differentiated area, since they understand translation from a reductionist and linguistic perspective (cf. Cadieux & Esselink 2004). In addition, some other professionals and scholars also prefer to separate the idea of localization from any other translation mode because of the broad range of adaptation processes it implies, the particularities in the professional practice and the type of product being translated (Pym 2016; Méndez González 2015; Jiménez-Crespo 2013; Mata Pastor 2005).

On the other hand, the term localization does not add anything new to the concept of translation defended by scholars such as Bernal-Merino (2006; 2015) and

O'Hagan & Mangiron (2013). This is the same position that Vázquez Rodríguez (2018) takes himself, acknowledging that the term localization is widely used in professional spheres and, as such, it can be used in research as well, as a label for the translation practice that deals specifically with video games, software and websites. Localization and translation should not be separated, but rather, the first can be included in the latter as another translation mode with defining characteristics.

Having said that, is localization another mode within the field of AVT or is it a completely different area? In both cases, a multimodal product with very similar characteristics—except for the interactive dimension—is translated. Game localization might liken to AVT in the sense that it encompasses some other modalities itself, such as dubbing for cinematic scenes or subtitling dialogues, in addition to some other practices to modify legal, technical or external contents, or even accessibility practices.

However, it should be taken into account that AVT is a very broad field under which any multimodal text can find its particular translation practice. From my point of view, it all depends on the perspective from which both processes are considered. As a professional practice, game localization can be understood as the hypernym under which different translation modes can be gathered. Localization seems to seek for a clear differentiation within the industry, related to the concept of transcreation (O'Hagan & Mangiron 2013). Although no clear definition has been established so far with empirical studies to validate the term transcreation (Bernal-Merino 2015). But that lies outside the scope of these pages. In academic circles, AVT is conceived as the process of adapting any kind of multimodal product, among which video games are included. As O'Hagan & Mangiron (2013) rightly summarize, and as pointed out above,

The emergence of new media resulting from the convergence of technologies is seeing the previously separate domains of localization and AVT come together to cater for the new type of products needing to be prepared to go global. Whether AVT subsumes localization or vice versa remains to be seen, although it is now widely acknowledged that AVT is fast gaining a foothold within Translation Studies. (O'Hagan & Mangiron 2013: 106)

Given the ambiguity in setting one field or practice within the other, I follow Vázquez Rodríguez's (2018) approach, who proposes to adapt AVT research practices to include the interactive dimension and playability in an empirical study, in order to determine the repercussion that both might have in the translation of interactive audiovisual products. Thus, it is not necessary to establish a com-

Laura Mejías-Climent

pletely different paradigm for localization, but just to adapt the existing research approaches used in the AVT field.

When focusing on the analysis of video games, an additional semiotic channel needs to be taken into account in the configuration of the audiovisual product under study (Mejías-Climent 2017). Consequently, some differences in the translation of a movie and a video game can be identified with descriptive analysis. As the following case of study shows, there are clear differences in the product as a whole. But similarities in some specific areas are more prominent than differences, especially in the case of a graphic adventure, which is, ultimately, a movie offering choices to the viewer (thus, a type of narrative and audiovisual content including interaction).

3 Dubbing as an audiovisual translation mode within the process of localization

As is widely known, dubbing “consists of replacing the original track of a film’s (or any audiovisual text’s) source language dialogues with another track on which translated dialogues have been recorded in the target language” (Chaume 2012: 1). This translation mode is included in the localization of AAA video games, those with a large budget whose developer can afford a full localization including the adaptation of box and docs, in-game text and audio files and dialogues.

There are certain differences in the dubbing of a video game compared to that of a movie. In particular, the following aspects can be mentioned (Mejías-Climent 2019): There is no linear script, but numerous dialogue strings that can be grouped, depending on the character, the setting or some other criteria. Thus, there is no division in takes or loops, as it happens in countries such as Spain or Italy (*anelli*), where cinematographic scripts are traditionally divided to facilitate the dubbing actors’ task. No dubbing symbols are generally used, although they can be introduced later in the studio by the dubbing director; no TCRs are used, since there is no linear development of events in a video game. Finally, in most cases no images are used to perform the dubbing in the studio. Sound engineers use audio waves sometimes, but in the case of translators, they will never have access to images accompanying the dialogue strings.

Apart from these differences, the results in the dubbing of a modern video game and that of a movie seem to be very close—with some striking but rare exceptions, such as *Arizona Sunshine* (Vertigo Games, Jaywalkers Interactive, 2016) or *Age of Pirates* (Akella, 2006).

Dubbing is a historical and extended practice in countries such as Spain, France or Italy, among others, in which all dubbed products must meet a series of quality standards to be consumed successfully by what could be considered an ideal spectator (Chaume 2007). Among them, synchrony is one of the most prominent characteristics of dubbing.

3.1 Characteristics of cinematographic dubbing: synchronies

Dubbing synchronies represent coherence between what is heard (in this case, a soundtrack containing dubbed dialogues) and seen on the screen. As Chaume (2007: 73) points out, “respect for mouth articulation (phonetic or lip-sync), body movements (kinetic synchrony) and the duration of the translation to match the lines spoken by the screen actors (isochrony), constitute one of the cornerstones of dubbing.”

As far as traditional audiovisual products are concerned (this is, those in which there is no interaction: movies, series, TV shows, etc.), the three types of synchronization above mentioned have been studied in depth: lip-sync, kinetic synchrony and isochrony (Chaume 2012: 68). In the case of dubbing into Spanish, the implementation of the three synchronies depends on the configuration and nature of the audiovisual product and the acoustic and visual codes involved, among other aspects (such as genres and historical conventions). Especially the paralinguistic code and on- and off-sound codes (acoustic channel), and the photographic and the kinetic codes, and the types of shots (visual channel) determine to a greater extent the level of precision with which each of these three types of synchronies needs to be applied (Chaume 2004).

For example, in an extreme close-up, the character’s lip movement is clearly visible onscreen. Thus, the translator and dialogue writer need to use a similar articulation in the translation with, at least, the same number of open vowels and bilabial and labio-dental consonants (Chaume 2012: 73). However, in a long shot, there is no need to use lip-sync, and even isochrony is irrelevant, since the characters’ faces and bodies might not be visible clearly.

This happens in a linear and pre-configured audiovisual product, in which the acoustic and visual codes are determined beforehand. In a video game, however, interaction opens the visual configuration of the audiovisual product to a wider number of options; hence, synchronization does not necessarily work in the same way.

Laura Mejías-Climent

3.2 Characteristics of dubbing in video games: restrictions

Dubbing synchronies play an important role in the dubbing of a video game as well. Video games are the most complex example of a modern audiovisual product. As such, they are very close to traditional movies in some aspects, especially when it comes to cinematic scenes. However, the materials available for translation are not the same as in a movie: neither a traditional script nor a final video is facilitated (Mejías-Climent 2019). Consequently, the process takes place differently and synchronies are rather understood as a series of restrictions (Pujol Tubau 2015: 197).

These restrictions can be transmitted to translators in a maximum number of characters or words per string. They can also be indicated depending on the type of string, being dialogues and sound content more restrictive (thus, the translation needs to resemble the original length as much as possible), and in-game dialogues more flexible.

Nevertheless, many other factors come into play when determining restrictions in the dubbing of a video game: different localization vendors work differently, as shown in Mejías-Climent (2019). In addition, the different agents taking part in the whole process of dubbing play different roles when applying restrictions (thus, synchronies) to the translated text.

It is in the dubbing studio when up to five types of synchronies can be identified and applied to the translated text for the dubbing of a video game, in contrast with the three synchronies described for traditional audiovisual products (see §3.1). Dubbing actors and directors would typically apply the three types of synchronies described for traditional movies if videos were available. However, this is not usually the case, and only sound waves are available in some AAA projects, so they tend to imitate the original sound waves as much as possible, in order to assure a well synchronized dubbed dialogue. Here, up to five levels of restriction can be set, depending on the type of string they dub. It is the role of sound engineers to make the dubbed audio files resemble the original ones as much as possible, according to five levels of restriction that can be understood as the five types of synchrony used in the dubbing of a video game (Mejías-Climent 2017: 105).

Wild no time restriction applies.

Time constraint (TC) the translated utterances must be the same length as the originals, with a 10% or 20% margin.

Strict time constraint (STC) the translated utterances must be exactly the same length as the original ones, ignoring any internal pauses or specific intonation.

Sound-sync (SS) the translated utterances must be exactly the same length as the original ones, including internal pauses and intonation.

Lip-sync the translated text must be exactly the same length as the original, including pauses, and must resemble the lip articulation.

These five synchronies can be associated with game situations (Mejías-Climent 2019: 90). Game situations alternate continuously in any video game (Pujol Tubau 2015: 150). They are a direct consequence of the interactive dimension (interactive channel) and imply different conditions for interaction, depending not only on the genre, but also on the nature of every single video game.

Typically, in action-adventure video games, cinematic scenes stop interaction completely, since they are closed video clips based on a cinematographic configuration. Game action implies full interaction, the dynamic moment in which the player makes the game develop fully. Dialogues are dialectical exchanges and can be considered a situation in between cinematics and action: they can stop interaction partially, restricting the player's action to a few camera movements, for example, or not interfering with interaction at all. Finally, tasks are instructions transmitted to the player that can also take place during full interaction or stopping it completely. In the following section, the relationship between game situations and game synchronies will be described, as well as the methodology used to analyze dubbing as one of the AVT modes included in the process of localizing a graphic adventure.

4 Methodology

In a previous project (Mejías-Climent 2019), a relationship between game situations and dubbing synchronies in action-adventure video games was established. The empirical analysis showed that in the three action-adventure video games that were analyzed,

- tasks are always dubbed applying no restriction (wild sync), since they are transmitted through off-screen voices;
- game action is a relatively flexible situation, since full interaction does not always permit the highest level of visibility of the characters onscreen.

Laura Mejías-Climent

Thus, TC usually applies, with a few cases of wild sync for off-screen voices;

- dialogues are a hybrid situation in terms of interaction (they vary greatly from one video game to another). The result is that the five types of synchrony can be found in dialogues, being TC the most frequent;
- cinematics tend to resemble movies as much as possible. This is also evident in the type of synchrony most commonly used in them: lip-sync. In addition, wild sync is always used for off-screen voices.

In this case, a different video game has been analyzed to determine if this relationship between game situations and types of synchrony is also present in a game subgenre closer to traditional movies: the graphic adventure *Detroit: Become Human*. More specifically, the dubbing of the cinematic scenes in this game will be discussed in terms of lip-synching, in order to determine if there is a clear distinction between the dubbing of a movie and the dubbing of the cinematic scenes in a graphic adventure.

To do so, an empirical analysis was carried out within the framework of descriptive translation studies, following the methodology used in [Mejías-Climent \(2017; 2019\)](#). This can be labeled as an exploratory study requiring further research, since, to the best of our knowledge, no previous empirical studies on the particularities of dubbing in video games have been conducted. To reduce the scope of the study, the specific phenomenon of dubbing synchronies was analyzed in each game situation identified in said video game. More specifically, the game was played classifying in an Excel sheet every game situation that alternated throughout the story and annotating the type of synchrony identified in each situation. The game was played first in Spanish until the main goal was achieved. The same path was reproduced again in English, looking for the types of synchrony used in both the translated segments and the original ones ([Toury 1995](#)). The game was played for ten hours both in Spanish and in English. A total of 20 hours of gameplay was analyzed.

This methodology reveals some limitations that were already acknowledged in [Mejías-Climent \(2019: 315–317\)](#). First of all, the validity of the game situations taxonomy could be questioned, as dialogues represent a completely heterogeneous situation in terms of interactive options. Nonetheless, while game action, cinematics and tasks have been analyzed, the characteristics of dialogues have been redefined in the particular case of the video game presented here to suit the features of a graphic adventure. Indeed, dialogues have been considered as

dialogic quick-time events (see §5), which creates a more adapted taxonomy of game situations to conduct this empirical analysis.

Secondly, an interactive product will always entail a certain level of arbitrariness when displaying its components, which might make slight differences arise if the study was replicated. However, in terms of percentages, the impact of these changes on the results obtained would mean unrepresentative variations, so the general conclusions should not be affected.

Finally, a strong limitation imposed by empirical analyses of video games is the difficult access to the linguistic components (such as written scripts of the plot used for translation purposes), as strong non-disclosure agreements prevent the video game assets from being freely accessible or distributed. To overcome this, the corpus was analyzed while playing, since this is the only way to access the full game, and only some transcriptions of representative examples were made. The gameplay, nonetheless, was recorded and the time codes of the recording were annotated in the Excel sheet to track down all the game situations easily. Closer collaboration between the industry and academia could enrich empirical analysis like this in the future.

5 The case of *Detroit: Become Human*

The video game *Detroit: Become Human* was released in 2018 for Play Station 4, and in 2019 for PC. Its director, David Cage, is also the founder of the studio in which this game was developed, Quantic Dream, specialized in interactive storytelling.

The work by David Cage has always created controversy among the most purist players because of the high level of narrative all his games contain, at the expense of a fully interactive and ludic experience. The storytelling seems to be more important than mechanics based on a rapid response and quick reflexes from the (active) player (Altozano Dayo 2017). With his games, Cage exhorts the player to “play the story” combining continuous cinematic scenes with interactive dialogues and quick time events. Nevertheless, despite what some players claim, Cage states that he aims to achieve a full and realistic immersion rather than the narrative display (Altozano Dayo 2017).

Quick-time events (QTEs) are one of the most defining features in David Cage’s work. A QTE represents an action that is completed automatically after pressing a certain button in a limited time. They typically occur during cinematics and, if the specific button is pressed as indicated by the game, the scene continues developing successfully (Altozano Dayo 2017: 131). QTEs are a useful tool to make the

Laura Mejías-Climent

story develop combining action and cinematics (thus, interaction and movies). On the one hand, a QTE is like watching a movie, except for certain buttons that the spectator is required to press if he/she wants the story to continue. On the other hand, the most purist players also see it as an interruption of pure playability (Altozano Dayo 2017: 132).

Be it as it may, the truth is that QTEs are a recurrent tool included by Cage in all of his games to make the story develop making the player participate in an unpredictable but limited way. In this analysis, QTEs are considered as dialogues: they usually introduce a question or an answer that the player has to choose in a limited time as part of a longer dialogue. Pure QTE are not very common in *Detroit: Become Human*, neither are they considered a game situation to be analyzed, because they do not imply any sort of dubbing – in the middle of a fight, the player needs to press \times , \triangle or \circ to hit the opponent or to cover him/herself, but no linguistic content is related to the QTE in such cases (thus no dubbing sync applies).

This video game, as any other by David Cage and Quantic Dream, is intended to make the player become part of the story through very simple mechanics based on basic movements, constant dialogues and numerous cinematic scenes and dialogical QTEs (as opposed to “action” QTEs, which are not analyzed here, as mentioned). The story revolves around a dystopian Detroit City in which androids start feeling and behaving beyond what machines are expected to do, causing some divergent androids to rebel against humans and fight for their rights. The player controls three characters alternatively. The story is divided into sequences. After each one, a blank tree diagram shows the different possibilities that the player could have considered with his/her choices. The 10 hours played in each language add up a total of 35 sequences.

5.1 Game situations in Detroit

As stated in §4, the game was played for 10 hours in each language, obtaining 696 registers distributed in 35 sequences (See Table 8.1).

The most repetitive game situation, as the numbers show, are cinematic scenes, followed by QTE-dialogues, game action, and finally, tasks. This represents a clear illustration of the nature of the game: a graphic adventure focuses on the narrative content. The story is the core of the game and those game situations that carry the narrative weight are more frequent than fully interactive moments (game action). This contrasts with the number of game situations obtained in the aforementioned studies (Mejías-Climent 2017; 2019), in which action-adventure

Table 8.1: Game situations identified in 10 hours gameplay of *Detroit: Become Human*

Game situation	Num. of registers	
Tasks	58	8.33%
Game action	180	25.86%
QTE-Dialogues	212	30.46%
Cinematic scenes	246	35.34%

games were analyzed. In all of them, game action was the most repetitive situation throughout the interactive audiovisual product. We consider it *repetitive* instead of *long* because the time span has not been measured. While in a movie time codes and the duration of particular phenomena can be traced, in a video game, interactivity make some situations last as much as the player wishes (game action), while other do have a pre-established length (cinematics). This is why alternation and repetition have been analyzed instead of time span.

Regarding the types of synchrony found in each group of game situations, the following data was obtained in *DTB*:

- Tasks are transmitted through in-game (written) text exclusively. Therefore, their translation is never dubbed but always written onscreen, with a single exception: an introductory sequence in the main menu, during which a woman in a close-up shot talks directly to the player to guide him/her through the main settings of the game, before the actual story begins. This single task is dubbed lip-synching.
- Game action: most of the game action is dubbed using wild sync (42 cases both in Spanish and in English). STC and TC are also frequent (31 and 22 cases respectively for the Spanish version, and 31 and 21 in English). Lip-sync is only used in 2 cases in Spanish, but in 6 cases in English. There are 80 moments in which no dialogues are heard during game action.
- QTE-Dialogues: with a few exceptions, most dialogical QTEs are dubbed using lip-sync, thus resembling the dubbing used in traditional movies. Only 1 example of TC, 4 of STC and 2 examples of SS have been found both in Spanish and English. There are 5 moments using off-voices (wild sync).

Laura Mejías-Climent

- Cinematics: as well as dialogical QTEs, cinematic scenes are dubbed using lip-sync almost exclusively (221 cases in Spanish; 222 in English), with only 1 example of wild, TC, STC and 5 of SS in Spanish; and 1 example of wild, TC, STC and only 4 of SS in English.

These results illustrate the nature of the subgenre of a graphic adventure, at least, in the case of *DBH*: dialogues using QTEs and cinematic scenes are the most repetitive game situations in a video game aimed at recreating the atmosphere of a movie, but always reminding the player that he/she is the main character in charge of taking all the decisions to make the story develop. Additionally, the use of lip-sync as the most frequent type of synchrony in the dubbing of such a game – 424 cases in Spanish and 429 in English out of 696 registers – shows the closeness of the final configuration of dubbing in a graphic adventure and in a traditional movie, in terms of the types of synchrony applied. As in any non-interactive audiovisual material, wild sync is also frequent—54 registers both in Spanish and English, because off-voices are used repeatedly to make the story develop.

A significant difference with non-interactive movies, however, is the constant use of text written on the screen, especially, to give instructions to the player, replacing dubbing. A total of 148 out of 696 cases, both in Spanish and English, have been registered as containing no dialogues; 80 took place during game action, 11 represented silent cinematic scenes and 58 were tasks in which the instructions were transmitted through in-game text. This in-game instructional text will never be used in a non-interactive movie, as no action is required from the viewer.

5.2 The dubbing of cinematic scenes

As described in §5.1, the highest level of restriction, lip-sync, is the most frequently applied in the dubbing of the graphic adventure *DBH*. This is the most complex type of synchrony, since the reproduction of the articulatory movements in the translated text is not always compatible with an accurate translation. In movies, it is reserved almost exclusively for close-ups and extreme close-ups (Chaume 2012).

In the case of *DBH*, the most restrictive type of synchrony is noticeable in almost all the cinematics and the dialogical QTEs. Here, a strong similarity with the dubbing of any non-interactive movie that should be noted is that the characters' lips in *DBH* are animated in English using the articulatory lip movements of

the actors dubbing the original version. In a movie, we can see real actors speaking onscreen, whose utterances are replaced later by the translated utterances performed by the dubbing actors in the target language. In a video game, it has been argued (Méndez González 2015: 106) that the “original version” might not be strictly the first one, since all video game characters need a human actor/actress to provide them with his/her voice. Nonetheless, in AAA modern video games such as *DBH*, characters’ movements and even speech articulation are recreated with the motion capture technique (Turnes 2020), using sensors on a human actor’s body to capture his/her movements and recreate them in a digital model to animate a virtual character (Kines 2000). In such a way, the characters’ lips in *DBH* accurately reproduce the articulatory movements of the words they utter in English, as real actors in a non-interactive movie do. In this case, the slight differences between the exact lip movements of the original version in English and the lip-synched Spanish sentences are visible in *DBH* as it happens in any movie.

Regarding the level of accuracy in lip-synching, it should be noted that, as explained above, in most cases dubbing actors do not have images to support their performances. This might result in a slightly less accurate lip-synching than in a traditional movie that, nonetheless, remains unnoticed by most spectators, although further reception studies in the field of video game dubbing would be quite revealing. In the case of *DBH*, lip-synchrony appears quite accurate, even though some open vowels, and bilabial and labio-dental consonants do not coincide exactly in some cases.

To determine how accurate lip-sync is achieved in *DTB*, the following examples reproduce the dubbed and original dialogues in a cinematic scene, a dialogical QTE and a second cinematic scene dubbed using lip-sync. Those paralinguistic features included in the acting (coughing, onomatopoeic expressions, etc.) have been reproduced with the dubbing symbol (G) used in Spanish dubbing scripts. Pauses are represented with “/” and whenever a non-dialogical QTE is required to make the action continue, the symbol (QTE) has been introduced.

(1) Cinematic scene 298

Connor:	¿Teniente? / [QTE]	Lieutenant? / [QTE]
Anderson:	(G)	(G)
Connor:	Despierte, teniente.	Wake up, lieutenant!
Anderson:	(G)	(G)
	[QTE]	[QTE]
Connor:	¡Soy yo, Connor!	It’s me, Connor!

Laura Mejías-Climent

- Anderson: (G) (G)
 Connor: Por su bien, haré que vuelva I'm going to sober you up,
 a estar sobrio. Se lo advierto: for your own safety. I have
 será poco agradable. to warn you, this may be
 unpleasant.
 Anderson: (A LA VEZ) (G) ¡Déjame en (SIMULTANEOUSLY) Hey!
 paz, puto androide! ¡Sal de Leave me alone, you fuckin'
 mi casa, joder! android! Get the fuck outta
 my house!
 Connor: Lo siento, teniente, pero I'm sorry, lieutenant, but
 lo necesito. [QTE] Le I need you. [QTE] Thank
 agradezco de antemano su you in advance for your
 colaboración. cooperation.
 Anderson: (P) Eh, ¡lárgate de aquí, (P) Hey! Get the fuck outta
 joder! [QTE] here! [QTE]

(2) QTE-Dialogue 302

- Connor: Lo entiendo... Tampoco creo I understand... It probably
 que tuviera mucho interés... wasn't interesting anyway...
 Han encontrado el cadáver A man found dead in a
 de un hombre en un burdel sex club downtown... Guess
 del centro... Ya resolverán el they'll have to solve the case
 caso sin nosotros... without us...
 Anderson: Oye, no me vendrá mal You know, probably
 tomar un poco el aire. En el wouldn't do me any
 armario de la habitación hay harm to get some air...
 ropa. There're some clothes in
 the bedroom there.
 Connor: Iré a por ella. I'll go get them.

(3) Cinematic scene 305

- Connor: ¿Qué se quiere poner? What do you want to wear?
 Anderson: (G) (Tose) (G) (Coughing)
 Connor: ¿Se encuentra bien, Are you alright, lieutenant?
 teniente?
 Anderson: (G) Sí... sí... (G) De (G) Yeah... yeah... (G) Won-
 maravilla... solo... dame derful... Just a... Give me
 cinco minutos, ¿vale? five minutes, okay?
 Connor: Claro. Sure.
 Anderson: (G) (G)

In the three examples above, all the elements included in video game lip-synching are reproduced precisely, as in the case of a non-interactive movie: the length of the translated utterances is the same as the original ones; intonation and paralinguistic features are reproduced exactly too. Also open vowels (*a*, *e*, *o*, marked in red) and most labial and bi-labial consonants (marked in blue) are reproduced as much as possible. This makes this level of restriction quite close to how it is done in a non-interactive movie, even though translators and most probably dubbing actors did not have access to the final videos. Most times, videos have not been produced yet when dubbing takes place, although this could not be confirmed in the particular case of *DBH*.

The use of the highest level of restriction in the translated text of a video game implies that all previous levels have been taken into account as well: length of the utterances (TC, STC) and intonation and pauses within utterances (SS). Thus, isochrony is applied, as it is always the case in a non-interactive movie, being this the most valued quality standard among Spanish spectators (Chaume 2007).

Some cases of cinematographic kinetic synchrony are also found in the dubbing of *DBH*. This is most probably not done intentionally in the dubbing studio, since videos are not available. Some deictic expressions have been either omitted, as in (2), or reproduced literally, in the same position within the sentence, to assure the correspondence with the image.

In (2), *There're some clothes in the bedroom there* has been translated as '*En el armario de la habitación hay ropa*'. The second occurrence of *there* has been omitted, but the reference to the bedroom should be enough to make the possible body movements of the character coherent with a reference to it, wherever it is located in the house.

Another example was found in the game situation 23 (see example (4)), a limited game action during which the player's actions were restricted to camera movements. The seller refers to the android in front of him with the following utterance:

- (4) Game action 23

Seller: En este momento tenemos una oferta especial para esta gama de 7999 dólares, con financiación sin intereses a 48 meses.

At the moment we're doing a special promotion on this entire range at \$7999, with a 48-months interest free credit.

The deictic *this* coincides with the character onscreen pointing at the android in front of him. Since it has been translated reproducing the same length and structure of the sentence, kinetic synchrony has been applied here as well.

Laura Mejías-Climent

These are just a few examples that illustrate the accuracy with which lip-sync has been applied in the analyzed video game, similar to what Chaume (2016) did in a descriptive and “qualitative analysis according to the episode’s adherence to a checklist of dubbing standards” in the TV series *The West Wing* (1999) (Chaume 2016). His work, focusing on quality standards in dubbing (Chaume 2007), as well as some other studies on the language of dubbing such as those by Piñero (2009); Ferriol (2013); Freddi & Pavesi (2009) and Sierra (2008), among many others, could serve as useful models for further research in the field of video game dubbing, compared to dubbing in non-interactive material.

Empirical and reception studies would especially benefit greatly from this pilot analysis on the characteristics of dubbing synchronies in video games, which represents a revealing starting point to trace considerable similarities in the dubbing of interactive and non-interactive material, in spite of the different materials available when translating, and the differences in the translation and localization processes.

6 Final remarks

The main aim of this chapter was to review the most prominent similarities and differences between cinematographic dubbing and that of a video game belonging to the subgenre of graphic adventures, within the main genre of adventure video games.

Video games represent the most complex example of a multimodal text and, as such, they share many of the characteristics of a non-interactive audiovisual product. The particularity of the added interactive dimension in video games makes them a specific case, which undergoes a process of localization when exported to other cultures, beyond a mere linguistic transfer.

Within the industry, professionals support the idea of localization being an independent field, different from AVT, since it implies some other adaptation processes including a more creative approach for translators and the modification of linguistic and non-linguistic contents of the localized product. Within academia, however, there is no clear evidence that justifies the need to separate localization from AVT. Both fields deal with the translation of multimedia and multimodal products, and both encompass a series of translation modes such as dubbing or subtitling, among many others, required for the different types of audiovisual products and their particularities.

Further research is needed on the convergences and differences of the process of dubbing a movie and a video game, and on localization and AVT modes in

general. Notwithstanding this, the aim of these pages was to point out that the result in the dubbing of a graphic adventure and a traditional movie is very similar in terms of dubbing synchronies.

Three types of synchronies have been described in the dubbing of non-interactive audiovisual products (Chaume 2004; 2007; 2012). A new taxonomy of five dubbing synchronies is also necessary in video games (Mejías-Climent 2017; 2019), given the idiosyncrasy of the interactive audiovisual text. However, in the specific case of cinematic scenes and dialogical QTEs of a graphic adventure such as *Detroit: Become Human*, the most restrictive dubbing synchrony in video games, lip-sync, is the one used most frequently and encompasses all the three dubbing synchronies described for non-interactive products.

It should be acknowledged that a case study will never be enough to identify translation tendencies (Toury 1995), neither to close the debate of the link between AVT and localization. Nonetheless, this study can be understood as no more than a starting point after which further research is needed, first, enlarging the corpus of adventure video games. Other game genres would need to be explored as well, to look for translation tendencies in the use of dubbing synchronies, and to determine the most characteristic game situations for each genre. Audiovisual translation and localization are two convergent fields and professional practices, which seem to share more similarities than differences, although much is yet to be done with further research.

Abbreviations

AVT	Audiovisual Translation
DBH	<i>Detroit: Become Human</i>
QTE	Quick-Time Event
TC	Time constraint
STC	Strict time constraint
SS	Sound-sync

References

- Altozano Dayo, José. 2017. *El videojuego a través de David cage*. Sevilla: Héroes de Papel.
- Bernal-Merino, Miguel Ángel. 2006. On the translation of video games. *JoSTrans: The Journal of Specialised Translation* 6. 22–36.
- Bernal-Merino, Miguel Ángel. 2015. *Translation and localization in video games: Making entertainment software global*. London: Routledge.

Laura Mejías-Climent

- Cadieux, Pierre & Bert Esselink. 2004. GILT: Globalization, internationalization, localization, translation. *Globalization Insider* 11(1.5). 1–5.
- Chaume, Frederic. 2004. *Cine y traducción*. Madrid: Cátedra.
- Chaume, Frederic. 2007. Quality standards in dubbing: A proposal. *Tradterm* 13. 71–89.
- Chaume, Frederic. 2012. La traducción audiovisual: Nuevas tecnologías, nuevas audiencias. In Cristina Bosisio & Stefania Cavagnoli (eds.), *12° congresso dell’associazione Italiana di linguistica applicata*, 143–159. Perugia: Guerra Edizioni.
- Chaume, Frederic. 2013. The turn of audiovisual translation: New audiences and new technologies. *Translation Spaces* 2. 107–125.
- Chaume, Frederic. 2016. Dubbing a TV drama series: The case of *The West Wing*. *inTRAlinea* 18. 1–12.
- Chaume, Frederic. 2018a. An overview of audiovisual translation: Four methodological turns in a mature discipline. *Journal of Audiovisual Translation* 1(1). 40–63.
- Chaume, Frederic. 2018b. Is audiovisual translation putting the concept of translation up against the ropes? *JoSTrans: The Journal of Specialised Translation* 30. 84–104.
- Esselink, Bert. 2000. *A practical guide to localization*. Amsterdam: John Benjamins.
- Ferriol, José Luis Martí. 2013. *El método de traducción*. Castelló de la Plana: Universitat Jaume I.
- Freddi, Maria & Maria Pavesi. 2009. *Analysing audiovisual dialogue: Linguistic and translational insights*. Bolonia: Clueb.
- Gambier, Yves. 2003. Screen transadaptation: Perception and reception. *The Translator* 9(2). 171–189.
- Granell, Ximo, Carme Mangiron & Núria Vidal. 2015. *La traducción de videojuegos*. Sevilla: Bienza.
- Hurtado Albir, Amparo. 2011. *Traducción y traductología: Introducción a la traductología*. Madrid: Cátedra.
- Jiménez-Crespo, Miguel. 2013. *Translation and web localization*. London: Routledge.
- Kines, Melianthe. 2000. *Planning and directing motion capture for games*. https://www.gamasutra.com/view/feature/131827/planning_and_directing_motion_.php (6 July, 2020).
- Mangiron, Carme & Minako O’Hagan. 2006. Game localization: Unleashing imagination with “restricted” translation. *JoSTrans: The Journal of Specialised Translation* 6. 10–21.

- Mata Pastor, Manuel. 2005. Localización y traducción de contenido web. In Detlef Reineke (ed.), *Traducción y localización: Mercado, gestión y tecnologías*, 187–252. Las Palmas de Gran Canaria: Anroart.
- Maxwell-Chandler, Heather. 2005. *The game localization handbook*. Massachusetts: Charles River Media.
- Mejías-Climent, Laura. 2017. Multimodality and dubbing in video games: A research approach. *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 17. 99–113.
- Mejías-Climent, Laura. 2019. *La sincronización en el doblaje de videojuegos: Análisis empírico y descriptivo de los videojuegos de acción-aventura*. Universitat Jaume I. (Doctoral dissertation).
- Méndez González, Ramón. 2015. *Localización de videojuegos: Fundamentos traductológicos innovadores para nuevas prácticas profesionales*. Vigo: Servizo de Publicacións Universidade de Vigo.
- Muñoz Sánchez, Pablo. 2017. *Localización de videojuegos*. Madrid: Síntesis.
- Neves, Joselia. 2005. *Audiovisual translation: Subtitling for the deaf and hard-of-hearing*. University of Roehampton. (Doctoral dissertation).
- O'Hagan, Minako & Carme Mangiron. 2013. *Game localization: Translating for the global digital entertainment industry*. Amsterdam: John Benjamins.
- Piñero, Rocío Baños. 2009. *La oralidad prefabricada en la traducción para el doblaje: Estudio descriptivo-contrastivo del español de dos comedias de situación: Siete vidas y Friends*. Universidad de Granada. (Doctoral dissertation).
- Pruys, Guido Marc. 2009. *Die Rhetorik der Filmsynchronisation*. Köln: Eigenverlag.
- Pujol Tubau, Miquel. 2015. *La representació de personatges a través del doblatge en narratives transmèdia: Estudi descriptiu de pel·lícules i videojocs basats en El senyor dels anells*. Universitat de Vic. (Doctoral dissertation).
- Pym, Anthony. 2016. *Teorías contemporáneas de la traducción: Materiales para un curso universitario*. 2nd edn. Tarragona: Intercultural Studies Group.
- Sierra, Juan José Martínez. 2008. *Humor y traducción: Los simpson cruzan la frontera*. Castelló de la Plana: Universitat Jaume I.
- Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.
- Turnes, Yova. 2020. *GamerDic*. <http://www.gamerdic.es/web/sobre-la-web/> (6 July, 2020).
- Vázquez Rodríguez, Arturo. 2018. *El error de traducción en la localización de videojuegos: Estudio descriptivo y comparativo entre videojuegos indie y no indie*. Universitat de València. (Doctoral dissertation).

Chapter 9

Analysing the Dimension of Mode in Translation

Ekaterina Lapshinova-Koltunski

Universität des Saarlandes

The present chapter applies text classification to test how well we can distinguish between texts along two dimensions: a text-production dimension that distinguishes between translations and non-translations (where translations also include interpreted texts); and a mode dimension that distinguishes between spoken and written texts. The chapter also aims to investigate the relationship between these two dimensions. Moreover, it investigates whether the same linguistic features that are derived from variational linguistics contribute to the prediction of mode in both translations and non-translations. The distributional information about these features was used to statistically model variation along the two dimensions. The results show that the same feature set can be used to automatically differentiate translations from non-translations, as well as spoken texts from the written texts. However, language variation along the dimension of mode is stronger than that along the dimension of text production, as classification into spoken and written texts delivers better results. Besides that, linguistic features that contribute to the distinction between spoken and written mode are similar in both translated and non-translated language.

1 Introduction

In the present contribution, we analyse translation as a product which possesses a number of linguistic characteristics expressed in its linguistic features. These linguistic features make translation look different from other language products. Translation variation is influenced by various dimensions, i.e. language, register, text production or expertise (Lapshinova-Koltunski 2017). They are related



to the constraint dimensions as defined by Kotze (2019: p. 346) who sees translation as one of constrained language varieties. These varieties are probabilistically conditioned by various interacting dimensions that allow for modelling their variation. Mode¹ and text production are amongst the five dimensions described by Kotze (2019: p. 346). The focus of this paper is on the variation in English-to-German translation that involves the dimension of mode, i.e. variation between spoken and written language production. We believe that such variation manifested by the linguistic features of written and spoken translations (also referred to as 'translations' and 'interpretations' respectively), e.g. preferences for modality meanings, proportion of nominal or verbal phrases and others. These features should allow us to analyse and model the dimensions involved. Methodologically, we focus on quantitative distributions of these linguistic features reflected in the lexico-grammar of texts.

In the following, we analyse language variation in translation products that include both translations and interpretations. Our focus is on **mode** in translation product – differences between English-to-German translations vs. interpretations. We also analyse differences between translated and non-translated texts in German. These differences correspond to the variation along the dimension of **text production**. Based on existing studies in the area of translationese, interpretese and variational linguistics (see §2) we expect that the variation along the dimension of mode should be stronger than that along the dimension of text production. We are interested in the linguistic features that contribute to the distinction between spoken and written mode in both translated and non-translated language.

The remainder of the paper is organised as follows: §2 provides an overview of the related work and theoretical background; we give details on the data and methods used in our analyses in §3; §4 is dedicated to our results and analysis; we conclude and point to some issues for discussion and future work in §5.

2 Related work and theoretical background

2.1 Translationese

We rely on the studies on translationese (Baker 1993; Toury 1995; Bernardini & Ferraresi 2011; Teich 2003; among numerous others) showing that translated texts have certain linguistic characteristics in common which differentiate them from original, non-translated texts. These differences, however, do not point to

¹Kotze (2019: p. 346) calls this constraint 'Modality and register'

the quality of the texts, as claimed by Gellerstam (1986) and empirically shown by Kunilovskaya & Lapshinova-Koltunski (2019). Translationese is rather a statistical phenomenon, and the differences are reflected in the distribution of lexicogrammatical, morpho-syntactic and textual language patterns that can be organised in terms of more abstract categories often called features of translations. These include *simplification* (Toury 1995), *explicitation* (Olohan & Baker 2000; Øverås 1998), *normalisation* and *shining-through* (Bernardini & Ferraresi 2011; Teich 2003; Scott 1998) and *convergence* (Laviosa 2002). Since the differences between translated and non-translated texts are of statistical character, they can be uncovered automatically. Recent studies on translationese employ automatic detection techniques using various feature constellations. One of the first works in this area is (Baroni & Bernardini 2006). They use n-grams for wordforms, lemmas and parts-of-speech (POS) which represent lexical and grammatical features in a supervised scenario² to differentiate between translated and non-translated texts. Ilisei et al. (2010) use a number of simplification-related features to succeed in differentiating between translated and non-translated texts with machine-learning algorithms. A number of translationese indicators have been applied in an unsupervised approach to automatic classification between translations and originals by Volansky et al. (2015). Linguistically interpretable features were used by Kunilovskaya & Lapshinova-Koltunski (2020), who automatically differentiate translated Russian and German from originals in both languages. They proceed bottom-up in their feature definition and try to identify translationese effects based on the results of corpus analysis.

2.2 Variational linguistics

We refer to studies in variational linguistics, such as Systemic Functional Linguistics (SFL, Halliday 2004; Halliday & Matthiessen 2014) and genre or register studies (Biber 1995; Neumann 2013). Following these studies, language varies according to the context of use. To account for a functional organisation of language, the framework offers contextual configurations, i.e. three variables characterising the level of context: *Field*, *Tenor* and *Mode* of discourse³. These variables

²Supervised machine learning, also referred to as text classification, is an approach for discovering groupings in multivariate data sets. In a supervised scenario, we know what groupings we can expect in the data, and the question is whether the data under analysis support these groupings (see Baayen 2008: p. 118). In an unsupervised scenario, we do not know what groupings exist in the data and an algorithm tries to identify any groupings by extracting features and patterns on its own.

³Note that Mode of discourse does not correspond to mode of production that we use to differentiate between spoken and written text'. Mode of discourse is related the role of the language

Ekaterina Lapshinova-Koltunski

correspond to sets of specific lexico-grammatical features. Field of discourse is realised in term patterns or functional verb classes. Tenor of discourse is realised in stance used by speakers or modality expressed by modal verbs. Mode of discourse relates to Theme-Rheme structure and cohesive relations. Linguistic features inspired by SFL and genre or register studies have been used in the analysis of contextual variation of translated texts. For instance, [Evert & Neumann \(2017\)](#) apply them for intralingual and cross-lingual variation in both translated and non-translated text. [Lapshinova-Koltunski & Martínez Martínez \(2017\)](#) use various categories of cohesive relations (related to the parameter of Mode) to automatically differentiate between spoken and written texts in English and German. This is one of the few works known to us that analyse differences between spoken and written texts with machine learning techniques. The authors succeed in automatic identification of the dimension of mode in non-translated texts.

2.3 Interpretese

In terms of the dimension of mode in translation, there are fewer studies on interpretese ([Kajzer-Wietrzny 2012](#); [Defrancq et al. 2015](#); [He et al. 2016](#); [Bernardini et al. 2016](#); [Ferraresi & Miličević 2017](#); [Dayter 2018](#); [Bizzoni & Teich 2019](#)). They show that interpreted texts possess linguistic features that differentiate them not only from translated texts but also from other language products. In our work, we aim to analyse the differences not only between interpreted and translated texts, but also between interpreted, non-interpreted and non-translated texts. With this goal in mind, we follow work by [Shlesinger & Ordan \(2012\)](#) who claim that modality (corresponding to our notion of mode dimension) exerts a stronger effect than ontology (corresponding to our notion of text production). This means that the dimension of mode (i.e. whether a text is spoken or written) has more influence than the dimension of text production (whether a text is a translation or an original).

3 Methodology

3.1 Hypotheses and research questions

Our research questions are based on our assumptions given in §1 above.

in the interaction.

Research Question 1 (RQ1) First of all, we are interested in language variation along the dimension of text production. So, we would like to find out if we can automatically differentiate between translations and non-translations independently of the mode production (whether spoken or written).

Research Question 2 (RQ2) This research question is related to the analysis of mode in translation – differences written and spoken translations. We aim to find out if we can automatically detect mode in translation.

Research Question 3 (RQ3) We would like to know if it is easier to automatically detect text production (differentiate between translations and non-translation) or mode (spoken and written translation) with an assumption that variation along the dimension of mode is stronger than that along the dimension of text production.

Research Question 4 (RQ4) We are also interested in the linguistic features that contribute to the distinction between spoken and written mode. Specifically, we want to find out if the same features are responsible for the variation between translations and interpretations and between written and spoken texts.

3.2 Corpus resources

For our analyses, we use written and spoken data that is derived from the European Parliament, so that all the subcorpora belong to the same register. We include transcribed interpretations and translations (spoken and written translations) from English into German (INTER and TRANS) and comparable German originals – transcriptions of European Parliament speeches by native speakers of German and published written speeches in German (GO-SP and GO-WR). The spoken part (transcribed speeches in German and interpretations from English into German) is taken from EPIC-UdS (Karakanta et al. 2019), whereas the written part (published written speeches in German and official translations from English into German) is taken from Europarl-UdS (Karakanta et al. 2018). We provide details on the size of the subcorpora in terms of number of texts (txt), sentences (sent) and tokens (token) in Table 9.1. As seen in the table, the ‘spoken’ subcorpora are much smaller.

All the texts in the subcorpora at hand were automatically annotated with information on token, lemma and part-of-speech based on the Universal Dependency framework (Nivre et al. 2019; Straka & Straková 2017). They are encoded in

Table 9.1: Size of the subcorpora under analysis

subcorpus	txt	sent	token
TRANS	575	169,016	3,994,177
INTER	137	3,409	61,631
GO-SP	165	4,076	59,896
GO-WR	1072	427,775	8,954,768
TOTAL	1,949	604,276	13,070,472

CWB and can be queried with the help of Corpus Query Processor (CQP, [Evert & Team 2019](#)), which is a part of the Corpus Workbench (CWB, [Evert & Hardie 2011](#)). They are also available in CQPWeb⁴ supported by CLARIN-D. These annotations facilitated extraction of features for our analysis as described in §3.3. The accuracy of our feature extraction is thus dependent on the accuracy of the automatic annotation. The respective model performance is 99.9% for words, 80.9% for sentence borders, 91.7% for universal parts-of-speech and 95.4% for lemmas⁵.

3.3 Features

In our approach, we use a set of features derived from variational linguistics (see §2.2 above). As already mentioned above, the frameworks offer three context parameters for language variation that correspond to various lexico- grammatical patterns. Table 9.2 illustrates the features used, as well as language patterns they represent within a text. The first column in the table contains the corresponding contextual parameter of variation, the second column includes examples of features formulated in abstract categories, and the third column shows examples of language patterns serving as operationalisations for the features.

Overall, we use 17 lexico-grammatical patterns. We include four patterns related to the Field of discourse (see Table 9.2). They are associated with the abstract categories of processes and participants and are linguistically realised in nouns and verbs, the distribution of content words and also *ung*-nominalisations. The next four patterns are included within the parameter of Tenor. They are related to roles and attitudes of participants, and are realised linguistically in modality expressed by modal verbs such as *can*, *may*, *must* that we group according to their meanings (3 patterns). Tenor is also related to evaluation used

⁴<http://corpora.clarin-d.uni-saarland.de/cqpweb>

⁵See http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_20_models for details.

Table 9.2: Features under analysis

parameter	feature	language pattern
FIELD	participants and processes	nominal and verbal parts-of-speech, content words, <i>ung</i> -nominalisations
	modality	modal meanings: obligation, permission, volition
TENOR	evaluation	evaluative patterns (<i>more importantly/ it is important to say</i>)
MODE	textual cohesion	personal and demonstrative pronouns; general nouns (<i>fact, plan</i>); conjunctions; logico-semantic relations: additive, adversative, causal, temporal, modal

by speakers to convey personal attitude to the given information, e.g. evaluative patterns like *very important, it is important to say*. They represent the fourth patterns in this parameter. The final nine patterns are related to Mode of discourse, i.e. the role and function of language in a particular situation, the symbolic organisation of a text. They are realised as cohesive relations at the textual level, for instance coreference via pronouns (2 patterns) or general nouns (1 pattern), distribution of conjunctions (1 pattern) or discourse relations via conjunction (5 patterns). All these features were used in previous works on translationese (see e.g. [Lapshinova-Koltunski 2019; 2017](#)).

The frequencies of these features are automatically extracted from corpora. We use the functionality of Corpus Query Processor mentioned above. This query tool allows definition of language patterns in the form of complex regular expressions based on string and part-of-speech restrictions. The query tool delivers text instances along with their frequencies in the texts and subcorpora in which they occur. The extracted distributional information is saved in matrices for further use for statistical analysis.

3.4 Methods

We use Weka ([Witten et al. 2011](#)), an open source tool for statistical analysis and visualisation for our analyses. To answer the first two research questions (RQ1 and RQ2), we apply text classification using Support Vector Machines (SVM, [Vapnik & Chervonenkis 1974; Joachims 1998](#)) with a linear kernel. Classification with

Ekaterina Lapshinova-Koltunski

SVM is a supervised scenario in machine learning. We label our data with the information on classes represented in our case by text production (translations vs. non-translations) and mode (spoken vs. written), collect the information on the language patterns outlined in §3.3 from the corpora described in §3.2, and see if our corpus data support the predefined classes.

We apply separate binary classification tasks for text production (RQ1) and mode (RQ2). The result of a linear SVM is a hyperplane (a line separating two classes) that separates the classes as best as possible, and allows a clear interpretation of the results. The classes defined in this study include translations and non-translations in the first classification task, and spoken and written modes in the second. The performance scores of classifiers are judged in terms of precision, recall and F-measure. They are class-specific and indicate the results of automatic assignment of class labels to certain texts.

To answer the third research question (RQ3), we compare the scores resulting from the two classifications in RQ1 and RQ2. If the scores are higher in the second classification task, variation along the dimension of mode is stronger than the variation along the dimension of text production (in line with our assumption).

We use methods of feature selection to answer the fourth question (RQ4). Attribute selection derived from machine learning is used to automatically select attributes (the language patterns we use) that are most relevant to the predictive modeling problem (prediction of a class membership). In the data, there is always a mixture of attributes with some being more relevant for making predictions and the others being less relevant. The process of selecting attributes in the data helps to reduce their number to those relevant for the specific prediction task. The attribute evaluator is the technique by which each attribute in the dataset is evaluated in the context of the output class (mode in our case). We use the best-first strategy (the best attribute is added at each round) which uses an iterative algorithm (starts with an arbitrary solution to a problem and attempts to find a better solution at every step). This is a correlation-based technique that evaluates the value of a subset of attributes by considering the individual predictive ability of each of these attributes along with the degree of redundancy between them. Subsets of language patterns that are highly correlated with the class (mode) and at the same time have low intercorrelation are preferred over others (see [Hall 1998](#): for more details). We then compare the lists of resulting language patterns for the class in the two data subsets (mode in non-translation and mode in translation). Our assumption is that if there are any overlaps in the lists, this would indicate that the same/similar linguistic features are responsible for the prediction of mode in both translated and non-translated texts.

4 Results

4.1 RQ1

In the RQ1 analyses, we define three classification tasks. All classes are defined on the basis of the two text production types under analysis – translations/ interpretations and non-translations/-interpretations in German. In the first classification task, we automatically separate written translations (TRANS) from written non-translations (GO-WR). In the second classification task, we automatically separate spoken translations (INTER) from spoken originals (GO-SP). And finally, in the third classification task, we do not differentiate the mode but attempt to assign both translations and interpretations one class (TRANS+INTER) and both written and spoken originals (GO-WR+GO-SP) – the other. The performance of the classifier that automatically separates two juxtaposed classes is evaluated with a 10-fold cross-validation step. We judge the performance scores in terms of precision, recall and F-measure. These scores are specific for each class (text production type) and indicate the results of automatic assignment of production type labels to certain texts in our data. In the case of precision, we measure how many cases in the data correspond with the positive labels given by the classifier. For example, there are 137 spoken translations in our data. If the classifier assigns INTER labels to 137 texts, and all of them really belong to the subcorpus of spoken translations, then we will achieve a precision of 100%. If 37 texts turned out to be non-translations in German and were wrongly classified into the INTER class, we would have a precision of 73% only. With recall, we measure if all translated texts were actually assigned to the INTER class. So, if we have 137 translated texts, we would have the highest recall if all of them are assigned the INTER label. If only 100 out of 137 available in the data were assigned to the INTER class (and the rest to the GO-SP class), we would have a recall score of 73% of. F-measure combines both precision and recall and is given by their harmonic mean. The results of the classification performance (in terms of precision, recall and F-measure) are presented in Tables 9.3 – 9.5 below. Figure 9.1 provides bar plots of the weighted average of the F-Measure for the three classification tasks.

Overall, we achieve an accuracy of 86.5%, with an average F-measure of 85.7% in the first classification task for translations and non-translations. As seen from Table 9.3, non-translated texts are better identified by the classifier than the translated ones (F-measure of 90.5% vs. 76.6%). However, translations are identified with better precision (97.3% vs. 83.4%), whereas the texts originally written in German achieve higher recall (99.1% vs. 63.1%). This means that more texts in the dataset were labelled by the model as originals, with more translations being wrongly recognised as originals than originals being wrongly recognised as

Table 9.3: Classification results for the first text production type distinction in %

	Precision	Recall	F-Measure
TRANS	97.3	63.1	76.6
GO-WR	83.4	99.1	90.5
Weighted average	88.2	86.5	85.7

translations. We look into the confusion matrix available in the Weka output to see that 212 translations (out of 575) were labeled as non-translations. At the same time, only 10 non-translations were wrongly assigned the TRANS label.

Table 9.4: Classification results for the second text production type distinction in %

	Precision	Recall	F-Measure
INTER	77.8	61.3	68.6
GO-SP	72.7	85.5	78.6
Weighted average	75.0	74.5	74.0

In the second text production classification task (interpretations vs. speeches originally produced in German), we achieve an accuracy of 74.5% with the average F-measure of 74.0%, pointing to the fact that text production distinction in the spoken texts is harder to make than in the written ones in the dataset at hand. The scores in Table 9.4 reveal that again, translations are recognised with better precision than non-translations (77.8% vs. 72.7%), but recall is higher for non-translations: 85.5% vs. 61.3%. This means that more texts in the dataset were recognised by the model as original speeches, and thus, more translations were wrongly recognised as originals than originals were wrongly recognised as translations, as we also observed in the first case.

If we combine the spoken and the written data to differentiate between translations and non-translations, we achieve an accuracy of 80.04% and an average F-measure of 77.7% (see Table 9.5). These overall scores are lower than in the first task (distinction between written translations and non-translations) and higher than in the second one (distinction between spoken translations and non-translations), which was foreseeable as the data in the this task is a mixture of

Table 9.5: Classification results for the third text production type distinction in %

	Precision	Recall	F-Measure
TRANS+INTER	98.5	46.1	62.8
GO-WR+GO-SP	76.2	99.6	86.4
Weighted average	84.4	80.0	77.7

the first two. However, we observe an overall increase in the precision for translations along with an overall increase of recall for non-translations. The analysis of the confusion matrix shows that only five non-translations were erroneously classified into the class of translations, whereas more than a half of translations (53.9%) were erroneously labelled as non-translations. In other words, the translations in our data seem to be readily but inappropriately recognised as non-translations by the non-translation class, whereas non-translations are not accepted as translations by the modelled translation class. This indicates that non-translations represent a more diverse class, displaying more variation than translated texts, with the latter being a subset of non-translations in terms of the features underlying classification. In terms of translationese, this points to convergence of translations.

The results in Table 9.5 suggest that it is easier to model non-translated texts regardless of the mode they belong to (F-measure of 86.4%). Although written originals achieve the best result (F-measure of 90.5%), mixing them with spoken non-translations (whose F-measure equals 78.6%) results in a drop of 4.1% against the result for the written mode and an increase of 7.8% against the spoken mode. For translations, mixing both modes results in an F-measure of 62.8% with a drop of 13.8% against the written mode (76.6%) and a drop of 5.8% against the spoken mode (68.6%).

The results of the three classifications suggest that we can automatically tease apart translations from non-translations regardless of the mode production. At the same time, the task is easier, when only written texts are involved.

4.2 RQ2

We perform the same analysis steps for the differentiation between spoken and written modes as we did for translation and non-translation in §4.1. We again decide for a three-fold task in the mode analysis: (1) classification of non-translated

Ekaterina Lapshinova-Koltunski

spoken and written texts (GO-SP vs. GO-WR); (2) classification of translated spoken and written texts (INTER vs. TRANS) and (3) classification of spoken and written texts with both translations and non-translations taken together (GO-SP+INTER vs. GO-WR+TRANS). In the last task, we do not sort texts according to the text production type, defining the task as one of finding the overall variation along the dimension of mode.

Table 9.6: Classification results for the first mode distinction in %

	Precision	Recall	F-Measure
GO-SP	80.5	100.0	89.2
GO-WR	100.0	96.3	98.1
Weighted average	97.4	96.8	96.9

We achieve an overall accuracy of 96.77% with an average F-measure of 96.9% in the classification into spoken and written non-translations (see Table 9.6). Interestingly, written texts are better classified than the spoken ones (98.1% vs. 89.2% of F-measure). At the same time, we observe asymmetries in precision and recall: the classification of spoken texts delivers 80.5% for precision with 100% recall, whereas the classification of written texts works with perfect precision (100%) but with lower recall (96.3%). This means that some written originals were erroneously recognised as spoken texts, but none of the spoken texts were recognised as written texts. This indicates that some written texts in our data may contain features considered specific to spoken language.

Table 9.7: Classification results for the second mode distinction in %

	Precision	Recall	F-Measure
INTER	81.5	100.0	89.8
TRANS	100.0	94.6	97.2
Weighted average	96.4	95.6	95.8

The mode distinction in translations also achieves high accuracy (95.7%) with an average F-measure of 95.8% (See Table 9.7). Again, we observe a higher F-measure for the written translations than for the spoken ones (97.2% vs. 89.8%). Similarly to the first classification task, interpretations are recognised with better recall (100% vs. 94.6%) and translations are identified with better precision (100%

vs. 81.5%). The confusion matrix shows that around 5.4% of the translations were erroneously labelled as interpretations.

Table 9.8: Classification results for the third mode distinction in %

	Precision	Recall	F-Measure
GO-SP+INTER	81.6	99.7	89.7
GO-WR+TRANS	99.9	95.9	97.9
Weighted average	97.1	96.5	96.6

In the third classification task, we achieve 96.5% of accuracy and an F-measure of 96.6%. Similarly to the other mode distinction tasks, written texts, regardless of their text production type, achieve a better F-measure than the spoken ones (97.9% vs. 89.7%), with higher precision observed for the written texts (99.9% vs. 81.6%) and a higher recall for the spoken ones (99.7% vs. 95.9%), see Table 9.8. Mixing both text production types for the mode distinction task results in an intuitively insignificant drop in the observed scores.

The results show that we can automatically detect mode in translation, and the results of such a classification are comparable with the results on mode distinction in non-translated German. We achieve very good classification results in all tasks on mode distinction.

4.3 RQ3

We compare the three F-measure scores resulting from the three classification tasks within RQ1 – differentiation between translation and non-translation⁶ with the three F-measure scores from the three classification tasks within RQ2 – differentiation between spoken and written texts⁷. For this, we summarise the results of all these classification tasks in Figures 9.1 and 9.2. The first figure contains average F-measure scores for the text production type distinction, whereas the second figure illustrates the F-measure scores for the mode distinction.

As seen from the graphs, it is easier to detect mode than text production type in our dataset given the same feature set. These results confirm our assumption that variation along the dimension of mode is stronger than that along the dimension of text production.

⁶We use the weighted average F-measure from Tables 9.3, 9.4, 9.5.

⁷We use the weighted average F-measure from Tables 9.6, 9.7, 9.8.

Ekaterina Lapshinova-Koltunski

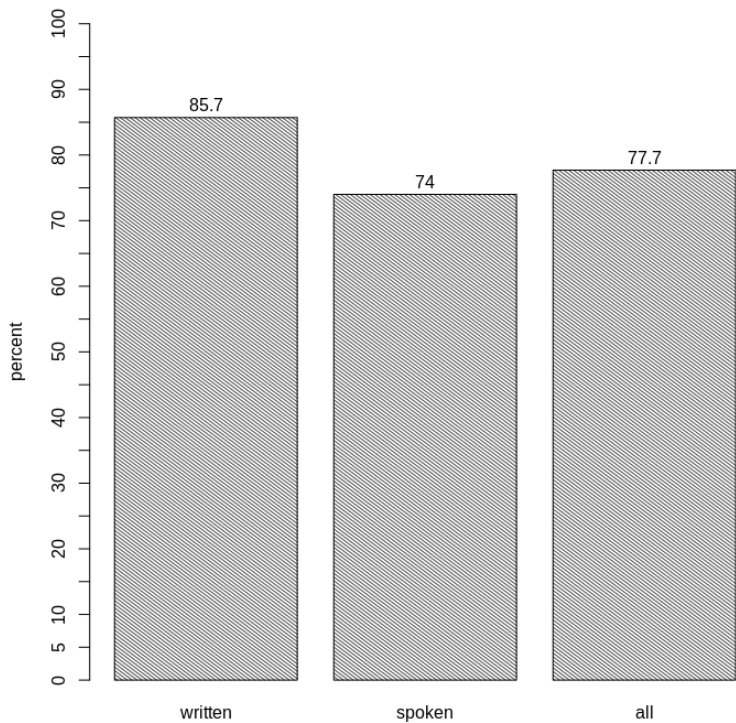


Figure 9.1: Classification results for text production distinction in % (weighted average)

4.4 RQ4

As explained in §3.4, we use automatic attribute selection to test if the same language patterns are responsible for prediction of mode in both translated and non-translated texts. In both cases, the language patterns should contribute to the classification of the two classes: spoken and written.

We start with the evaluation of the language patterns relevant for the mode prediction task in non-translations. We use cross-validation with 10 folds which records in how many folds each of the attributes (our language patterns) appeared in the best subset found. We select attributes that appeared in at least one fold, which results in a list containing 10 language patterns⁸: content words (10),

⁸The figure in brackets indicates the number of folds the feature appears in.

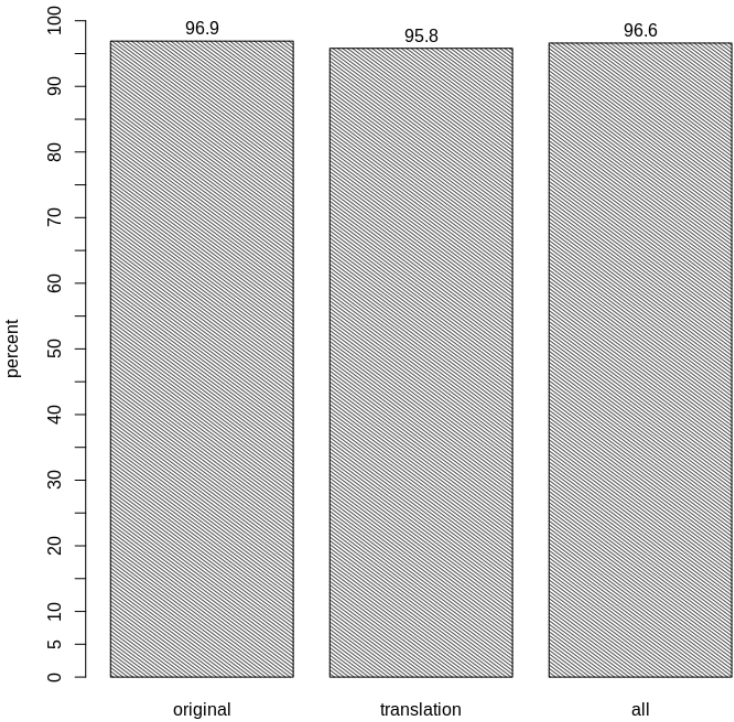


Figure 9.2: Classification results for mode distinction in % (weighted average)

nominal parts-of-speech (10), verbal parts-of-speech (9), *ung*-nominalisations (10), obligation (7), additive relations (2), adversative relations (7), modal relations (2), personal pronouns (10) and demonstratives (9).

The same procedure is applied on the dataset of translations. The list of language patterns contributing to the mode distinction here contains seven items: content words (4), nominal parts-of-speech (10), verbal parts-of-speech (3), *ung*-nominalisations (10), temporal relations (10), personal pronouns (4) and demonstratives (2).

For a better visualisation, we outline the language patterns selected for the mode distinction in both non-translated and translated texts in Table 9.9. We also relate them to the more abstract feature categories as well as contextual parameters introduced earlier (see Table 9.2 in §3.3 for an overview).

Table 9.9: Features contributing to the mode distinction

non-translated			translated		
par.	feat.	lang.pattern	lang.pattern	feat.	par.
Field	participants, processes	content words	content words	participants, processes	Field
Field	participants	nominal pos	nominal pos	participants	Field
Field	processes	verbal pos	verbal pos	processes	Field
Field	processes	ung-nom.	ung-nom.	processes	Field
Tenor	modality	obligation			
Tenor	cohesion	pers. pron.	pers. pron.	cohesion	Mode
Mode	cohesion	dem. pron.	dem. pron.	cohesion	Mode
Mode	cohesion	additive			
Mode	cohesion	adversative			
			temporal	cohesion	Mode

As seem from the Table 9.9, the two lists have an overlap of six language patterns, while the first list contains three language patterns not included in the second list (modal verbs of obligation, additive and adversative relations). However, the second list is not entirely a subset of the first one, as it contains one language pattern which is not included into the first list (temporal relations). We mark the non-overlaps in grey in the table. In terms of abstract linguistic features, participants, processes and cohesion contribute to the mode distinction in both translated and non-translated texts, which correspond to the contextual parameters of Field and Mode. However, in the texts originally produced in German, there is also modality corresponding to Tenor, which is not distinctive for mode in the translations. It is also interesting to see that although discourse relations contribute to the mode distinction in both translations and non-translations, they differ in the logico-semantic types in each list.

Since the majority of the features overlap (6/10 and 6/7), we suggest that the same features (especially if interpreted in terms of abstract categories) are responsible for the variation between translations and interpretations and between written and spoken texts. The overlap in the features common for the distinction of mode may trace back to the register the texts in the dataset belong to – they are all speeches from the parliamentary debates.

5 Conclusion and discussion

The present study focuses on the variation in English-to-German translation along the mode dimension. Translation variation is reflected in the linguistic features that we were able to analyse with language patterns derived from variational linguistics. We extracted the distribution of these patterns in spoken and written texts that included both texts originally spoken or written in German, and translations and interpretations. This distributional information was used to statistically model variation along the text production dimension (translations vs. non-translation) and the mode dimension (spoken vs. written). Our results show that we are able to automatically tease apart translation from non-translations, as well as spoken texts from the written texts using the same feature set. However, it turned out to be easier to automatically differentiate between spoken and written texts regardless of their production type, which confirms our assumption that language variation along the dimension of mode is stronger than that along the dimension of text production. We are also able to find out which linguistic features contribute to the distinction between spoken and written mode in both translated and non-translated language.

This brings our findings in accordance with Shlesinger & Ordan (2012)'s claim that mode exerts a stronger effect than text production. This means that the difference between spoken and written texts is stronger than that between translations and non-translations. In this way, the interpretations in our dataset show more similarities to the speeches originally spoken in German than to the written translation, making interpretations more 'spoken' than 'translated'.

At the same time, we realise that our study also has a number of limitations. First of all, we use a feature set inspired by variational linguistics. Although it has been applied in the analysis of translationese in a number of previous studies, it was originally developed for the analysis of register variation that also includes variation along the dimension of mode. However, many of the language patterns in our set are extensively applied in the analysis of translationese (e.g. cohesive markers) as well.

Another drawback of the present study is the limitation of the corpus data – it includes political speeches only. Yet, whereas there are many translation corpora which could be used for such an analysis, it is hard to find comparable interpreted data.

In the future, we should extend the features and the data to further investigate the specifics of translated and interpreted texts. It will be also interesting to have a closer look at the features contributing to the mode distinction and perform a qualitative analysis of these features.

Ekaterina Lapshinova-Koltunski

Abbreviations

RQ – research question
 pos – part-of-speech
 pers. pron. – personal pronoun
 dem pron. – demonstrative pronoun
 ung-nom. – ung-nominalisation
 par. – parameter
 feat. – feature
 lang.pattern – language pattern

References

- Baayen, R. Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In G. Francis Baker M. & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: Benjamins.
- Baroni, Marco & Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3). 259–274. DOI: [10.1093/lc/fqi039](https://doi.org/10.1093/lc/fqi039).
- Bernardini, S. & A. Ferraresi. 2011. Practice, description and theory come together: Normalization or interference in Italian technical translation? *Meta* 56. 226–246.
- Bernardini, S., A. Ferraresi & M. Miličević. 2016. From EPIC to EPTIC—exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28. 61–86.
- Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Bizzoni, Yuri & Elke Teich. 2019. Analyzing variation in translation through neural semantic spaces. In *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC) at RANLP-2019*. Varna, Bulgaria: ACL.
- Dayter, Daria. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM* 16(2). To appear, 241–264.
- Defrancq, B., K. Plevoets & C. Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics*, 195–222. New York: Springer International Publishing.

- Evert, Stefan & Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the corpus linguistics 2011 conference*. Birmingham, UK: University of Birmingham.
- Evert, Stefan & Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts : A multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions* 300. 47.
- Evert, Stefan & The CWB Development Team. 2019. *CQP query language tutorial*. Tech. rep. Version Version 3.4.15.
- Ferraresi, A. & M. Miličević. 2017. Phraseological patterns in interpreting and translation. Similar or different? In G. De Sutter, M.-A. Lefer & I. Delaere (eds.), *Empirical translation studies. New methodological and theoretical traditions*, vol. 300 (Trends in Linguistics. Studies and Monographs [TiLSM]), 157–182. Berlin: Mouton de Gruyter.
- Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (eds.), *Translation studies in Scandinavia*, 88–95. Lund: CWK Gleerup.
- Hall, Mark A. 1998. *Correlation-based feature subset selection for Machine learning*. University of Waikato. (Doctoral dissertation).
- Halliday, M. A.K. 2004. *An introduction to functional grammar*. London: Arnold.
- Halliday, M. A.K. & C. M.I. M. Matthiessen. 2014. *Halliday's introduction to functional grammar*. 4th edn. London: Routledge.
- He, He, Jordan Boyd-Graber & Hal Daumé III. 2016. Interpretese vs. Translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 971–976. San Diego, California: Association for Computational Linguistics. DOI: [10.18653/v1/N16-1111](https://doi.org/10.18653/v1/N16-1111). <http://www.aclweb.org/anthology/N16-1111>.
- Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor & Ruslan Mitkov. 2010. Identification of translationese: A supervised learning approach. In A. Gelbukh (ed.), *Proceedings of CICLing-2010*, vol. 6008 (LNCS), 503–511. Springer, Heidelberg.
- Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec & Céline Rouveirol (eds.), *Machine learning: ecml-98*, 137–142. Berlin: Springer.
- Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Unpublished PhD thesis. Poznan, Poland: Uniwersytet im. Adama Mickiewicza. (Doctoral dissertation).

Ekaterina Lapshinova-Koltunski

- Karakanta, Alina, Katrin Menzel, Heike Przybyl & Elke Teich. 2019. Detecting linguistic variation in translated vs. Interpreted texts using relative entropy. In *Empirical investigations in the forms of mediated discourse at the European parliament, thematic session at the 49th poznan linguistic meeting (PLM2019), poznan*.
- Karakanta, Alina, Mihaela Vela & Elke Teich. 2018. EuroParl-UdS: Preserving and extending metadata in parliamentary debates. In *ParlaCLARIN workshop, 11th Language Resources and Evaluation Conference (LREC2018)*. Miyazaki, Japan. http://lrec-conf.org/workshops/lrec2018/W2/pdf/10_W2.pdf. published.
- Kotze, Haidee. 2019. Converging what and how to find out why: An outlook on empirical translation studies. In Lore Vandevoorde, Joke Daems & Bart Defrancq (eds.), *New empirical perspectives on translation and interpreting* (Routledge Advances in Translation and Interpreting Studies), 333–371. Routledge.
- Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 47–56. Varna, Bulgaria: Incoma Ltd., Shoumen, Bulgaria. <https://www.aclweb.org/anthology/W19-8706>.
- Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of LREC-2020*. Marseille, France.
- Lapshinova-Koltunski, Ekaterina. 2017. Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies: New methodological and theoretical traditions*, vol. 300 (TILSM series), 207–234. TILSM series. Berlin: Mouton de Gruyter.
- Lapshinova-Koltunski, Ekaterina. 2019. Exploring linguistic differences between novice and professional translators with text classification methods. In Lore Vandevoorde, Joke Daems & Bart Defrancq (eds.), *New empirical perspectives on translation and interpreting* (Routledge Advances in Translation and Interpreting Studies), 215–239. London: Routledge.
- Lapshinova-Koltunski, Ekaterina & José Manuel Martínez Martínez. 2017. Statistical insights into cohesion: Contrasting English and German across modes. In Markéta Janebová, Ekaterina Lapshinova-Koltunski & Michaela Martinková (eds.), *Contrasting English and other languages*, 130–163. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Laviosa, Sara. 2002. *Corpus-based translation studies, theory, findings, application*. Amsterdam: Rodopi.

- Neumann, Stella. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin, Boston: Mouton de Gruyter.
- Nivre, Joakim et al. 2019. *Converging what and how to find out why an outlook on empirical translation studies*. Tech. rep. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 333–371. <http://hdl.handle.net/11234/1-1983>.
- Olohan, M. & M. Baker. 2000. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1. 141–158.
- Øverås, L. 1998. In search of the third code. An investigation of norms in literary translation. *Meta* 43. 557–570.
- Scott, N. 1998. *Normalisation and readers' expectations: A study of literary translation with reference to lispector's A hora da estrela*. doctoral dissertation. Liverpool: University of Liverpool. (Doctoral dissertation).
- Shlesinger, Miriam & Noam Ordan. 2012. More spoken or more translated?: Exploring a known unknown of simultaneous interpreting. *Target* 24. 43–60.
- Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 88–99. <http://www.aclweb.org/anthology/K17-3009>.
- Teich, Elke. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.
- Toury, Gideon. 1995. *Descriptive translation studies – and beyond*. Amsterdam: John Benjamins Publishing Company.
- Vapnik, Vladimir N & A Ja Chervonenkis. 1974. *Theory of pattern recognition*. Moscow: Nauka.
- Volansky, Vered, Noam Ordan & Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1). 98–118. DOI: [10.1093/llc/fqt031](https://doi.org/10.1093/llc/fqt031).
- Witten, Ian H., Eibe Frank & Mark A. Hall. 2011. *Data mining: Practical Machine learning tools and techniques: Practical Machine learning tools and techniques* (The Morgan Kaufmann Series in Data Management Systems). Amsterdam: Elsevier Science.

Empirical studies in translation and discourse

Set blurb on back with \BackBody{my blurb}

