Empirical investigations into the forms of mediated discourse at the European Parliament

Edited by

Marta Kajzer-Wietrzny

Adriano Ferraresi

Ilmari Ivaska

Silvia Bernardini

Translation and Multilingual Natural Language Processing

language science press

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 investigations into the forms of mediated discourse at the European Parliament

Edited by

Marta Kajzer-Wietrzny

Adriano Ferraresi

Ilmari Ivaska

Silvia Bernardini

language science press

Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.). 2022. *Empirical investigations into the forms of mediated discourse at the European Parliament* (Translation and Multilingual Natural Language Processing). Berlin: Language Science Press.

This title can be downloaded at:
redefine\lsURL

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: XeLaTeX

redefine \publisherstreetaddress
redefine \publisherurl

Storage and cataloguing done by redefine \storageinstitution

Freie Universität Berlin

Contents

Using European Parliament data in translation and interpreting research: An introduction

Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini

1 Background

Ever since its inception, Corpus-based Translation Studies (CTS) have been preoccupied with systematic and rigorous investigations of translations in the search for linguistic characteristics that set them apart from original texts (Laviosa 1998, Olohan & Baker 2000, Kenny 2001, Kruger & van Rooy 2010, Redelinghuys & Kruger 2015, De Sutter & Lefer 2019). Interpreting scholars followed suit, and despite the far more time-consuming and complex compilation process, corpus research on interpreting steadily progresses (Shlesinger 1998, Shlesinger & Ordan 2012, Bendazzoli & Sandrelli 2005, Kajzer-Wietrzny 2012, Defrancq 2015, Defrancq & Plevoets 2018, Kajzer-Wietrzny & Ivaska 2020, Dayter 2021).

The number of studies taking advantage of the machine-readable format of corpora and investigating vital research questions at textual level keeps growing both in Translation and in Interpreting Studies. At the same time, both interpreting and translation corpora are becoming more multifaceted (Bernardini 2011, Castagnoli 2020), allowing comparisons between translations and their source texts (parallel perspective), between translations and comparable original texts in the same language (monolingual comparable perspective), and sometimes across multiple translations of the same source text (multi-parallel perspective). They also are far richer in annotation levels and metadata (Reynaert et al. 2021) making increasingly more advanced multifactorial analyses possible.

Although progress is clearly visible across both translation modes, interpreting will always involve further layers of complexity, due to the necessity to tran-

Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini. 2022. Using European Parliament data in translation and interpreting research: An introduction. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, iii–x. Berlin: Language Science Press. DOI: ??

scribe data and account for spoken language-specific traits. At the beginning, interpreting corpora were mostly comparable. Today, most are parallel and aligned at sentence level, with a few also including alignment with corresponding translated texts and original videos (EPTIC, Ferraresi & Bernardini 2019), or even sound-to-text alignment at word level (PINC, cf. Chmiel et al. in this volume). Scholars compiling their corpora make use of such technological advancements as speech recognition to speed up the transcription process like in EPTIC, PINC or PETIMOD (Ferraresi & Bernardini 2019; Koržinek & Chmiel in press, Corpas Pastor & Sánchez Rodas in this volume) or speaker identification to disambiguate interpreter voices e.g. in PINC (Koržinek & Chmiel in press, Chmiel et al. in this volume). Corpora are tagged for Parts Of Speech (POS), lemmas, dependencies and features of orality. The level of granularity varies, from simple orthographic transcription and annotation to very specific orality traits, e.g., pause length.

Investigations in Corpus-based Translation Studies and Corpus-based Interpreting Studies have initially focused on translation or interpreting "universals", to later look at recurrent shared phenomena through new lenses, like those of "language mediation"(Ulrych & Murphy 2008) or "cognitive constraints" (Lanstyák & Heltai 2012). Kotze's framework of constrained varieties (2019: 346), in a way, unites the two by classifying constraints into five "interacting and overarching dimensions", i.e., language activation; modality and register; text production; proficiency; task expertise. This approach makes it possible to adopt a broader perspective and may help shedding more light on which linguistic features typically associated with translation result from bilingual activation in general, or from the process of reworking a text. Furthermore, translation and interpreting scholars have also worked at the interface between corpus linguistics and other linguistic approaches in order to explore in greater depth the complexity linked to translation and interpreting of sensitive social issues.

Parliamentary data have been used extensively and for many years in corpus-based linguistic research. Due to its multilingual nature, European Parliament data (Tiedemann 2012) in particular have been used widely in translation research, and still offer today a wealth of unique opportunities to investigate constraints that can affect linguistic production. The European Union institutions, in general, are likely to be the richest source of multilingual and multimodal texts which are spoken, written and re-written for various recipients in diverse forms depending on the communicative goal. The activities connected with the European Parliament plenaries involve Members of the European Parliament either delivering a speech impromptu or upon earlier preparation, usually based on existing documentation at various stages of completion/translation. All speeches, be they written-up and then read out or delivered impromptu, are transcribed

into verbatim reports. Both cases involve adaptation to a different modality. The oral speeches are interpreted simultaneously and the reports until 2011 were also translated. Thus, the European Parliament data constitute a valuable source of texts that in Kotze's (2019: 346) classification of constrained varieties could be categorized as bilingual and/or dependent/mediated, "in the sense that a prior text delimits and shapes the[ir] production". In addition to videos with multilingual audio tracks, the European Parliament website provides information about speakers and topics of the debate. From a methodological perspective, the EP material also guarantees a great degree of homogeneity, as translations and interpretations are consistently performed by experienced professionals, and speeches in various modes are delivered in the same institutional setting (Monti et al. 2005), which is particularly valuable in corpus studies, where data comparability is frequently a challenge. Content-wise, the EP plenaries provide a diversity of topics and a wide range of speakers and interpreters. Issues discussed at the plenaries range from mundane and bureaucratic to terminologically dense or highly sensitive, providing ample opportunities for investigation of interpreting or translation challenges.

For the most part, research on spoken and intermodal mediated discourse at the European Parliament plenaries has been scattered and no single volume has attempted to capture the complexity of language mediation in the two modes in this very specific context. In this volume we focus on quantitative and qualitative spoken and intermodal mediated discourse looking either solely at interpreting at the European Parliament plenaries, or at both interpreting and translation, but never at written translation alone. This ties in with the specific spoken/intermodal nature of the plenaries at the EP, where speeches are first spoken or read out and interpreted and are only then transcribed and (until a few years ago) translated.

2 Spoken mediated discourse

The first three chapters in the section on spoken mediated discourse, i.e., interpreting, adopt a linguistically oriented perspective, looking at convergence between orators and interpreters, analysing formality of mediated and non-mediated texts and investigating predictors of interpreters' fluency.

In the first chapter, using EPIC-G (Bernardini et al. 2018) Defrancq and Plevoets examine speeches delivered by Members of the European Parliament and their interpretations first, to theoretically determine whether MEPs or interpreters have more expertise in the linguistic genre of the Parliament. The empirical part con-

centrates on key 3- and 4-grams, which help to identify the dominant group shaping the linguistic features of the genre. Their results suggest that MEPs adopt some of the interpreters' patterns, thus supporting Pöchhacker's (2005) idea that in an interpreter-mediated encounter all interactants influence each other's communicative behaviour.

In the second chapter, Ivaska, Ferraresi and Kajzer-Wietrzny draw on EPTIC to examine speeches read out and delivered impromptu at the European Parliament by native English speakers to draw a list of linguistic features contributing to formality or informality. Next, they use a human-validated dataset of formality features to examine differences between interpreted and non-interpreted texts. The outcomes point to a higher level of formality of interpreted texts.

Chapter three, by Chmiel, Korzinek, Kajzer-Wietrzny, Janikowski, Jakubowski and Polakowska, introduces PINC — the Polish Interpreting Corpus — a corpus of European Parliament speeches and their interpretations. Its rich metadata make the corpus unique, insofar as it includes, e.g., interpreter identification and very fine-grained text-to-speech alignment. The study in which the corpus is exploited proves that fluency is modulated by the source text speech and articulation rate, as well as the target text compression rate, and that the majority of interpreters produce interpretations which are longer than the source texts. Interpreter identification further made it possible to discover individual differences in compression rate.

Chapter four in the volume adopts a more qualitative approach to address sensitive, and hence challenging issue for interpreters, i.e., migration. Analysing an ad-hoc interpreting corpus comprising transcripts of speeches and their interpretations, Anghelli and Mori investigate the topic of migration through the lens of contrastive qualitative discourse analysis. They evaluate which strategies are employed by interpreters to preserve, alter or distort politicians' intentions and to detect cues mitigating and/or intensifying the pragmatic intent of the original speakers during plenary sessions devoted to migration.

3 Intermodal investigations

The section on intermodal comparisons begins with chapter five, in which Lefer and De Sutter carry out a corpus study of the French rendition of English concatenated nouns in simultaneous interpreting and written translation. Using parallel corpus data extracted from the European Parliament Translation and Interpreting Corpus (EPTIC), they model the French renditions of English concatenated nouns with regression analysis, attempting to establish which factors affect the

use of equivalent vs. non-equivalent renditions. The outcomes highlight the key commonalities between the two modes and prove that the cognitive sources in Halverson's gravitational pull model can be successfully researched with a multifactorial design.

In chapter six, Mikolič Južnič and Pisanski Peterlin examine sentence-initial connectors in mediated and non-mediated spoken and written Slovene by comparing the Slovene section of EPTIC, two monolingual reference corpora of Slovene, and a subsection of a comparable Slovene corpus of parliamentary discourse. The results show notable differences between the two modes of production, and at the same time reveal that other factors impact on results, such as genre and mediation status.

In chapter seven, Przybyl, Karakanta, Menzel and Teich investigate the effects of mediation and mode in a data-driven, exploratory approach to detecting linguistic features typical of translation/interpreting. The approach employs simple word-based n-gram language models combined with the information-theoretic measure of relative entropy used as a method of corpus comparison. In addition to confirming previous findings from the literature, the authors detect new features, such as a tendency towards more general lexemes in the verbal domain in interpreting, and features related to nominal style in translation.

Chapter eight by Corpas Pastor and Sánchez Rodas presents an NLP-enhanced analysis of shifts in the named entities in an English<>Spanish subcorpus of the translation and interpreting corpus of the Committee on Petitions of the European Parliament. The outcomes suggest that tendencies such as normalisation, transformation and simplification depend on the language direction, the mediation mode, and the semantic category of the named entity.

4 Issues and open challenges

This volume presents a unique collection of papers on mediated discourse either in its spoken form or both spoken and written. Looking at the contributions, it is hard not to notice that to some extent they reflect the current dominant research avenues undertaken by interpreting and translation scholars working with data other than the European Parliament plenaries. Despite the very specific context of production, the volume thus makes it possible to make reflections which have a bearing on CBTS and CBIS at large.

First, the analysed interpreting and inter-modal corpora are relatively small (so much so that they have been referred to as "nanocorpora" Collard & Defrancq (2016)), especially when evaluated from the perspective of monolingual corpus

linguistics research. Although voice recognition does facilitate spoken corpus creation, the processes needed to verify its output are still extremely time consuming. Equally challenging is the alignment of source and target texts, as finding one-to-one correspondences between spoken source and interpreted texts is not always trivial. Due to the small size, the need to incorporate richer metadata in corpus design also becomes crucial (**ReynaertMackenDeSutter2021**). It is only thanks to metadata that analyses can account for a number of fixed factors, while at the same time controlling for random effects related to individual variation, such as interpreter ID. Even though awareness of the problem is higher than in the past, the number of studies trying to account for the problem of variation is still proportionally low.

The problem could, in part, be solved with more data. It seems, however, that in the case of spoken and inter-modal analyses, collecting and pre-processing the required amount of data lies beyond the capacity of a single scholar. And yet small, individually compiled corpora still constitute the majority of datasets analysed in translation and interpreting studies. This volume shows a more optimistic tendency in this respect. The corpora used in a number of contributions presented here are the result of cooperation between scholars: examples include the EPTIC corpus (Ferraresi & Bernardini 2019), which is a joint effort of a few teams scattered across Europe, and the Europarl-UdS corpus (Przybyl et. al in this volume), which makes use of data collected in other centres (Ghent and Poznan) and enriches them with more data and annotation layers. The way forward probably lies in coming up with a shared and customizable corpus format that could work for more than one research group, and could make data exchange between groups a more common practice. It is only in such a way that corpus-based translation and interpreting research can escape the problem of nano-size.

Compiling and investigating corpora that allow for the analysis of spoken mediated discourse and intermodal comparisons will always constitute a greater challenge than corpora of written texts. The present volume illustrates a number of ways in which this challenge can be approached in the context of qualitative and quantitative studies, both corpus-based and corpus-driven.

References

Bendazzoli, Claudio & Annalisa Sandrelli. 2005. An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In Heidrun Gerzymisch-Arbogast & Sandra Nauert (eds.), *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional*

Translation, 149. https://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html (5 December, 2012).

Bernardini, Silvia. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS - A Journal of Professional Communication* 26. 2–13. http://hdl.handle.net/11250/2393975.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 21–42. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-6199-8_2.

Castagnoli, Sara. 2020. Translation choices compared: Investigating variation in a learner translation corpus. In *Translating and comparing languages: Corpus-based insights: Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference*, vol. 6, 25. Presses universitaires de Louvain.

Collard, Camille & Bart Defrancq. 2016. How to use a nanocorpus. Enriching corpora of interpreting. Paper presented at Corpus Linguistics in the South: Doing corpus linguistics with large and small corpora 2016. https://lib.ugent.be/catalog/pug01:8518823.

Dayter, Daria. 2021. Variation in non-fluencies in a corpus of simultaneous interpreting vs. Non-interpreted English. *Perspectives* 29(4). 489–506.

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* 0(0). 1–23. DOI: 10.1080/0907676X.2019.1611891.

Defrancq, Bart. 2015. Corpus-based research into the presumed effects of short EVS. *Interpreting* 17(1). 26–45. DOI: 10.1075/intp.17.1.02def.

Defrancq, Bart & Koen Plevoets. 2018. Over-uh-load, filled pauses in compounds as a signal of cognitive load. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 43–64. Singapore: Springer.

Ferraresi, Adriano & Silvia Bernardini. 2019. Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In Irene Doval & M. Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*, vol. 90 (Studies in Corpus Linguistics), 123–139. Amsterdam/Philadelphia: John Benjamins.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Adam Mickiewicz University, Poznan. (Doctoral dissertation).

Kajzer-Wietrzny, Marta & Ilmari Ivaska. 2020. A Multivariate Approach to Lexical Diversity in Constrained Language. *Across Languages and Cultures* 21(2). 169–194.

Kenny, Dorothy. 2001. *Lexis and creativity in translation.* Manchester: St. Jerome Publishing.

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*, 333–371. London.

Kruger, Haidee & Bertus van Rooy. 2010. The features of non-literary translated language: A pilot study. In Richard Xiao (ed.). http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/Kruger_Van_%20Rooy.pdf (5 December, 2012).

Lanstyák, István & Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13(1). 99–121. DOI: 10.1556/Acr.13.2012.1.6.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4). 557–570.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli & Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta: Journal des traducteurs/Meta: Translators' Journal* 50(4). DOI: https://doi.org/10.7202/019850ar.

Olohan, Maeve & Mona Baker. 2000. Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across languages and cultures* 1(2). 141–158. http://www.akademiai.com/index/M7862H62GMLKH525.pdf (5 December, 2012).

Pöchhacker, Franz. 2005. From operation to action: Process-orientation in interpreting studies. *Meta* 50 (2). 682–695.

Redelinghuys, Karien & Haidee Kruger. 2015. Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics* 20(3). 293–325. DOI: 10.1075/ijcl.20.3.02red.

Reynaert, Ryan, Lieve Macken & Gert De Sutter. 2021. Building a new-generation corpus for empirical translation studies: The Dutch Parallel Corpus 2.0. In Vincent X. Wang, Lily Lim & Defeng Li (eds.), *New perspectives on corpus translation studies*, 75–100. Singapore: Springer.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: Journal des traducteurs* 43(4). 486–493. DOI: 10.7202/004136ar.

Shlesinger, Miriam & Noam Ordan. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24(1). 43–60. DOI: 10.1075/target.24.1.04shl.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. Istanbul: European Language Resources Association (ELRA).

Ulrych, Margherita & Amanda Clare Murphy. 2008. Descriptive translation studies and the use of corpora: Investigating mediation universals. In C. Taylor Torsello, K. Ackerley & E. Castello (eds.), *Corpora for university language teachers*. Bern: Peter Lang. http://hdl.handle.net/10807/2116 (24 July, 2018).

Chapter 1

Ut interpres: Linguistic convergence between orators and interpreters in the European Parliament

Bart Defrancq[a] & Koen Plevoets[a]
[a]EQTIS, Ghent University

This paper combines a theoretical and an empirical approach to the analysis of converging linguistic features in speeches held by Members of the European Parliament and interpretations in the same Parliament. The theoretical approach seeks to determine which group has more seniority and therefore more expertise in the linguistic genre of the Parliament. The empirical analysis concentrates on key 3- and 4-grams used in speeches and interpretations to determine which group's usage is more expert and can be considered the dominant group shaping the linguistic features of the genre. The two-pronged approach reveals that interpreters are the expert group and that for the items considered the case can be made that Members adopt interpreters' lexical patterns. The study thus provides complementary evidence for Pöchhacker's (2005) idea that in an interpreter-mediated encounter all interactants influence each other's communicative behaviour.

1 Introduction

Cicero's (46 BCE) self-reported translation method "*nec ut interpres, sed ut orator*" ('not as an interpreter, but as an orator') is widely quoted in the translation literature as one of the oldest examples of a functionalist approach to translation (see for instance Nord 2013). It does not seem to have met with the same kind of enthusiasm in Interpreting Studies. That is of course perfectly understandable, considering the negative view it carries on interpreters (although the Latin *interpres* covers both translators and interpreters, as well as mediators and exegetes).

Bart Defrancq & Koen Plevoets. 2022. *Ut interpres*: linguistic convergence between orators and interpreters in the european parliament. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 1–27. Berlin: Language Science Press. DOI: ??

In this study we will subvert Cicero's quote and ask ourselves if there is evidence that orators speak *ut interpretes*, like interpreters, and more in particular simultaneous interpreters.

Since the 1990s, the theoretical work on simultaneous interpreting has increasingly made room for functionalist thinking, albeit at a slower pace than in other areas of interpreting research. Pöchhacker (1994) made a first comprehensive attempt at transferring functionalist theories of translation to conference interpreting, categorising and describing its various *skopoi*. Major empirical landmarks by Diriker (2004) and Monacelli (2009) followed, illustrating simultaneous interpreters' agency during conference assignments. Yet, for all the progress that was made, it seems that the functionalist approach has not yet been exploited to its full potential.

In Pöchhacker's (2005) interactant model of interpreting, shown in Figure 1, interpreting is described in terms of an interaction between (at least) three participants, each coming to the interaction with their perspective on the interaction and the interactants, embedded in their socio-cultural background.

As Pöchhacker admits, the model fits situations of triadic communication best. If we were to apply the model to simultaneous interpreting in a conference, it would certainly have to include more interactants, and, crucially, more interpreters. Interaction obviously also takes place between boothmates and even with colleagues in other booths, directly or through the *chef d'équipe*. This aspect of conference interpreting is clearly under-represented in the literature and has only been thoroughly investigated by Duflou (2016) with respect to turn-taking.

Ultimately, the interactant model is designed to be a framework to describe communicative behaviour. As Pöchhacker (2005) puts it:

> The 'interactant model of the situation' […] seeks to show the multiple dynamic relationships which make up the communicative situation as it 'exists' for a given interactant and shapes his or her communicative behaviour. Pöchhacker (2005: 688)

The communicative behaviour that has been under most scrutiny in the literature is, quite understandably, the interpreter's, with a specific focus on communicative behaviour that runs counter various interpretations of the so-called *conduit model*, i.e. the historic normative view that interpreters produce linguistic output based on linguistic input abstaining from interfering in the communication between primary participants (Diriker 2004; Monacelli 2009; Bartłomiejczyk 2016: 2020). Similarly, in studies that focus on linguistic properties (see §4), the focus has been on the interpreters' output and how it is shaped by aspects of the

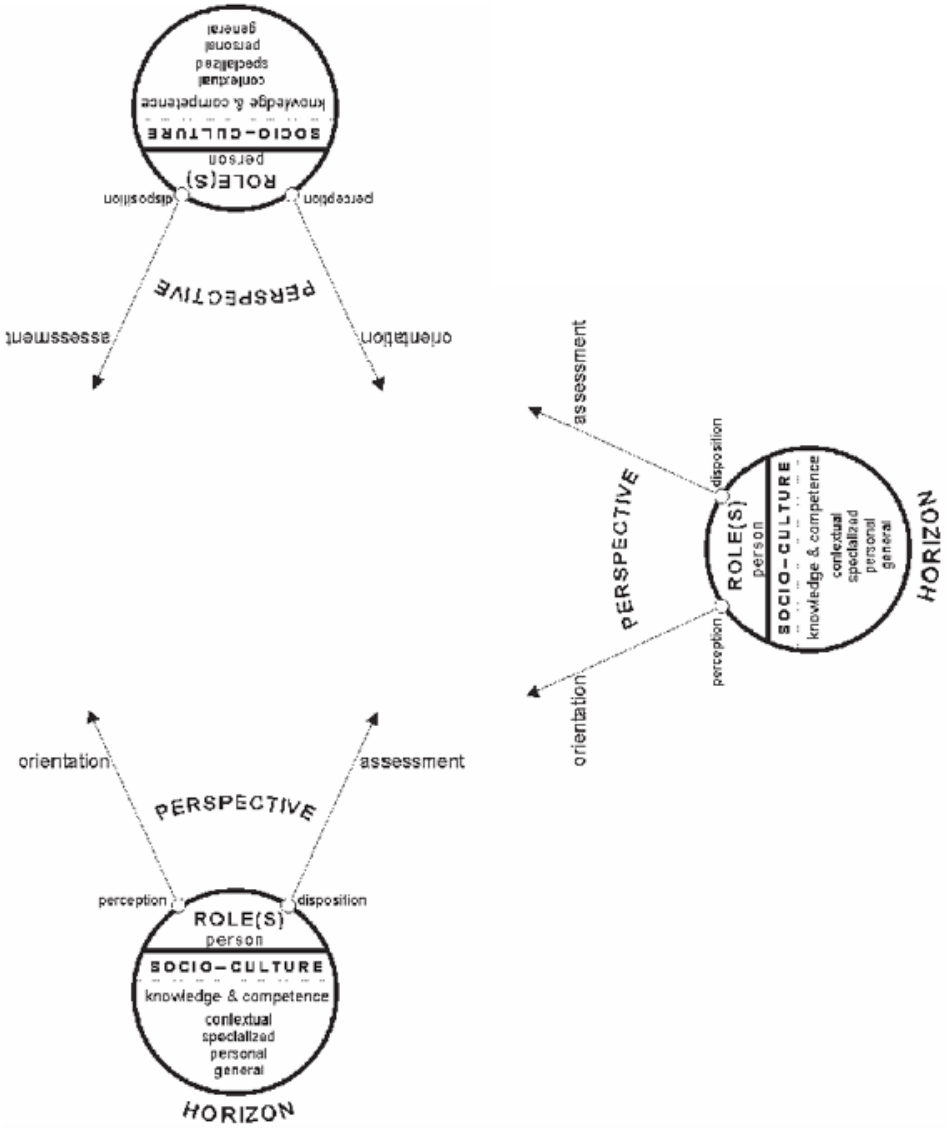


Figure 1: Pöchhacker's interactant model (Pöchhacker 2005: 689).

communicative situation, including the other interactants. A fairly representative view in that respect is the one voiced by (Bartłomiejczyk 2016), explaining that interpreters at the European Parliament (EP) acquire keywords and expressions due to prior exposure to the primary participants and to the boothmate (similar views are held by Duflou 2016 and by Henriksen 2007 for the European Commission's DG SCIC):

> Secondly, the EP discourse is characterised by a large degree of repetitiveness, which concerns certain phrases that might well be described as clichés as well as keywords. One of such keywords is, for example, *solidarity*, which collocates with the adjectives *European* and *multinational* [...]. This is conducive to experienced interpreters building up a large repertoire of ready-made translation solutions, which may be worked out individually or copied from boothmates. (Bartłomiejczyk 2016: 57)

Acquired knowledge is obviously part of an interactant's perspective, which, in turn, is part of the communicative situation.

Of course, what holds true for interpreters, also holds true for the other interactants. Their exposure to interpreters' output is likely to impact their perspective and, as a result, their communicative behaviour. This dimension is, however, poorly represented in the literature. Apart from a systematic study of references to interpreters and interpreting by members of the European Parliament's (MEP) speeches (Bartłomiejczyk 2017), and a series of quality surveys (for an overview, see Kurz 2001), the perspectives of primary participants as shaped by their interactions with interpreters in conference situations is hardly explored.

In this paper we set out to explore precisely that dimension. We will first review the concept of linguistic convergence and some of our own studies on lexical patterns and potential linguistic convergence between Members of the European Parliament (MEPs) and interpreters in the European Parliament. This will lead us to the research questions at the end of §2 These will focus on the potential role of interpreters in shaping the linguistic patterns of MEPs. To gain a better understanding of the EP context and to provide a theoretical answer to our research questions, we review the relevant research on MEPs in general and Dutch-speaking MEPs in particular and on EP interpreters and the Dutch booth (§3 and 4 respectively). §5 presents the quantitative and qualitative methods and results of a detailed study of the n-grams or lexical bundles that were also the subject of investigation in our previous studies. §6, finally presents the conclusions.

2 Linguistic convergence between MEPs and interpreters

According to various sociological and sociolinguistic theories, such as Communication Accommodation Theory (CAT; Giles 1973; **GilesOgay2007**), the theory of Discourse Communities (DC; Swales 1990) and of Communities of Practice (CoP; Wenger 1998), linguistic convergence is to be expected in contexts where individuals or groups frequently interact, as a way to construct group identity or as a way for one individual or group to gain social acceptance by the other. In DC and CP, linguistic convergence is theorised in terms of genre: communities develop structured linguistic repertoires or genres, made up of repeated linguistic patterns (Miller & Kelley 2016), that need to be acquired by new group members. Interestingly, MEPs have been described as a Discourse Community (Calzada-Pérez 2007) and the EP booths as Communities of Practice (Duflou 2016)[1]. In Defrancq (2018) and Defrancq & Plevoets (Forthcoming), a theoretical analysis of the European Parliament as a discourse community is put forward, that not only includes the MEPs, as proposed by Calzada-Pérez (2007), but also the interpreters in the EP booths.

To test the theoretical model, an analysis of lexical patterns used by Dutch-speaking MEPs and the Dutch booth in the EP was conducted (Defrancq & Plevoets Forthcoming) and output from both groups was compared with the output of Dutch-speaking members of national parliaments (for an identification of the items used, see §5). A Correspondence Analysis led us to conclude that there is indeed a degree of linguistic convergence between Dutch-speaking MEPs and the Dutch booth in the EP: members of national parliaments and the Dutch EP booth appear at the extreme ends of the linguistic spectrum, while MEPs position themselves in between. However, the case could not be made that MEPs and the Dutch booth constitute a single group from a linguistic point of view in opposition to national parliamentarians. Interestingly, it also appeared that the group of MEPs shows striking signs of internal convergence: diatopical variation (Belgian Dutch vs. Netherlandic Dutch) is considerably lower among MEPs than among members of national parliaments. MEPs thus seem to converge on the use of a hybrid variety that shares some properties with national Dutch-speaking parliamentarian registers and others with the EP's Dutch booth.

In Defrancq & Plevoets (Forthcoming) we refrained from claims about the direction of the observed convergence. However, all theories of communication that account for it are based on the idea that some individuals or groups are dominant, in that their linguistic repertoire tends to be emulated by other individuals

[1]Duflou considers language booths as separate communities of practice but all booths collectively as one too.

or groups and not the other way around. CAT holds that individuals and groups create, maintain or decrease social distance through linguistic, paralinguistic and non-verbal communicative strategies. Individuals or groups accommodate, i.e. shift to features that are more similar to the features of the other, in order to maximise social integration with the other individuals or groups, making the latter the dominant force in convergence. Similarly, in the theory of Discourse Communities and Communities of Practice, newcomers to the community are assumed to seek to assert their membership by proving their grasp of the community's specific genre, the dominant or expert group being the insiders that already have knowledge of the genre.

Bartłomiejczyk's (2016) above-mentioned quote seems to prioritise MEPs as the dominant or expert group in the genre makeup of the European Parliament: interpreters are reported to acquire lexical patterns from MEPs (and from more senior interpreters), but MEPs are not reported to acquire lexical patterns from interpreters. However, the Correspondence Analysis we presented in Defrancq & Plevoets (Forthcoming) seems, at first sight, to give some credit to the idea that MEPs adapt to linguistic patterns used by interpreters: MEPs combined position is situated between the positions of national MPs and the EP booth. The idea is not unreasonable: Dutch-speaking MEPs are likely to listen a fair amount of time to their interpreters, to be exposed to linguistic patterns interpreters use and are therefore also likely to adopt these patterns. As a result, they might position themselves closer to the interpreters than members of national parliaments who lack that kind of exposure. It is important to note that the idea of MEPs adapting to interpreters is not incompatible with Bartłomiejczyk's (2016) proposal. Bartłomiejczyk hypothesises a cross-linguistic pattern of accommodation mediated by translation, whereas our assumption relates to adaptation within one single language.

Alternative accounts for the observed convergence are possible: it could be argued that the MEPs are most representative of the EP genre and that the booth's outward position in the Correspondence Analysis reflects linguistic routines that are influenced by the challenging circumstances in which they produce output. As a group, MEPs present less diatopical variation, which is a clear sign of convergence taking place within that group. There is also evidence adduced by Ferraresi & Miličević (2017) from the Italian EP booth that suggests that interpreters' lexical patterns are less idiomatic than those of speakers of the same language and that this could be due to source text interference, cognitive load or a combination of both. Combined, this evidence appears to contradict the idea that interpreters may be the linguistically dominant group in the EP.

To be completely on the safe side, conclusions on linguistic convergence and the direction of convergence should be based on a longitudinal analysis of MEPs and their linguistic output. Unfortunately, the data collected for the corpus that was used for this study does not allow for such an analysis. We therefore propose to study convergence synchronically in terms of output features of different groups in the framework of linguistic theory that maps diachronic evolution to sychronic states (§5). Accordingly, the research questions of this study are the following:

- What are the profiles of Dutch-speaking MEPs and the Dutch booth in the EP in terms of seniority, exposure and output in the European Parliament? Seniority is an important variable in determining who is most likely to constitute the group with most experience in the EP genre. Exposure data is required to ascertain the possibility of accommodation of output features.
- Which group is the dominant or expert group in the EP in linguistic terms, i.e. is more likely to have shaped the features of the EP genre, while the other group is still in the process of acquiring those features?

The first question will be answered on the basis of an overview of the relevant literature on MEPs and interpreters, concentrating on the EP's Sixth and Seventh Term, i.e. the period between 2004 and 2014. The Sixth and Seventh Terms are the ones from which most of the data that we will use to answer the second question is drawn. The second question will be answered with a combination of quantitative and qualitative methods. As the corpora used for this study do not allow for diachronic analysis, linguistic expertise will have to be interpreted synchronically and comparatively. We will assume that the non-expert group has incomplete mastery of the genre: it is therefore unlikely to have acquired all linguistic features of the genre and likely to use the acquired features to a lesser extent than the dominant or expert group.

3 Members of the European Parliament

3.1 Seniority

The more than 700 MEPs are elected by universal suffrage according to rules laid down by the Member State's electoral authorities. During the Sixth Term and Seventh Term, which are directly relevant to this study, as most of our data in §5 were drawn from these, the EP consisted of 732 and 736 members respectively.

Electoral procedures vary across Member States. MEPs professional profiles also vary, but are predominantly situated in the legal and academic fields according to an analysis of biographies of MEPs in the Sixth term (2004–2009) by Beauvallet & Michon (2010). According to the same analysis, 81% are university graduates and 26% hold a PhD. Unlike in the early years of the EP, the EP mandate is for most MEPs (Sixth Term: 61%; Seventh Term: 66%) the first electoral mandate of their political career or the first mandate beyond the local level Beauvallet & Michon 2010; Beauvallet et al. 2013). Beauvallet & Michon (2010) conclude that the EP is a breeding ground for a new national political class as it offers most MEPs their first paid full-time job as a politician. Belgian and Dutch MEPs, who are directly relevant to our research, differ considerably: only 7% of Dutch MEPs in 2004 had previous experience beyond the local level, compared to 42% of the Belgian MEPs.

Roughly half (52% in 2004) of the MEPs are newly elected with each 5-year electoral cycle and 12% of them had left office and was replaced by newcomers before the end of the parliamentary term (Whitacker 2014). This means that a sizeable number of MEPs had limited experience on the job. In the Sixth term the average length of the EP stint was 6.6 years for the pre-2004 Member States and 6.3 and 5.7 years for Belgian and Dutch MEPs respectively (Beauvallet & Michon 2010).

3.2 Exposure

Even though membership of the EP does not require particular foreign language skills, linguistic competences are an asset in the EP. Among the 141 MEPs she interviewed, Wright (2007) quotes several of them pointing out that MEPs who do not master English as a *lingua franca* are likely to be marginalised in the political process. The EP offers simultaneous interpreting from 24 into 24 official languages during plenaries. For group and committee meetings interpretation is offered for the languages requested by participants.

To determine to what extent MEPs are exposed to interpretation we should be able to estimate how many of the contributions to plenaries, committee and group meetings are held in languages that they are unlikely to understand. No such data are available for committee and group meetings. The literature on plenaries provides us some clues, but caution is due in interpreting the figures. The most direct source of information are corpora of EP proceedings, such as Europarl (Koehn 2005). However, Europarl is built with the purpose of ensuring roughly equal numbers of data per language and does not reflect the proportions

of languages actually used. One Europarl sub-corpus, extracted from the 1996–1999 plenaries by Cartoni, Cartoni et al. (2013) reports corpus sizes for 5 of the then ten official languages which are claimed to reflect the actual language use. The three major languages, i.e. English, French and German each account for 25 to 29% of the data, while Dutch reaches 17% and Spanish and Italian 14 and 12% respectively. These figures do not include the other languages that were official at the time (Portuguese, Greek, Finnish, Danish and Swedish) and are therefore exaggerated. In a study based on data drawn from 62 plenaries in 2006 (with 21 official languages), Cucchi (2007) reports that English represents 21% of the data: 12% of native English and 9% of nonnative. Other languages are not differentiated. It is important to note that the data are calculated on the basis of token counts, which does not automatically translate to speech counts. Long speeches held in one of the languages will result in higher proportions in the token count. English in particular is mostly used by the Commission representative during the plenary, who is given more speaking time than MEPs.

Only a very rough estimate can therefore be given of the amount of time Dutch-speaking MEPs will seek interpreting. Considering most of them know English well enough to do without interpretation and a fair share of them also understand French or German well enough, MEPs are likely to resort to interpretation for slightly over half of the plenary speeches. It is important to note that English speeches are not necessarily listened to directly. Wright (2007) reports that on one occasion she noticed that a considerable number of MEPs put on their headphones when an Irish MEP took the floor in native English after an intervention in nonnative English by a German MEP.

3.3 Output

Plenary speaking time is allotted to the political groups in accordance with their numerical strength. Individual MEPs are granted speaking time at their request. Among the many factors that determine an individual's likelihood of being allowed to hold a speech, EP seniority is one of the most significant Slapin & Proksch (2010). MEPs are thus likely to have spent a fair amount of time listening to their colleagues, either directly or through interpretation, before preparing and holding speeches of their own.

Wright (2007) reports an array of different attitudes among MEPs with regard to speech preparation: while many members make a point of using their own official language, some do not object to or even prefer the use of a *lingua franca* i.e. English or French, in plenaries for fear of not getting their views across through interpreting, a point also made more generally by Kurz & Basel (2009). It is also

customary for MEPs to articulate a few words in the language of previous speakers to whom they respond.

According to Wright (2007) MEPs whose mother tongue is one of the *lingua francas* (English and French) split into two sub-groups. Some members consciously adapt to the presumed needs of a nonnative audience, focusing on clear articulation and avoiding rhetorical and linguistic prowesses; others do not seem to be bothered. Among English-speaking MEPs, the latter group tends to be strictly monolingual and is reported to have trouble understanding the non-native English used in the European institutions. French-speaking MEPs of the latter sort are likely to be upset by the widespread use of English and the diminished status of French.

MEPs' language patterns have drawn considerable interest, not least the rhetorical and metaphorical devices put forward to construct a European identity (De Angelis 2011, Fløttum 2013). The availability of the written verbatim reports of MEPs speeches and their translations in particular has sparked detailed studies of specific patterns. The Europarl corpus is still by far the biggest translation corpus in terms of language scope and sheer size.

Comparing an English sub-corpus derived from Europarl (2006 plenaries) with a corpus of English TED talks, Lefer & Grabar (2015) find that some categories of evaluative prefixes are typical of EP discourse. In particular, prefixes expressing excess or insufficiency (*over-centralised*; *under-represented*) are significantly more frequent in EP discourse. Granger (2014) compares a bilingual (French and English) 2-million tokens' sub-corpus of Europarl with a corpus of journal editorials, pointing at the high frequencies of lexical bundles performing EP rituals related to interaction during the plenary (e.g. thanking the President or congratulating a colleague); expressing epistemic stance (e.g. *I'm delighted that, I must say that, I am sure that*) and directive stance (e.g. *we want to see, we have to make sure that, we need to, we must not, we have a duty to, we have to ensure that, there is a need for/to*). Those lexical bundles are typical of MEPs' speeches but their frequency seems to vary across languages, as the English data show higher frequencies than the French data.

It has been pointed out that the verbatim reports are sanitised versions of MEPs' speeches (Cucchi 2009) and do not accurately reflect MEPs' linguistic patterns. Several small corpora of the spoken versions of speeches have been compiled, both with and without interpretations (Bernardini et al. 2018, Cucchi 2007; Kajzer-Wietrzny 2012, Russo et al. 2006). From an analysis of the transcribed speeches (N=62) in her corpus, Cucchi (2009) concludes that general extenders (*e.g. and so on*) are less frequent in MEPs' speeches than in ordinary conversa-

tion and that they are used as a way of referring to information only MEPs have access to and can complete, strengthening their institutional identity.

4 EP interpreters

4.1 Seniority

Interpreters in the European Parliament are recruited as staff interpreters or as freelancers through competitions and accreditation tests. Competitions have become increasingly rare in the last decade. Accreditation tests are organised according to need. For the Dutch booth, due to looming personnel shortages, accreditations tests have been held annually over the last 10 years, with the exception of 2020, due to the Covid-19 crisis. Since 2004 accreditation tests are organised jointly by the European Commission's DG Interpretation (SCIC) and the European Parliament's DG LINC (Duflou 2016). Success rates are traditionally low, ranging between 20 and 30% (Duflou 2016), meaning that the influx of new interpreters is limited. For the Dutch booth, on a pool of ca. 17 staff and ca. 70 freelancers during the Sixth Term (Duflou 2016)[2], only 1 to 3 new freelancers are accredited per year. This does not mean that new freelancers are immediately integrated in the EP's workforce. Duflou (2016) states that the EP prioritises freelancers on the basis of institutional loyalty in order to be able to recruit experienced interpreters offering the required language combinations:

> [N]ewly accredited interpreters are mainly recruited during peak periods, and only a few of them, depending on their language combination and the outcome of their evaluation reports, will be recruited to work for DG INTE regularly. Duflou (2016: 145)

Compared to the MEP workforce with a turnover of over 50%, the interpreter pool appears much more stable. In the non-representative sample of interviewees Duflou (2016) drew from the Dutch booths with DG Interpretation (SCIC) and DG LINC, 24 had started interpreting for the EU 4 years or more before the interviews that took place over the period 2007–2010. 7 of those had started before 1980. 11 had less than 4 years of experience. Based on these – admittedly – partial data, it seems nevertheless safe to assume that on the whole the pool of EP interpreters can boast substantially more collective experience in parliamentary meetings, including the plenaries, than MEPs.

[2]The number of staff members has decreased over the last years. Heines (p.c.), the acting booth head, confirmed that there were only 9 staff interpreters left in 2020.

4.2 Exposure

It might seem trivial to state that interpreters are continually exposed to the output of MEPs, considering that they have to interpret it. Two caveats are nevertheless in order here: first, the Dutch booth does not cover the whole range of languages spoken in the EP. Exact figures are hard to find and vary from one plenary to another, but with an average coverage of just under 5 languages among the staff interpreters and 3.7 languages among freelancers (Duflou 2016), the Dutch booth is unlikely to cover more than 10 languages in the plenary, even though it sits three interpreters. For all other languages, the Dutch booth resorts to relay interpreting, in which case they do not listen to the MEPs but tune in on other booths. Second, during the Sixth Term Belgium elected 24 MEPs, 14 of which were Dutch-speaking, the Netherlands elected 27; in the Seventh Term, the figures were the following: Belgium 22 MEPs (13 Dutch-speaking); the Netherlands 25 (CVCEU.eu s.d.). The group of potential speakers of Dutch consisted thus in both terms of ca. 40 members on a total of 732 to 736 MEPs, which is less than 6%. In other words, interpreters' potential exposure to Dutch spoken by MEPs is marginal. Moreover, it is likely that Dutch speeches are welcomed as periods of rest with little attention paid to what is said and how it is formulated. In all, the Dutch booth can be safely assumed to have only very limited exposure to Dutch-speaking MEPs.

Rather, exposure and accommodation to booth mates, especially the more senior ones, is pervasive and well-documented by Duflou (2016). Less experienced interpreters are expected to listen to more experienced colleagues and to copy their renderings to the extent that not doing so is regarded as a reason to put their competence into question. This results in the creation and maintenance over long periods of time of a joint repertoire of expressions. One DG SCIC interpreter Duflou (2016) interviewed finds this "parroting"(p. 194) highly problematic.

4.3 Output

The linguistic patterns in EP interpretations have been investigated both in comparison with the source speeches they are based on and with the translations made of the verbatim reports of those source speeches. The available research focuses on a limited number of features: the handling of pragmatically challenging utterances, such as face-threatening acts or ideologically laden lexemes (Beaton-Thome 2013; Bartłomiejczyk 2016; **Bartlomiejczyk2020**; Magnifico & Defrancq 2017), translation universals such as simplification (Russo et al. 2006; Kajzer-Wietrzny 2012; Ferraresi & Miličević 2017; Bernardini et al. 2016) and explicitation (Kajzer-Wietrzny 2012; Defrancq et al. 2015); collocations and formulaic

expressions (Ferraresi & Miličević 2017; Aston 2018) and ideological homogenisation (Beaton-Thome 2007). Explanatory factors for specific patterns are usually sought in the area of source language interference and heavy cognitive load interpreters experience, which is held to be conducive to simplification and even to explicitation (Ferraresi & Miličević 2017), and in the area of interpreters' agency required for navigating a pragmatically complex and treacherous context.

The use of formulaic expressions has also drawn interest. EP interpreters appear to use formulaic expressions very frequently (Aston 2018) due to a combination of factors. First, the context of institutionalised procedures is conducive to the use of formulaic expressions and interpreters working in this context are thus exposed to high frequencies of them (Bartłomiejczyk 2016). Second, it is widely recognised that producing formulae allows interpreters to reduce cognitive load as formulae are retrieved as complete units from memory (Gile 1995; Setton 2011, **PlevoetsDefrancq2018**). Finally, as already explained, formulaic expressions are also part of the socialisation process newly accredited interpreters go through: to blend in, they are expected to adopt expressions used by boothmates (Bartłomiejczyk 2016; Duflou 2016). Among items with unusually high frequencies in the English booth of the EP, Aston (2018) list the performative expressions having to do with parliamentary rituals, but also stance items, such as I think we need to, to come up with a, we need to ensure that, when it comes to the. Incidentally but perhaps not coincidentally, there is quite some overlap with the lexical bundles Granger (2014) reports as typical of the written versions of speeches held by English-speaking MEPs during plenaries. This seems to confirm the linguistic convergence reported in §2 for a different set of MEPs and interpreters.

4.4 Intermediary conclusions

Considering the available information on MEPs and interpreters in the EP, a number of tentative conclusions can be reached about the likelihood of one group being linguistically dominant or expert. Much of the evidence is circumstantial. Hard evidence could only have been collected with observational methods over a long period of time, which is not the case here. However, it seems relatively safe to conclude that the level of expertise in the EP genre is probably higher in the booth than in the plenary room: MEPs have higher turnover rates and their sting typically does not last very long. Exposure to same-language linguistic output is likely to be far higher in the case of MEPs: they need to listen more often to their interpreters than the other way around. There is cross-linguistic quantitative and qualitative evidence that linguistic convergence takes place between

MEPs and interpreters that share the same language. Given the relative expertise and the relative exposure, it seems more likely that interpreters constitute the expert group and that MEPs conform to interpreters in terms of language patterns than the other way around.

It should be pointed out that some of these conclusions do not necessarily apply to post-2004 accession booths that practice both A-interpreting and retour. Interpreters in these booth are probably more likely to pick up patterns from MEPs who are native in the interpreters' B languages and native MEPs are much less likely to adopt patterns from retour interpreting.

5 Analysis of patterns

5.1 Operationalisations and assumptions

As explained in §2, a diachronic process, i.e. a group acquiring specific linguistic features of a genre, will be analysed through the lens of essentially synchronic data. The corpus used is EPICG which contains on the one hand 27 Dutch speeches from the EP's Sixth Term and 16 from the Seventh Term, spanning 4 years of plenaries (2008–2011). 33 different MEPs are included, 6 of which are included with more than one speech. The corpus also comprises 164 Dutch interpretations by an unknown number of interpreters from the Sixth and Seventh term, spanning 6 years of plenaries (2006–2011). Only for the 6 MEPs with multiple speeches would it be possible to study the adoption of certain features through time. Obviously, this is too small a sample. As for the interpreters, a lack of metadata forbids any comparable analysis.

This is why the corpus will be analysed as a synchronic state, rather than as a diachronic process. Translating diachronic processes to synchronic states is not uncommon in linguistics. In the area of grammaticalisation theory, synchronic variation is considered to be a "manifestation of (diachronic) change" (Lehmann 2005). Crucial to the representation of grammaticalisation is the so-called "cline" (Hopper & Traugott 2003), which represents both the diachronic evolution of single items through different stages of grammaticalisation and the relative position of multiple items in a synchronic state, including in a cross-linguistic perspective.

A similar projection will be made here. The acquisition of expertise in the community genre can be represented as a cline (in Figure 2), that both captures the stages individuals go through diachronically and allows us to compare individuals and groups synchronically:

One aspect of expertise will be singled out here, namely the use of a particular set of lexical patterns that belong to the EP genre. Logically, the expert group is

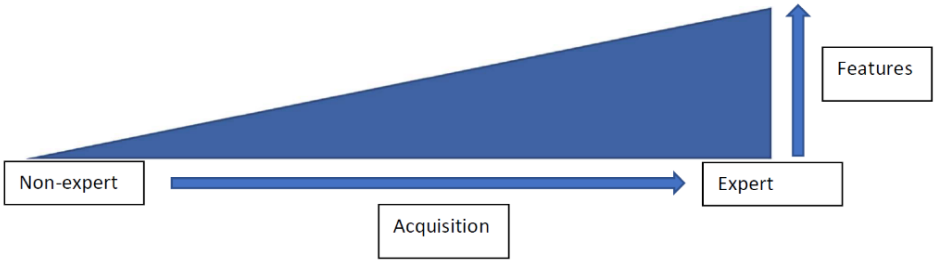


Figure 2: A cline of linguistic expertise

hypothesised to master more of these patterns than the non-expert or acquiring group and also to use those patterns more frequently than the latter.

One significant drawback of the EP data used in this study is that the linguistic features of the EP genre cannot be determined independently from the output of both groups under study. In addition, both groups are unequally represented in the data: interpretations add up to more than 70% of the EP data; MEPs to less than 30%. If linguistic features were to be extracted from the sum of the two data sets, interpretations would have a significant edge over MEPs speeches in determining the features of the EP genre.

To avoid bias, we will concentrate on the lexical patterns that both groups share as most typical of their respective outputs. Typicality will be determined through a keyness analysis and based on a comparison with the non-EP corpus, i.e. the corpus of speeches held in national parliaments.

Crucially, the expert group is assumed to master the EP genre more completely than the non-expert group and, therefore to use a broader range of typical patterns than the non-expert group. Consequently, we hypothesise that the patterns shared by both groups will make up a modest part of the patterns typical of the expert group, but a significant portion of the patterns typical of the non-expert group. We also hypothesise that the shared patterns will be used more frequently by the expert group.

5.2 Keyness in the EP sub-corpora

For the purpose of the Correspondence Analysis referred to in §2, we worked on a set of 181 3- and 4-gram types described in Defrancq & Plevoets (Forthcoming). This is also the set of lexical patterns that will be studied here. The 3- and 4-grams were selected from a set of frequent 269 types drawn from the Dutch sub-corpora in EPICG (Bernardini et al. 2018) and the CGNg (Oostdijk 2000), i.e.

a corpus of parliamentary speeches and debates in the Netherlands and Belgium, which is part of a larger corpus of spoken Dutch. The selection process is explained in Defrancq & Plevoets (2018). We excluded three types of items: syntactically ill-formed items (e.g. due to repetitions of the same word), items related to EU entities that could be considered as self-references (e.g. *verdrag van Lissabon* 'Lisbon treaty') and references to the debating context (e.g. *het woord is aan* 'has the floor'). Including the latter two categories would have artificially promoted the convergence hypothesis.

The 33 MEPs in our sample are Dutch and Dutch-speaking Belgians. They delivered the 43 speeches Dutch speeches contained in the sub-corpus. At the time the speeches were delivered MEPs had spent on average 85 months or around one and a half terms in the EP (one term is 60 months). Experience at the time of the speech ranges from 10 months up to 177.

The first step was to determine which 3- and 4-gram types were most typical of the Dutch speeches in the MEP sub-corpus (MEP). We therefore performed a comparative keyness analysis of these speeches with the data from the national parliaments (NAT). As our datasets are small, it is unadvisable to determine keyness based on significance tests (Likelihood ratio and Pearson chi-square test), as the results of such tests are impacted by the size of the available data (Gabrielatos 2018). Gabrielatos recommends the use of %DIFF and BIC for the comparison of frequencies in different corpora. %DIFF yields a measure of discrepancy between the relative frequencies of items, where high scores indicate large frequency differences. We set a threshold of 250 to select the items that are most key in MEPs speeches. The threshold was randomly chosen with an aim to obtain a set of approximately 25 items. The resulting list turned out to consist of 26 items, that can be found in Appendix A. BIC (Bayesian Information Criterion) is an alternative way of obtaining significance scores for keyness (Gabrielatos 2018). However, due to the small sizes of the sub-corpora, we found very few items that reached the significance threshold and made the choice to nevertheless proceed on the basis of the %DIFF scores.

The same procedure was repeated for the Dutch interpretations (INT). The resulting list contained 69 items and can be found in Appendix B. The longer list of key items in interpretations is not surprising: as our reference corpus is the sub-corpus of national parliamentary speeches, the longer list reflects the greater discrepancy between national parliamentarians and EP interpreters, confirming the outcomes of the CA in Defrancq & Plevoets (Forthcoming).

Crucially, at this stage we needed to check how many and which of the key 3- and 4-gram types in both sets were identical. This is shown in Table 1 and Appendices 1 and 2 (underlined items occur in both sets).

Table 1: Number of key items in the sub-corpora and their overlap. Keyness with regard to national parliaments.

%DIFF/NAT	MEP # (percentage shared)	INT # (percentage shared)	Shared between MEP and INT
>250	26 (65%)	69 (25%)	17

Interestingly, it turns out that almost two thirds of the items that are typical of Dutch MEP speeches are also among the key items in the Dutch booth. Conversely, only a quarter of the key items in interpretation are also key in Dutch MEP speeches. In other words, it not only appears that interpreters use a broader range of key 3- and 4-grams than MEPs, but that broader range also includes a significant portion of items that are key in MEPs' speeches. It therefore seems more likely that interpreters constitute the expert group in linguistic terms, while MEPs appear to be the group acquiring linguistic expertise. Additional food for this conclusion comes from the analysis of the nine items that are key in MEPs' speeches, but less so in interpretations. Of those nine, eight are still more frequent in the EP interpretations than in the national parliaments, four of which obtain a %DIFF score higher than 100. Conversely, of the 52 items reaching the keyness threshold in interpretations alone, only 17 are also more frequent in MEPs speeches than in national parliaments. In other words, all but one key items in MEPs speeches can be accounted for assuming they are adopted from interpreters, while not even half (17+17=34) of the key items in interpretations could be accounted for assuming these were adopted from MEPs. The data also contradict an alternative hypothesis in terms of interpreters' higher likelihood to use atypical patterns due to interference or cognitive constraints: if this were the case, the presence of so many of their key patterns in MEPs' speeches could not be accounted for.

Additionally, we compared the relative frequencies of the 17 shared items, assuming that these would be higher among the expert group. Table 2 shows that in all but two cases (underlined) the relative frequencies of key n-grams are indeed higher in the interpretations than among MEPs. Cases are shown according to their keyness score in the MEPs output (not shown here).

Due caution is needed in interpreting the figures because about half of the values in the MEP column represent 1 single occurrence in absolute numbers. Nonetheless, even in most of the remaining cases interpreters are found to use key items of the MEPs speeches even more frequently than the MEPs themselves. It is therefore reasonable to conclude that the interpreting booth in the EP is the

Table 2: Relative frequencies of key items.

	INT **Rel. Freq. /100k tokens**	**MEPs** **Rel. Freq. /100k tokens**
we moeten dus 'so we need to'	9.11	6.92
de veiligheid op 'the security on'	13.02	3.46
willen danken voor 'want to thank for'	6.51	3.46
van de veiligheid 'for the security'	13.02	10.39
de verenigde staten 'the United States'	9.11	38.08
we moeten niet 'we must not'	6.51	3.46
de bestrijding van 'the fight against'	7.81	6.92
en we moeten 'and we need'	22.13	10.39
ervoor zorgen dat 'make sure that'	22.13	10.39
in geval van 'in case of'	6.51	10.39
dus we moeten 'so we need to'	9.11	3.46
we moeten ook 'we also need to'	18.22	3.46
om ervoor te zorgen 'to make sure'	23.43	3.46
de bevoegdheden van 'the competences of'	9.11	6.92
ervoor te zorgen dat 'to make sure that'	24.73	6.92
van de gegevens 'of the data'	6.51	3.46
om te komen tot 'to reach'	9.11	3.46

expert group in linguistic terms, while MEPs show a lower degree of linguistic expertise. This is of course completely in line with the intermediary conclusions of §4.4. The positions of both groups can be set out against the expertise cline (Figure 3). Interpreters are represented by the straight cross and present the most typical use of the EP genre, while the group of MEPs, represented by the diagonal cross, has less expertise in the genre.

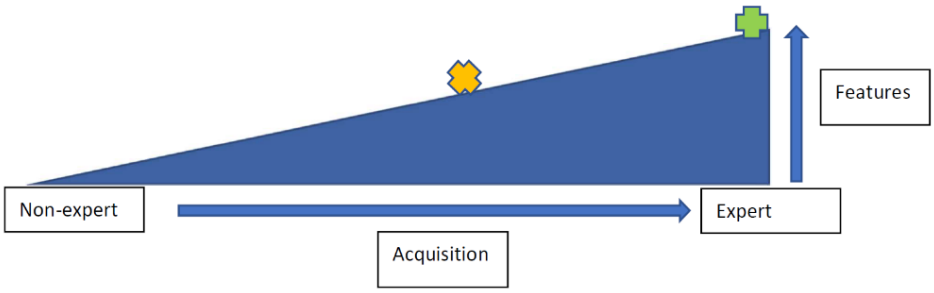


Figure 3: Positions of MEPs and interpreter on the linguistic expertise cline in the EP.

5.3 Functional analysis

A functional analysis carried out on the 17 items in Table 2 along the lines described in Biber (2004), reveals a number of interesting facts. Three of Biber's categories are present: referential n-grams, stance n-grams and a discourse organiser.

The discourse organiser is *in elk geval* ('anyway'), seemingly used to refute counter-arguments as irrelevant.

Six key n-grams are referential: *de veiligheid op, van de veiligheid, de verenigde staten, de bestrijding van, de bevoegdheden van, van de gegevens.* They represent topics covered by EU legislation, such as road safety, combating terrorism, data protection, international relations and institutional competences. It should be stressed that Dutch speeches and Dutch interpretations do not necessarily come from the same plenary sessions. It is pure coincidence that some topics were covered both in sessions from which speeches were downloaded and in sessions from which interpretations were drawn.

Ten key n-grams are stance expressions. They are exclusively attitudinal stance expressions of obligation and intention, clustering around verbs such as *moeten* ('need, have to, must'), *zorgen voor* ('make sure, ensure'), *komen tot* ('arrive at, reach'). Many of them occur with an adverbial connective or a conjunction (*en,*

dus, ook, om, dat).[3] The occurrence of such stance markers is plausible in a context of legislative procedure that is prescriptive in nature. What is distinctive of the EP is that a particular set of expressions is used very frequently to articulate such stance and that interpretation appears instrumental in promoting those expressions, including in MEPs speeches.

Interestingly, the clusters of stance expressions happen to be equivalents of some of the n-grams found to be typical of the written reports of English EP speeches and interpretations by the English booth. Table 3 lists the attitudinal stance markers reported in Granger (2014) for the written reports and in Aston (2018) for English interpretations, next to the ones from Table 2. (Parts between brackets are absent from the Dutch n-grams as these are in general shorter than the ones extracted by Aston and Granger.)

Table 3: Comparison of key n-grams across studies.

This study	**Granger (2014) English-speaking MEPs**	**Aston (2018) English booth**
we moeten dus, we moeten niet, en we moeten, dus we moeten, we moeten ook	we need to, we must not, we have a duty to, there is a need for/to	(as I think) we need to
ervoor zorgen dat, om ervoor te zorgen, ervoor te zorgen dat	(we) have to ensure that, (we) have to make sure that,	(we) need to ensure that
		when it comes to the
om te komen tot		to come up with a
	we want to see	

The cross-linguistic similarities clearly support the idea that the legislative purpose gears the EP genre towards expressions of intentional and deontic stance. However, it is impossible to deduce from the data presented by Aston (2018) and Granger (2014) which group uses the items involved the most. More research is

[3] One case (*willen danken voor* 'want to thank for') is probably connected to the ritual of thanking the President or another MEP. If that's the case, it should be withdrawn on the basis of the exclusion criteria mentioned in §5.1.

needed on the English items to substantiate our claim that interpreters shape the linguistic features of the genre in English as well.

6 Conclusions

In this study, starting from an observation made in earlier work (Defrancq 2018), we set out to determine which group, MEPs or interpreters, plays the determining role in the linguistic convergence that seems to take place in the European Parliament. Most prominent theories of socially determined linguistic change rest on the assumption that some individuals or groups adopt linguistic features typical of other, more dominant or experienced, individuals or groups. In order to find out which group was the most experienced in the EP, a two-pronged approach was taken. First, an analysis of EP seniority and potential linguistic exposure was conducted for MEPs and interpreters. It revealed that interpreters are probably more experienced in plenary dealings than MEPs, and are therefore also more likely to be experts in the EP genre and its linguistic features. Second, a detailed analysis was carried out on lexical patterns (3- and 4-grams) typical of the EP genre in Dutch, showing that, on the one hand, the Dutch booth uses a broader range of patterns with higher frequencies, and, crucially, that lexical patterns typically used by Dutch-speaking MEPs coincide to a very large extent with patterns used by the booth. Interpreters thus seem to shape aspects of the EP genre, which are to a certain extent adopted by MEPs. This supports Pöchhacker's (2005) interactant model in which all participants in an interpreter-mediated encounter are assumed to influence each other's communicative behaviour. However, our study shifts the traditional focus to interpreters influencing their audience.

A qualitative analysis showed that the overlapping patterns are related to topics covered by the EP plenaries and to intentional and deontic stance adopted by MEPs. Based on cross-linguistic similarities found in Granger (2014) and Aston (2018), we speculated that the stance category is promoted by the communicative purpose of the EP plenaries, i.e. produce legislation and that interpretation is instrumental in promoting a particular set of patterns to express that kind of stance, including in MEPs. It is possible that those patterns get promoted because they offer interpreters cognitive benefits: formulaic language is known to lower cognitive load.

Some of the limitations of the study have already been touched upon: the datasets they are based on are small. Larger datasets should be (compiled and) analysed to subtantiate our claims, preferably in several languages, as there is little reason to believe that the patterns we observed are language-specific (although the situation in post-2004 booths might differ). Another much needed

extension concerns the amount of exposure to speeches and interpretations outside the EP plenaries. MEPs are also exposed to interpretation in committee or political group meetings, but no data on these meetings have been collected so far. Committee meetings or political group meetings also place MEPs in a different context, with probably other types of interaction dynamics, which may also influence the linguistic features of their output. These different factors need to be explored to obtain a richer and more nuanced picture of the EP genre.

Appendix A Key items in MEPs speeches. Keyness with regard to national parliaments.

MEP/NAT	
misdaden_tegen_de_menselijkheid	in_geval_van
we_moeten_dus	de_huidige_situatie
de_veiligheid_op	in_elk_geval
willen_danken_voor	dus_we_moeten
van_de_veiligheid	we_moeten_ook
de_verenigde_staten	om_ervoor_te_zorgen
het_gebied_van	op_deze_manier
op_het_gebied_van	de_bevoegdheden_van
op_het_gebied	ervoor_te_zorgen_dat
we_moeten_niet	tot_nu_toe
de_bestrijding_van	van_de_gegevens
en_we_moeten	om_te_komen_tot
ervoor_zorgen_dat	

Appendix B Key items in interpretations. Keyness with regard to national parliaments.

INT/NAT	
de_veiligheid_op	om_te_komen_tot
van_de_mensenrechten	rekening_houden_met
van_doorslaggevend_belang	de_financiële_middelen
we_moeten_inderdaad	zien_we_dat
we_moeten_ervoor_zorgen	te_zorgen_dat
de_strategie_van	van_de_markt
veiligheid_op_de_weg	wil_ik_ook
we_moeten_dus	de_gevolgen_van
voor_de_patiënten	van_de_gegevens
de_mobiliteit_van	in_geval_van
de_veiligheid_in	van_de_bevolking
van_de_wereldgezondheidsorganisatie	voor_de_toekomst
willen_danken_voor	de_bevoegdheden_van
om_te_voldoen	de_verbetering_van
over_de_veiligheid	mannen_en_vrouwen
van_de_volksgezondheid	dat_weet_u
om_te_voldoen_aan	is_het_zo_dat
te_zorgen_voor	niet_alleen_maar
moeten_ervoor_zorgen	de_bescherming_van
op_de_weg	is_het_zo
moeten_ervoor_zorgen_dat	voor_het_feit
om_ervoor_te_zorgen	voor_het_feit_dat
we_moeten_ook	het_hebben_over
de_afgelopen_maanden	het_beleid_van
en_we_moeten	het_gaat_hier
ervoor_zorgen_dat	en_we_hebben
we_moeten_niet	ik_wil_ook
van_de_veiligheid	in_de_wereld
te_voldoen_aan	te_maken_met
ervoor_te_zorgen_dat	dan_wil_ik
het_arrest_van	in_verband_met
dus_we_moeten	te_komen_tot
de_bestrijding_van	het_principe_van
ervoor_te_zorgen	de_verenigde_staten
en_wij_willen	

References

Aston, Guy. 2018. Acquiring the Language of Interpreters: A Corpus-based Approach. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 83–96. Singapore: Springer.

Bartłomiejczyk, Magdalena. 2016. *Face threats in interpreting: A pragmatic study of plenary debates in the European Parliament.* Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Doctoral dissertation). https://wydawnictwo.us.edu.pl/sites/wydawnictwo.us.edu.pl/files/face_threats_in_interpreting_czw_st_e.pdf.

Bartłomiejczyk, Magdalena. 2017. The interpreters' visibility in the European Parliament. *Interpreting* 19(2). 159–185.

Beaton-Thome, Morven. 2007. Interpreted ideologies in institutional discourse. The case of the European Parliament. *The Translator* 13(2). 271–296. DOI: 10.1080/13556509.2007.10799241.

Beaton-Thome, Morven. 2013. What's in a word? Your 'enemy combatant' is my 'refugee'. The role of simultaneous interpreters in negotiating the lexis of Guantánamo in the European Parliament. *Journal of Language and Politics* 12(3). 378–399. DOI: https://doi.org/10.1075/jlp.12.3.04bea.

Beauvallet, Willy, Victor Lepaux & Sébastien Michon. 2013. Who are the MEPs? A statistical analysis of the backgrounds of members of European Parliament. *Etudes européennes* 14. 1–12.

Beauvallet, Willy & Sébastien Michon. 2010. L'institutionnalisation inachevée du Parlement européen: Hétérogénéité nationale, spécialisation du recrutement et autonomisation. *Politix* 89(1). 147–172.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC |Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 21–42. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-6199-8_2.

Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5 (1). 107–135. DOI: 10.1075/jhp.5.1.06bib.

Calzada-Pérez, Maria. 2007. *Transitivity in translating: The interdependence of texture and context*. Berlin: Peter Lang.

Cartoni, Bruno, Sandrine Zufferey & Thomas Meyer. 2013. Using the Europarl corpus for cross-linguistic research. In Marie-Aude Lefer & Vogeleer, Svetlana (eds.), *Interference and normalization in genre-controlled multilingual corpora* (Belgian Journal of Linguistics 27), 23–42. Amsterdam: Benjamins.

Cucchi, Costanza. 2007. An investigation of general extenders in a corpus of EU parliamentary debates. In *Proceedings of the Corpus Linguistics Conference CL2007, University of Birmingham, 27–-30 July*, 1–13. https://www.birmingham.ac . uk / Documents / college - artslaw / corpus / conference - archives / 2007 / 242Paper.pdf.

Cucchi, Costanza. 2009. Spoken discourse and identity in EU parliamentary debates. In Associazione italiana di anglistica. Congresso (ed.), *Forms of migration, migration of forms: Proceedings of the 23rd AIA Conference, Bari, 20-22 September 2007*, 449–465. Bari: Progredi.

De Angelis, Emma. 2011. The European Parliament's identity discourse and Eastern Europe, 1974-2004. *Journal of European Integration History* 17. 103–116.

Defrancq, Bart. 2018. The European Parliament as a discourse community : Its role in comparable analyses of data drawn from parallel interpreting corpora. *The Interpreters' newsletter* 23. 115–132. DOI: 10.13137/2421-714X/22401.

Defrancq, Bart & Koen Plevoets. 2018. Over-uh-load, filled pauses in compounds as a signal of cognitive load. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 43–64. Singapore: Springer.

Defrancq, Bart & Koen Plevoets. Forthcoming. Linguistic convergence in the European Parliament: A correspondence analysis of N-grams used by members of parliament and interpreters. In Sandra L. Halverson, Jun Pan & Jeremy Munday (eds.), *Translating and interpreting political discourse*. Amsterdam: Brill.

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies*, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.

Diriker, Ebru. 2004. *De-/Re-contextualizing conference interpreting*. Amsterdam: Benjamins.

Duflou, Veerle. 2016. *Be(com)ing a conference interpreter*. Amsterdam: Benjamins.

Ferraresi, Adriano & Maja Miličević. 2017. Phraseological patterns in interpreting and translation. Similar or different? In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies. New methodological*

and theoretical traditions, 157–182. Berlin: Mouton/De Gruyter. DOI: 10.1515/9783110459586-006.

Fløttum, Kjersti. 2013. *Speaking of Europe: Approaches to complexity in European political discourse*. Amsterdam: Benjamins.

Gabrielatos, Costas. 2018. Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor & Anna Marchi (eds.), *Corpus approaches to discourse: A critical review*, 225–258. Abingdon: Routledge.

Gile, Daniel. 1995. *Regards sur la recherche en interprétation de conférence*. Lille: Presses Universitaires de Lille.

Giles, Howard. 1973. Accent mobility: A model and some data. *Anthropological Linguistics* 15. 87–109.

Granger, Sylviane. 2014. A lexical bundle approach to comparing languages. Stems in English and French. In *Genre- and register-related discourse features in contrast. Special issue of languages in contrast*, vol. 14, 58–72.

Henriksen, Line. 2007. The song in the booth: Formulaic interpreting and oral textualisation. *Interpreting* 9i(1). 1–20.

Hopper, Paul & Elizabeth Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu. (Doctoral dissertation). https://repozytorium.amu.edu.pl/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, vol. 5, 79–86. Phuket: AAMT. http://mt-archive.info/MTS-2005-Koehn.pdf.

Kurz, Ingrid. 2001. Conference interpreting: Quality in the ears of the user. *Meta* 46 (2). 394–409.

Kurz, Ingrid & Elvira Basel. 2009. The impact of non-native English on information transfer in simultaneous interpretation. *Forum* 7 (2). 187–213.

Lefer, Marie-Aude & Nathalie Grabar. 2015. Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures : a multidisciplinary journal for translation and interpreting studies* 16(2). 187–208. DOI: 10.1556/084.2015.16.2.3.

Lehmann, Christian. 2005. Theory and method in grammaticalization. *Zeitschrift für Germanistische Linguistik* 32 (2). 152–187.

Magnifico, Cédric & Bart Defrancq. 2017. Hedges in conference interpreting: The role of gender. *Interpreting* 19(1). 21–46.

Miller, Carolyn & Ashley Kelley. 2016. Discourse genres. In Andrea Rocci & Louis de Sassure (eds.), *Verbal communication*, 269–286. Berlin: Mouton/De Gruyter.

Monacelli, Claudia. 2009. *Self-preservation in simultaneous interpreting. Surviving the role*. Amsterdam: Benjamins.

Nord, Christiane. 2013. Functionalist translation studies. In Millán, Carmen & Batrina, Francesca (eds.), *The Routledge handbook of translation studies*, 201–212. London: Routledge.

Oostdijk, Nelleke. 2000. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, 887–894. Paris: ELRA.

Pöchhacker, Franz. 2005. From operation to action: Process-orientation in interpreting studies. *Meta* 50 (2). 682–695.

Pöchhacker, Franz. 1994. *Simultandolmetschen als complexes Handeln*. Tübingen: Narr.

Russo, Mariachiara, Annalisa Sandrelli & Claudio Bendazzoli. 2006. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches extended analysis of EPIC. *Forum* 4 (1). 221–254.

Setton, Robin. 2011. Corpus-based interpretation studies (CIS): Reflections and prospects. In Arlet Kruger, Kim Wallmach & Jeremy Munday (eds.), *Corpus-based translation studies: Research and applications*, 33–75. London; New York: Continuum.

Slapin, Jonathan & Sven-Oliver Proksch. 2010. Look who's talking: Parliamentary debate in the European Union. *European Union Politics* 11 (3). 333–357.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Wenger, Etienne. 1998. *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

Whitacker, Richard. 2014. Tenure, turnover and careers in the European Parliament: MEPss as policy-seekers. *Journal of European Public Policy* 21 (10). 1509–1527.

Wright, Sue. 2007. English in the European Parliament: MEPs and their language repertoires. *Sociolinguistica Jahrbuch* 21. 151–165.

Chapter 2

Formality in mediated and non-mediated discourse: Bringing together human judgements and corpus-driven detection

Ilmari Ivaska[a], Adriano Ferraresi[b] & Marta Kajzer-Wietrzny[c]
[a]University of Turku [b]University of Bologna [c]Adam Mickiewicz University

Several works have suggested that both interpreting and translation tend to favour linguistic choices typically considered more formal. Observations concerning the degree of formality of these forms of language mediation, however, have been made within studies mostly focusing on different phenomena, such as standardization or conventionalization. In this paper, we report on a quantitative and qualitative analysis focusing specifically on formality in interpreted language. We take into account native and interpreted speeches delivered at the European Parliament (EP) and collected in the English subcorpora of the EPIC and EPTIC corpora. We compare the examined English varieties to one another, taking into account the mode of delivery of the speeches, i.e. whether they were read out from a written text or delivered impromptu. Our hypothesis is that interpretations of speeches read at the EP are located at the far end of the formality spectrum from speeches delivered impromptu by native English MEPs. Unlike previous work, we rely on formality indicators identified by triangulating human judgements and specific linguistic features derived bottom-up from a corpus, based on the MuPDAR[F] approach. The analysis provides partial support for the hypothesis, showing that interpreted texts are generally predicted as being characterised by a high level of formality, irrespective of the actual mode of delivery of their source text. We conclude by commenting on the linguistic features that were found to contribute the most to making interpreted speeches diverge from native ones.

Ilmari Ivaska, Adriano Ferraresi & Marta Kajzer-Wietrzny. 2022. Formality in mediated and non-mediated discourse: Bringing together human judgements and corpus-driven detection. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 29–61. Berlin: Language Science Press. DOI: ??

1 Introduction

As pointed out by Heylighen & Dewaele (1999), all speakers are likely to intuitively distinguish between formal and informal registers, whereby the prototypical formal end of the spectrum resembles the language used by a judge during a trial, and the prototypical informal end is marked by a relaxed conversation among friends. Graesser et al. (2014: 218) associate formality with the need to be "precise, coherent, articulate, and convincing to an educated audience", as opposed to informal settings such as oral conversation, which is "narrative, replete with embodiment words" and reliant on common background knowledge. Andrén et al. (2010: 224) link formality with the application of "officially standardized and recognized institutional conventions or prescriptions".

Although research on formality per se in interpreting and translation studies is still scarce, a lot of attention has been directed towards the related notions of standardization (Toury 1995) and conventionalization (Baker 1993). Thus, it has been suggested that translations tend to shun informal language use, e.g. by relying on "generally unmarked grammar clichés, and typical, common lexis instead of the unusual or the unique" (Mauranen 2008: 41). In turn, this tendency may be related to risk-aversion (Pym 2005), whereby, to transfer the source meaning, translators, and by extension interpreters, opt for conventional linguistic forms.

In this paper we report on a quantitative and qualitative analysis of formality of native and interpreted speeches delivered at the European Parliament and collected in the English subcorpora of the European Parliament Interpreting Corpus, or EPIC (Sandrelli et al. 2010), and the European Parliament Translation and Interpreting Corpus, or EPTIC (Ferraresi & Bernardini 2019). In the reported study we compare the examined English varieties to one another, further taking into account the mode of delivery of the speeches, i.e. whether they were read out from a written text or delivered impromptu. Our working hypothesis is that interpretations of speeches read at the European Parliament are located at the far end of the formality spectrum from speeches delivered impromptu by native English MEPs.

Our approach to analysing formality has been inspired by Jarvis' work on lexical diversity, particularly on the observation that while lexical diversity is an emic construct, in that "it relies crucially on the human interpretation of both form and meaning" (Jarvis 2017: 540), it has been traditionally studied following an etic approach, "concentrat[ing] on the identification, measurement, and description of forms and their distribution in a way that does not rely on meaning" (ibid.). In line with Jarvis' thinking, we find that formality, too, is a profoundly emic construct, and so any potential automated measurement should be based

on a holistic, human-informed definition of the construct (Jarvis 2013a; Jarvis 2013b; Jarvis 2017). Hence, in the reported study we rely on formality indicators identified through a data-driven, human-informed operationalization of linguistic formality.

We start with an overview of studies touching on issues of formality in interpreting and translation (§2), and then move on to a detailed description of the method that allowed us to, first, identify the linguistic features characteristic of texts that have been classified as formal by humans (§3), and then to discuss our results (§4). In §5, we conclude by summarizing our main results and suggesting ways in which the results of the present study could be validated and applied.

2 Formality in interpreting and translation studies

To the best of our knowledge, formality has so far not been the focal point of any analysis in Interpreting and Translation Studies, but rather a feature commented on in the context of studies on interpreting and translation investigating other linguistic phenomena. Since formality has so far merited mostly a mention in these papers, we decided to include both interpreting and translation in our literature overview to gain a better perspective on formality as a phenomenon in mediated discourse in general. This will be used as a backdrop for our findings regarding formality specifically in interpreting.

While no studies focus explicitly on formality in simultaneous interpreting, several reports mention features pointing to greater formality in this type of communication. Kajzer-Wietrzny (2012) compares the frequency of the optional connective "that" in translations and interpretations into English from Romance and Germanic languages, as well as spoken and written native English speeches. Findings point to the connective being more frequent in both spoken and written mediated texts. As *that*-omission might be related to informality (Biber 1995: 145), higher frequencies of the optional connective can be interpreted as a manifestation of formality. Moreover, in an interview preceding a case study of interpreting style, one of the interpreters reported using more formal language and higher register in his interpretations than in his non-interpreted communication (Kajzer-Wietrzny 2012).

From a completely different perspective, in an analysis of face-threatening acts during plenaries at the European Parliament, Bartłomiejczyk (2016) points to cases where interpreters made the target rendition less face-threatening and "considerably more formal by consistently avoiding any colloquial vocabulary" (Bartłomiejczyk 2016: 203) and by "using more formal and euphemistic language"

(Bartłomiejczyk 2016: 211). It follows that formality in interpreting might also be a by-product of strategies mitigating impoliteness.

Hale (1997: 57) found that interpreters working at court hearings, instead of approximating the register of the speaker, "by and large adapt their communicative style to what they perceive to be the expectations and/or limitations of the listener", raising or lowering the level of formality depending on the interpreting direction. She reports that the level of formality is raised by strategies such as the condensation of the source text and the omission of typically conversational features such as fillers, hesitations, repetitions and backtracking. On the other hand, the level of formality is lowered especially through lexical choices, e.g. translating a formal word with a colloquial item, or adding semantically empty words or phrases of pragmatic importance signaling "a certain level of familiarity" (Hale 1997: 47–50). Such changes of communicative style are key "in forming impressions", which in the context of court interpreting has an impact on the perception of the witness in court (Hale 1997: 53).

Another observation, potentially pertinent to the issue of formality in interpreting comes from a study carried out by Shlesinger (1989). She observed an equalizing effect in interpreting, whereby "the interpretation of texts which exhibited typically literate features" tended to be "shifted towards the oral end of the continuum, whereas the interpretation of texts that exhibited typically oral features shifted towards the literate end" (Shlesinger 1989 in Shlesinger & Ordan 2012: 54). Assuming that oral vs. literate features correspond to various degrees of formality, such findings stress the need to further investigate formality in interpreting.

Preference for formality also emerges from studies on written translation. In her study on the occurrence of contractions in literary translation and contemporary literary English writing, Olohan (2003) found differences "in terms of both variety of contracted forms encountered and frequency of occurrence of contractions" (Olohan 2003: 59); together with *that*-omissions, these can be seen as a "crude measure of informality" (Olohan 2004: 101). She observes variation between translators, which might depend on the source language, on the level of formality of the source text, on translator's style, and/or genre and narrative structure of the text. It also follows from the discussion of results that texts in the analysed translation corpus display more features typical of Biber's (1988) informational production, i.e. they are less involved, less generalised, more explicit and more edited.

Translations have also been found to display a more "unmarked formal register and a neutral standard variety of the language" (Moe 2010: 125). Looking at various types of register shifts, Moe (2010: 136) concludes that translators "shift

the style from the extremes towards neutrality", and that shifts towards increased formality constitute the largest number of shifts in all analysed texts. Such shifts of register, she argues, may make the text less appealing for the reader.

One of the few works that have looked closely at formality in translation is the study by De Sutter et al. (2012: 343), who use profile-based correspondence analysis and logistic regression to assess whether translations use more formal language than other texts. The first method allows the authors to measure and visualize the linguistic distances between the language varieties. Logistic regression makes it then possible to evaluate "the exact impact of the lects on the lexical choices" (De Sutter et al. 2012: 325). The authors select 10 lexical variables, consisting of a neutral and a formal variant of semantically equivalent expressions, and use them to explore formality distances between texts of different genres originally written in Dutch, and Dutch translations from either French or English. They conclude that translated texts and non-translated texts differ with respect to formality, but also that translations are not uniform: translations from French are more formal than those from English. Furthermore, text type turns out to be the most important variable. Interestingly, text types differ with respect to formality but "[t]here appears to be no formality differences between translated and non-translated" texts within specific text types, such as journalistic texts, non-fiction and instructions (De Sutter et al. 2012: 340). It must be noted though, that according to the authors "formality variation cannot be predicted successfully" on the basis of text types and source languages only, since the reported regression model explains 18% of the variation.

Comments on formality also emerge from studies on constrained language, where translation is frequently set against native and non-native language varieties. Kruger & van Rooy (2018: 237) observe that "non-native varieties and translated texts avoid informality features in written registers" and relate it to the authors' level of language proficiency. Less proficient users are more prone to risk avoidance. On the other hand, the tendency to increased formality is one of the shared features of two constrained varieties, i.e. translated and non-native indigenised varieties of English that distinguish them from the native variety (Kruger & van Rooy 2016: 26). This transpires from several shared tendencies, such as a greater mean word length than in the native varieties, lower frequency of the pronoun "it", less frequent use of emphatics (e.g. *really*, *for sure*, *a lot*), avoidance of possibility modals (e.g. *can*, *could*, *may*, *might*), avoidance of stranded prepositions, higher frequency of nominalizations and increased frequency of *wh*-questions (Kruger & van Rooy 2016: 37–41). Other features pointing to increased formality are observed in translation only (not in non-native varieties), such as the use of optional connective *that* or higher frequency of nouns.

As this overview should have shown, many side-remarks are found in various studies which suggest that both interpreting and translation seem to favour linguistic choices typically considered more formal. With a few exceptions (e.g. De Sutter et al. 2012), these indications still need to be tested in a study devoted specifically to formality.

3 Data and method

To test whether interpreting is more formal than non-mediated native spoken discourse, a general operationalization of formality is needed. Following in part the methodological architecture of Jarvis (Jarvis 2017: 548––549), we adopt a study design combining corpus-derived and human informant data. Specifically, focusing on speeches delivered at the European Parliament as a case in point, we: 1) ask human judges to evaluate a set of non-interpreted, native data in terms of their perceived overall formality; 2) train statistical classifiers on a different set of comparable non-interpreted native data to evaluate their formality, using a range of potentially relevant, linguistically defined featuresets; 3) use the trained classifiers to predict the formality of the human-evaluated data, and zoom in on the linguistic features that contribute most to the successful classification; 4) use this final model to analyse a set of comparable interpreted texts, so as to assess how they are positioned with respect to the non-interpreted texts in terms of formality features. Each of these steps and the associated datasets are described in the following sections.

3.1 Data

3.1.1 The European Parliament (Translation and) Interpreting Corpora

The corpus data analyzed in this study come from two related corpora, i.e. the European Parliament Interpreting Corpus (EPIC, Sandrelli et al. 2010) and the European Parliament Translation and Interpreting Corpus (Ferraresi & Bernardini 2019). Both of them are multilingual corpora of speeches given in the plenary sessions of the European Parliament, but while EPIC only includes transcriptions of original speeches and their interpretations, EPTIC also features the corresponding written-up versions of the original speeches (so called "verbatim reports") and their written translations. The languages currently represented in the corpora, as sources, targets or both of interpreted and translated texts, are English, Italian and Spanish (EPIC), and English, French, Italian, Polish and Slovene (EPTIC), for a total of around 580,000 tokens.

All the data used in this study are in English, and are extracted from the spoken component of the corpora, comprised of non-interpreted, original speeches (henceforth called "original"), and interpreted ones. The dataset can be divided into three subsets: (1) native original train data, (2) native original test data, (3) interpreted data with French, Italian, and Polish as source languages. Speeches were selected so as to provide a balanced dataset with respect to their mode of delivery, i.e. whether they were originally delivered impromptu or read out, a distinction which we hypothesize to be associated with formality differences (see §3.2.1). It should be noticed that in the case of interpreted texts, the mode of delivery refers to the original speech.

The transcripts follow the audio recordings, but given that the purpose of the present paper is to detect and analyse characteristic features of formality and not general differences between spoken and written language, we have excluded from the transcripts elements that are exclusive to spoken language, including hesitations, false starts, as well as empty and filled pauses (editing guidelines can be found in Appendix A). As suggested by one reviewer, such orality features might contribute to the perception of formality, yet their elimination was necessary so as to avoid excessive weight being assigned to them both in the human and the automatic evaluation task.

The texts were then parsed according to the Universal Dependencies (UD) annotation scheme using the Turku neural parser (Kanerva et al. 2018), and the subsequent analyses were conducted using the parsed data. Table 1 gives an overview of the corpus data used.

Table 1: Analysed dataset.

	Native train	**Native test**	**Interpreted fr>en**	**Interpreted it>en**	**Interpreted pl>en**
Impromptu	30 texts	10 texts	10 texts	5 texts	10 texts
	6,573 w	1,797 w	1,940 w	821 w	1,607 w
Read out	30 texts	10 texts	10 texts	5 texts	10 texts
	7,566 w	2,045 w	1,825 w	912 w	1,506 w
All	60 texts	20 texts		100 texts	
	14,139 w	3,842 w		8,611 w	

3.2 Method

3.2.1 Rationale and research questions: grounding formality in human perception

The definitions of formality have often been functional or contextual. By way of example, the notion of "formal" is defined by Atkinson (1982) as the opposite of "conversational" and by Andrén et al. (2010) as adherence to institutional conventions, while Heylighen & Dewaele (2002) refer to high and low contexts (for a detailed discussion, see Li et al. 2016). Methodologically, it is noteworthy that definitions of the theoretical construct of formality are often emic in nature, in that they rely on "human interpretation of both form and meaning" (Jarvis 2017: 540), taking holistically into account both function and form.

The operationalizations of formality, however, have traditionally been linked exclusively to form. Several scores based on linguistic indicators have been proposed to measure formality in texts, including the formality score, which takes into account the frequencies of various word classes (Heylighen & Dewaele 2002), the adjective density score (Fang & Cao 2009), or lists of formal and informal words created ad hoc (Abu Sheikha & Inkpen 2010). Within corpus-driven approaches, formality has also been used as an explanatory tool when interpreting differences across texts from different categories (e.g. registers, genres or varieties). For instance, in the wealth of studies making use of Multidimensional analysis (MD), formality is often used to explain the nature of linguistic differences observed between registers (e.g. Biber 1988; Conrad & Biber 2001; Biber 2012). In other words, it is interpreted as the reason why certain categories diverge from one another, as attested by differences in frequencies of linguistic elements. In both scenarios, the link between formality and the linguistic features that attest to it is an indirect one, inasmuch as assessments of (the degree of) formality have typically not been rooted in human perceptions: in Jarvis' terms, they are etic operationalizations.

In this paper, we explore a data-driven and human-informed operationalization of linguistic formality for European Parliament data. In the first part of the analysis, we take as our point of departure two text classes that, we hypothesize, are characterized by different levels of formality, i.e. speeches delivered impromptu (less formal) vs. read out (more formal). We then establish their order of perceived formality by means of human judgements, and use these judgments as a gold standard in a data-driven analysis of linguistic features distinguishing the more formal vs. less formal text class. As we are interested in potential differences between interpreted and non-interpreted communication, we explore a

range of linguistic features that have earlier been used to distinguish mediated from non-mediated language use (Volansky et al. 2015).

Our specific research questions are: 1) are read out speeches perceived as more formal than impromptu speeches? 2) Which linguistic features contribute to distinguishing these text classes? 3) Do interpretations differ from spoken native non-interpreted texts in terms of formality, as assessed by the linguistic features thus identified?

3.2.2 Human evaluations

Our aim was to follow the emic approach and the example set by Li et al. (2016), and so we wanted to root our text-based operationalization of formality – the distinction between speeches delivered impromptu and those read out loud – on human judgements. More specifically, we wanted to assess whether, and to what degree, human evaluators agree on formality differences between texts when they are not given any linguistic definition of formality itself.

In total, 55 individuals participated in an online survey where they were shown ten pairs of texts, and were asked to choose for each pair which text was more formal in their opinion. There was one impromptu text and one read out text in each pair, and to minimize any test effect, the pairs were assigned randomly, so that each impromptu text was paired at least once with every read out text, and vice versa. The order in which the texts were introduced, as well as their order in the display, changed randomly for each participant. Participants were not given any definition of formality in the beginning, but they were told where the texts stem from. At the end of the survey, they were also asked to elaborate on their notion of formality (the questionnaire form can be found in Appendix B). The survey was circulated among the authors' colleagues and students, and not advertised in any other way. It was conducted anonymously on the Webropol online survey platform (version 3.0[1]) in 2019. The texts evaluated in this way constitute the test set of our corpus data.

According to the background information provided, the first language of 53 participants out of 55 is other than English. Altogether, 48 participants reported having a higher education background in languages, linguistics, translation and/or interpretating, and all but two agreed that they feel comfortable studying or working in English. The age range of informants were: 18–25 years (27), 26–35 years (14), 36–45 years (10) and more than 45 years (4). One participant did not complete the survey and their answers were left out of the results.

[1] https://webropol.com

3.2.3 Featuresets considered

Our ultimate research question was to see whether interpreted language diverges from non-interpreted language in terms of formality, and to explore the linguistic features that contribute to this potential difference. Hence, we decided to compare a range of different featuresets which have been shown to consistently distinguish mediated from non-mediated language in general, and to assess which ones could also be related to formality differences.[2] As all the included featuresets were implemented using parsed and CONLL-U formatted data,[3] they are easily transferrable to data on any language where sufficient UD resources are available.

Various kinds of sequential n-grams are probably the most widely used featuresets (e.g. Baroni & Bernardini 2006; Koppel & Ordan 2011; Volansky et al. 2015). We, too, used normalized frequencies of unigrams, bigrams and trigrams of words, lemmas, parts-of-speech, as well as syntactic functions, for a total of 12 featuresets. For instance, example (1). consists of six unigrams (*I*, *want*, *to*, *ask*, *some*, *questions*), five bigrams (*I want*, *want to*, *to ask*, *ask some*, *some question*") and four trigrams (*I want to*, *want to ask*, *to ask some*, *ask some questions*), each of which is represented in the four different levels of annotation (word, lemma, POS, syntax).

(1)

I	want	to	ask	some	questions	(word)
I	WANT	TO	ASK	SOME	QUESTION	(lemma)
PRON	VERB	PART	VERB	DET	NOUN	(POS)
nsubj	root	mark	xcomp	det	obj	(syntax)

Positional tokens, that is, starts and ends of sentences, are other features that have been used to distinguish mediated from non-mediated texts (e.g. Volansky et al. 2015; Rabinovich et al. 2016). To that end, we considered first, second, penultimate and ultimate positions in sentences to see how often different items occur in these positions. Here, too, we used four parallel featuresets, one for each level of annotation. In the case of example (1), we looked at how often *I/I/PRON/nsubj* occurred in the first position of the sentence, *want*/WANT/*VERB/root* in the second position, and so on. The sentence boundaries stem from the original data structure of EPTIC.

[2] All the frequency data as well as the R scripts of the statistical analyses can be found here: https://osf.io/q75jw/

[3] https://universaldependencies.org/format.html

Character trigrams and other character n-grams have been shown to work well when distinguishing mediated from non-mediated texts (e.g. Popescu 2011; Volansky et al. 2015). However, as pointed out by Volansky et al. (2015: 113), character-based features are difficult to interpret in a linguistically meaningful manner, especially when shorter n-grams are considered. Due to this difficulty, we opted for a compromise solution which allowed us to include this feature without totally sacrificing interpretability, we decided to focus exclusively on the trigram level and to limit the focus to individual words and their boundaries. For instance, the sequence *I want* consists of the character trigrams *_I_*, *_wa*, *wan*, *ant* and *nt_*.

More recently, dependency bigrams have been introduced as reliable, scalable and yet linguistically interpretable features when distinguishing mediated from non-mediated texts (Ivaska & Bernardini 2020; Ivaska et al. In print). Unlike typical POS bigrams, dependency bigrams are not necessarily sequential and provide information on the constituent words/lemmas/POS, their order in text, as well the nature of the syntactic relation linking them.

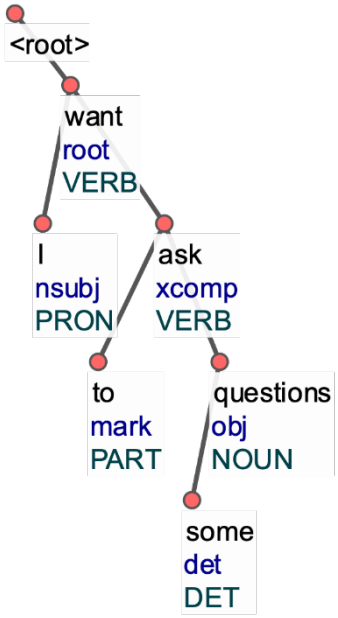


Figure 1: Tree representation of example 1.

For instance, example (1) (visualized in Figure 1) consists of the sentence *I want to ask some questions*, split into the following dependency word bigrams: *I*NODE_nsubj_*want*HEAD, *want*HEAD_xcomp_*ask*NODE, *to*NODE_mark_*ask*HEAD,

*ask*HEAD_obj_*questions*NODE, *some*NODE_det_*questions*HEAD). We used three dependency bigram featuresets, defined on the word level, the lemma level, and the POS level.

3.2.4 Feature selection and model training

For each of the 20 featuresets considered in the train dataset we conducted a feature selection procedure. First, following the example of Volansky et al. (2015), we limited each featureset to the 300 most common features. Then, to tease apart those features that actually contribute to the classification task on text formality, we conducted a Boruta feature selection (Kursa & Rudnicki 2010) for each featureset. According to Kursa & Rudnicki (2010) feature selection is helpful in predictive model building, as modern datasets are frequently rich in irrelevant variables that may decrease models' accuracy. Hence selecting a "small (possibly minimal) feature set giving best possible classification results is desirable for practical reasons" (ibid). Boruta helps limiting the dataset to only the most relevant variables.

Boruta introduces randomness to the data by duplicating all variables and randomly permutating the duplicates' values (here, feature frequencies). It then makes use of the random forest algorithm (Breiman 2001) and builds a classification model for the task at hand (here, the identification of texts presumably characterized by different formality), compares the actual features' performance to the randomized features, and suggests as important only those features that consistently outperform the randomized duplicates. Random forest was chosen as the statistical method, as it has originally been created to solve issues related to data including few observations with many predictors (ibid.), much like ours. The size of our final featuresets is summarized in Table 2.[4]

We then trained separate forest-based classifiers for each featureset using the same train dataset consisting of impromptu and read out files (labelled *impromptu_001*, *read_001* etc.) but only the selected features as predictors. We used the *ranger* implementation of random forests throughout the analyses (Wright & Ziegler 2017) and trained a prediction model for each featureset. We trained the models with ranger's probability function to obtain the likelihood of each prediction instead of just the most likely label. Whenever the classification is discussed in terms of the predicted labels, we have used 0.5 as the cutting point.

[4]Due to the relatively small train dataset, the word trigram featureset ended up being too sparse for reliable feature selection, and it has thus been left out of all the subsequent analyses.

Table 2: The featuresets used in the formality classification task.

Category	Featureset	Final number of features (out of considered)
Sequential n-gram	word 1-gram	12 (of 300)
	lemma 1-gram	14 (of 300)
	POS 1-gram	3 (of 15)
	syntax 1-gram	7 (of 40)
	sequential word 2-gram	11 (of 300)
	sequential lemma 2-gram	6 (of 300)
	sequential POS 2-gram	11 (of 187)
	sequential syntax 2-gram	13 (of 300)
	sequential word 3-gram	NA
	sequential lemma 3-gram	6 (of 300)
	sequential POS 3-gram	12 (of 300)
	sequential syntax 3-gram	14 (of 300)
Positional frequencies	positional word	7 (of 300)
	positional lemma	7 (of 300)
	positional POS	10 (of 55)
	positional syntax	11 (of 97)
Character n-gram	character 3-gram	18 (of 300)
Dependency n-gram	dependency word 2-gram	7 (of 300)
	dependency lemma 2-gram	9 (of 300)
	dependency POS 2-gram	9 (of 300)

3.2.5 Model validation and MuPDAR[F] analysis

The rest of the analysis follows the logic of the Multifactorial Prediction and Deviation Analysis using Regression / Random Forests (MuPDAR[F], e.g. Gries & Deshors 2014; Gries & Adelman 2014 using regression; e.g. Deshors & Gries 2016; Gries & Deshors 2020 using random forests), a two-phase analysis where a certain phenomenon (in our case: original mode of delivery) is modelled in train data and, provided that the model predicts the phenomenon well, the same model is used to predict the phenomenon in different data that diverges from the train data in some respect. In our case, original models are trained on non-interpreted data and they are used to predict interpreted data. Analysing the deviations oc-

curring in the predictions on the second dataset, based on the model trained with the first dataset, gives a detailed insight on the ways in which the two datasets diverge from one another. The method has been used successfully when explaining how L1 and L2 users of English or L2 users with different L1 backgrounds diverge from each other (e.g. Gries & Deshors 2015 on dative alternation differences between EFL and ESL learners; Wulff & Gries 2019 on L1-related variation in verb–particle constructions in L2 English), but also to contrast translated with non-translated English (Kruger & De Sutter 2018 on *that*-omission).

In the present study, we used the obtained forest models of each featureset to predict the original mode of delivery in the test data. We then selected the model of the best performing featureset and used that to predict the mediated data. Looking at the direction of the deviation provides an overall view on the role of formality in mediated language use: if texts delivered originally impromptu are predicted more often falsely as having been read out loud than the other way around, the results can be seen to indicate that the mediated texts are indeed relatively more formal than non-mediated ones. Comparing the results of the machine prediction with the human judgements (cf. §3.2.2.) makes it possible to validate (or reject) the applicability of the different featuresets as indicators of a formality difference.

As a final step, MuPDARF logic allows for further analysis of the observed deviations. To this end, we built a final forest model on the erroneously predicted mediated data. Here, we followed the logic of Deshors & Gries (2016): we had as the response a numeric variable that indicated how far off the prediction was from being correct. On the other hand the predictors were the features included in the final model, as well as the constraining language. The values of the variable range from −0.5 to 0.5, where negative values represent cases where a read out text was predicted as impromptu, and positive values the opposite. The further away from zero the value, the more erroneous the prediction.

In short, we first trained a range of forest-based classifiers to distinguish impromptu speeches from those read out in the non-mediated native English variety; these models involved a categorical response variable (impromptu vs. read). We then evaluated which of the featuresets make it possible to fit the most accurate model. Having selected the most viable model for the native English non-mediated speeches, in line with the MuPDARF approach, we trained another model to predict to what extent mediated/interpreted speeches deviate from the outcomes of the native English non mediated variety. In the latter case the response variable was a numeric one, i.e. the observed deviation.

4 Results

4.1 Human evaluations of formality differences

Table 3 shows for each of the 20 texts how many times (out of 54) each read out text was labelled as more formal, and each impromptu text as less formal than the other text in the same (shuffled) pair.

Table 3: Perceived formality of read out and impromptu texts.

Text ID	Subjects labelling text as more formal	Text ID	Subjects labelling text as less formal
read_05	53 (98.1%)	impromptu_07	51 (94.4%)
read_04	49 (90.7%)	impromptu_05	50 (92.6%)
read_09	49 (90.7%	impromptu_01	49 (90.7%)
read_01	47 (87.0%)	impromptu_08	49 (90.7%)
read_10	47 (87.0%)	impromptu_06	46 (85.2%)
read_08	45 (83.3%)	impromptu_02	45 (83.3%)
read_02	44 (81.5%)	impromptu_03	44 (81.5%)
read_03	44 (81.5%)	impromptu_09	44 (81.5%)
read_07	41 (75.9%)	impromptu_10	41 (75.9%)
read_06	38 (70.4%)	impromptu_04	35 (64.8%)

In all cases, more than half of the respondents perceived read out texts as more formal and impromptu texts as less formal, with percentages of agreement among raters above 80% for 16 texts out of 20, equally split across the two categories.

53 out of 55 respondents further provided answers to the final question of the survey asking them to illustrate the reasons for their choices regarding the more formal text. These answers were categorized bottom-up to gain an understanding of the linguistic or stylistic features that respondents associated with formality or informality. Table 4 reports on these categories, with percentages indicating the proportion of answers mentioning them, as well as examples of the specific features mentioned by respondents.

In §4.2.1 we compare and discuss some of these features, especially those concerning lexis and syntax, with respect to those emerging from the text-based analysis of formality differences.

Table 4: Features associated with text formality.

Category	**Features (e.g.)**
Discourse (35.7%)	Coherence/cohesion Impersonal style Lack of repetitions
Lexis (32.2%)	Rare vocabulary Terminology Formulae (e.g. greetings)
Syntax (28.0%)	Complex sentence structure Sentence length Passive forms
Contents (4.1%)	Hard facts (vs. opinions)

4.2 Corpus-based identification of formality differences

4.2.1 Best-predicting features of formality differences

The 20 trained forest-based classifiers were used to predict formality differences within the test data, corresponding to the 20 texts which were also used in the human evaluation experiment. The performance of each classifier was evaluated in terms of precision, i.e. the probability that the model classifies texts correctly as read out or impromptu, and recall, i.e. the ability of the model to find all instances of read out and impromptu texts. Table 5 reports precision and recall values alongside the resulting F-measure (the harmonic mean of the other two values), which is conventionally used to assess the accuracy of classification models. For space reasons, results are only reported for the 10 best-scoring models.

With 6 featuresets among the overall top 10, the category of sequential n-grams seems to perform better than other categories, and especially n-grams of length 1, i.e. unigrams (of syntactic functions, POS and words). Syntax-based features also perform well, both when used as part of sequential n-grams and within positional-based featuresets. Among the top 3 models, dependency POS bigrams get the highest F-measure together with syntax 1-grams, while at the same time packing more linguistic information than the other two best-scoring featuresets (i.e. unigrams of syntactic functions and POS respectively), since a) they take into account syntactic functions in the form of dependencies, and b)

Table 5: Classification accuracy of the 10 best-scoring models.

Featureset	Precision	Recall	F-measure
syntax 1-gram	0.8	0.8	0.8
dependency POS 2-gram	0.8	0.8	0.8
POS 1-gram	0.7	0.875	0.778
positional POS 1-gram	0.76	0.716	0.737
word 1-gram	0.8	0.667	0.727
sequential syntax 3-gram	0.8	0.571	0.667
sequential syntax 2-gram	0.7	0.636	0.667
sequential POS 2-gram	0.6	0.667	0.632
positional syntax	0.58	0.658	0.616
positional lemma	0.6	0.545	0.571

are based on POS. In view of the small differences in terms of F-measure as well as the higher linguistic and functional interpretability of the featureset, which was also observed in previous work (cf. §3.2.4.), we selected the model based on dependency-defined POS bigrams as the most suitable classifier for subsequent analyses.

Nine features (Table 6) contribute to the distinction between read out and impromptu texts in our analysis. The five features more frequently observed in texts originally read out, which we hypothesized to be characterized by greater formality, all involve nouns. In particular, three of these involve a determiner used in conjunction with a noun, where the noun is typically related to the topic of the discussion, and the determiner is used to increase precision (e.g. *all exports*, *the threat*, *some enlargements*). Proper nouns serving as a direct object of a verb are often used in the formula of expressing thanks (*thank Iñigo*), and nouns coordinated by a conjunction are frequently observed in the formula opening a speech (*ladies and gentlemen*). The latter feature is also observed in excerpts studded with terminology (e.g. *scrutiny and control* or *deficit and debt*). Altogether, these observations tie in both with previous literature and with comments made by respondents in our survey. Heylighen & Dewaele (1999) noted that nouns and noun phrases are typically more frequent in formal texts, and both formulaic expressions and use of specialized terminology were noted by respondents as also associated with increased formality.

The remaining features are more frequent in impromptu texts, which we relate to informality. Three of them involve verbs, and specifically verbs having

Table 6: Features distinguishing impromptu and read out texts illustrated by examples.

POS dependency bigram	**Example**	**Register where more frequent**
DETNODE_det.predet_NOUNHEAD (determiner predetermining noun)	I believe Thailand should be completely delisted from all poultry meat exports until they can prove they have the infrastructure	Read
DETNODE_det_NOUNHEAD (determiner determining noun)	The threat to Europe 's health from the rapid spread of disease is real and present.	Read
VERBHEAD_obj_PROPNNODE (verb having a proper noun as a direct object)	Let me thank Íñigo and the coordinators for taking the decision	Read
NOUNHEAD_conj_NOUNNODE (nouns connected by coordinating conjunction)	I genuinely wonder, ladies and gentlemen, where the human rights regaining of confidence of the markets is the basis for a stable, sustainable growth and jobs	Read
DETHEAD_nmod_NOUNNODE (determiner modifying noun)	many of them raise concerns about some of the previous enlargements	Read
NOUNHEAD_advmod_ADVNODE (adverb acting as modifier of noun)	that's the agricultural sector here in Europe that room over there	Impr.
PRONNODE_nsubj_VERBHEAD (pronoun acting as subject to verb)	It is time we took action and showed that we support the Iranian opposition	Impr.
ADJHEAD_conj_VERBNODE (adjective coordinated with verb)	Now this is not sustainable, and it's not fair	Impr.
NOUNHEAD_advcl_VERBNODE (noun acting as adverbial clause modifier of verb)	It's a very good initiative but the Parliament, as Mr de Jong says, has some concerns.	Impr.

pronouns as subjects (*we took*) and verbs associated to adjectives or nouns (*sustainable [...] it's*; *initiative [...] says*); the fourth one involves adverbs modifying nouns (*sector here*). These features are consistent with Heylighen & Dewaele's observation that verbal structures tend to be preferred in informal contexts to nominal structures (1999), and that informal style is often associated with deictic expressions, which are exemplified here by pronouns (*we took*) and adverbs such as *there* or *here*. Inflected verbs, too, are suggested by Heylighen & Dewaele (1999) to be "intrinsically deictic because they refer implicitly to a particular time through their tense (...), and to a particular subject through their inflection". Highly context-dependent, deictic expressions are bound to decrease precision and increase involvement, which renders the text more informal.

The next Section reports on the results of the MuPDAR[F] analysis based on the model featuring the dependency bigrams discussed here.

4.2.2 MuPDAR[F] analysis

The model based on POS dependency bigrams, which achieved the best compromise between classification accuracy and linguistic interpretability in classifying the non-mediated dataset, was applied to the mediated dataset. Following the MuPDAR[F] approach (Gries & Deshors 2020), the aim of this second set of analyses is that of assessing whether and how the two datasets of native and interpreted texts differ from one another in terms of formality-related features.

Table 7: Prediction accuracy of interpreted data.

	Source language			
Prediction	**fr**	**it**	**pl**	**Total**
Impromptu: correct	4 (20%)	2 (20%)	6 (30%)	12 (24%)
Read out: correct	7 (35%)	3 (30%)	7 (35%)	17 (34%)
Read out: erroneous	6 (30%)	3 (30%)	4 (20%)	13 (26%)
Impromptu: erroneous	3 (15%)	2 (20%)	3 (15%)	8 (16%)
Total	20 (100%)	10 (100%)	20 (100%)	50 (100%)

Table 7 reports data on the model's prediction accuracy in the interpreted dataset, expressed as the number of interpreted texts which were identified correctly and erroneously as deriving from an impromptu or read out source text (henceforth called "impromptu" and "read out" texts for brevity). Since the baseline is constituted by predictions on non-mediated texts, the higher the accuracy,

the more mediated texts can be hypothesized to be similar to their non-mediated counterpart. Conversely, the higher the degree of deviation, the more mediated texts can be seen as different from non-mediated ones in terms of formality.

Impromptu interpreted texts get the lowest percentage of correct predictions (24%), and the highest percentage of incorrect ones (26%), pointing to the fact that they are most often predicted as being read out. Only rarely does the opposite scenario occur, i.e. that the model predicts a read out interpreted text as being impromptu (16% of cases). The picture that emerges is thus one where mediated texts are generally predicted as being read out, irrespective of the actual mode of delivery of their source text.

If, as we hypothesized, and as the human evaluation seems to confirm, the impromptu vs. read out distinction reflects a distinction in terms of formality, the analysis suggests that a) mediated texts tend to differ from their non-mediated counterparts in terms of levels of formality, and b) deviations occur in the direction of mediated texts being more formal than non-mediated ones.

In order to assess which features contributed to the erroneous predictions, we built another model that focused solely on the erroneous predictions. The response variable was the gravity of the prediction error – how far the prediction was from being correct – and the predicting variables were the ones that were earlier identified as contributing to distinguish the impromptu from the read out texts. The contribution of the different variables was measured in terms of the permutations-based variable importance values reported as part of the model. Figure 2 includes only those features with positive values in permutation, i.e. those which contributed positively to modelling. In other words, Figure 2 displays which features tend to be used by interpreters differently with respect to native speakers, leading to different levels of formality in interpreted texts.[5]

The most important dependency bigram which distinguishes correctly and erroneously labelled texts is constituted by pronominal subjects in a preverbal position. As can be seen in the four leftmost elements of Figure 3, these bigrams are clearly more frequent in the correctly labelled impromptu texts than in the other predicted data. When compared with the reference data, i.e. impromptu and read out non-interpreted speeches, the distribution follows the reference data, suggesting that the interpreted impromptu texts predicted erroneously as read out are indeed more formal in this respect than other impromptu texts. Interestingly, the variable behavior is not bidirectional, as the read out texts that

[5]Another way of interpreting the results of this analysis would be that features found in the final model but not here do not behave differently in falsely labelled and correctly labelled texts.

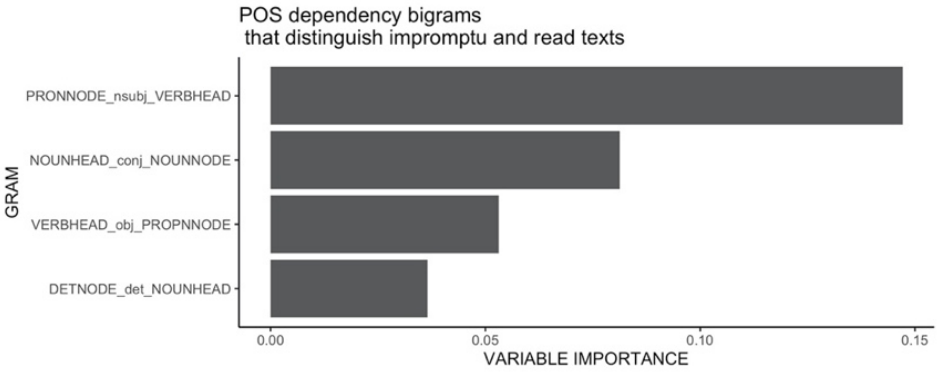


Figure 2: Importance of variables that contribute to erroneous predictions.

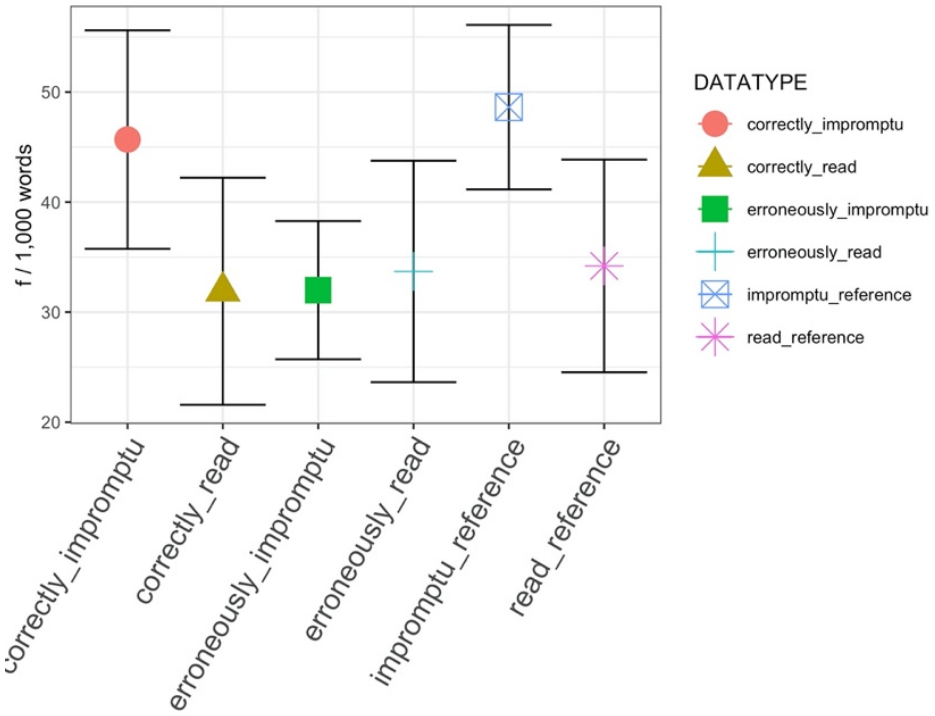


Figure 3: Normalized frequencies of PRONNODE_nsubj_VERBHEAD dependency bigram.

have been predicted erroneously as impromptu, pair with the correctly predicted read out texts.

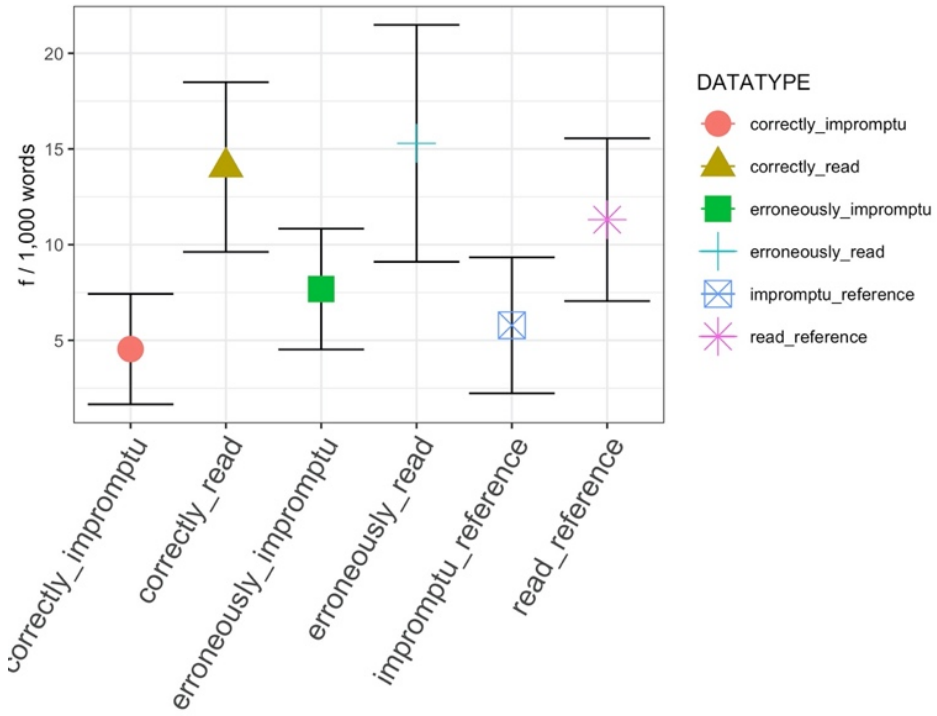


Figure 4: Normalized frequencies of NOUNHEAD_conj_NOUNNODE dependency bigram.

The dependency bigram scoring second in terms of variable importance reflects the use of coordinated noun phrases. As indicated in Figure 4, the structure is more frequent in the correctly labelled read out texts than in the correctly labelled impromptu texts, and this tendency reflects the pattern in the reference data. Impromptu texts that are labelled erroneously as being read out behave similarly to the correctly labelled read out texts, while texts labelled erroneously as impromptu are closer in this regard to the texts labelled correctly as impromptu. The effect of this variable is bidirectional, as it distinguishes both texts with increased formality (predicted erroneously as read out) and those with decreased formality (predicted erroneously as impromptu).

Dependency bigrams featuring proper nouns as postverbal objects is the third most important distinguishing feature for the erroneously labelled texts. They

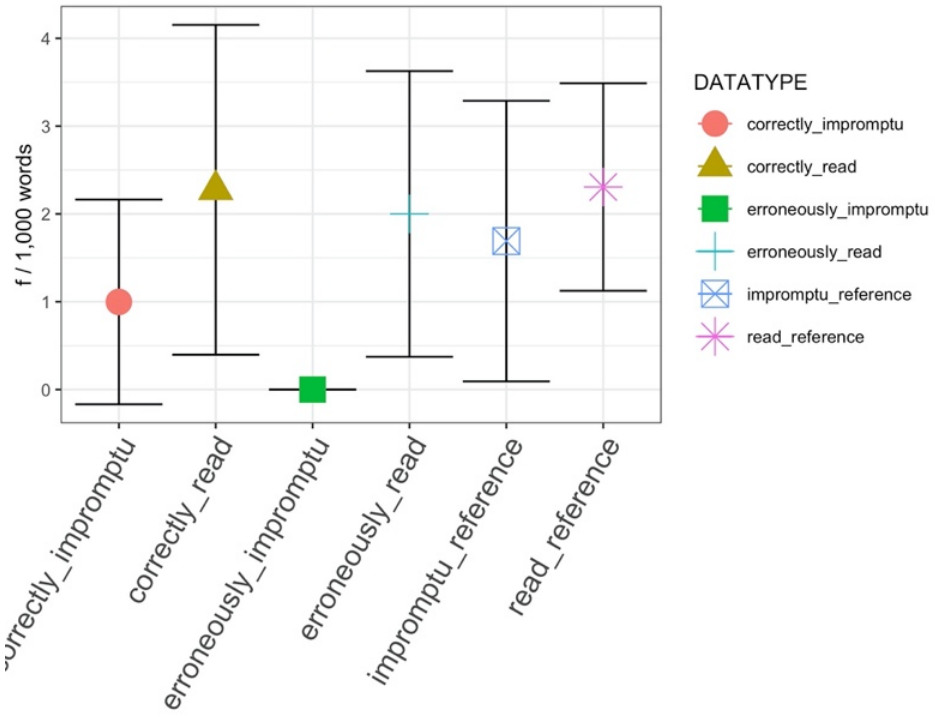


Figure 5: Normalized frequencies of VERBHEAD_obj_PROPNNODE dependency bigram.

occur more frequently in the correctly labelled read out texts than in the correctly labelled impromptu texts, and the impromptu texts labelled erroneously as read out are grouped together with the actual read out texts (Figure 5). This grouping also reflects the reference data, where these bigrams are more frequent in the read out texts than in the impromptu ones. It should be noted, however, that the feature is relatively rare (only 1.5 / 1,000 words on average), and it does not occur a single time in the texts labelled erroneously as impromptu.

The fourth most important dependency bigram reflects the use of prenominal determiners. Such determiners are relatively more common in the correctly labelled read out texts than in the correctly labelled impromptu texts (see Figure 6). This tendency reflects that of the reference data, even though determiners seem to be overall more frequent in the interpreted data than the reference data. In this case, the erroneously labelled texts actually behave relatively similarly to the correctly labelled ones, but the in impromptu texts labelled erroneously as

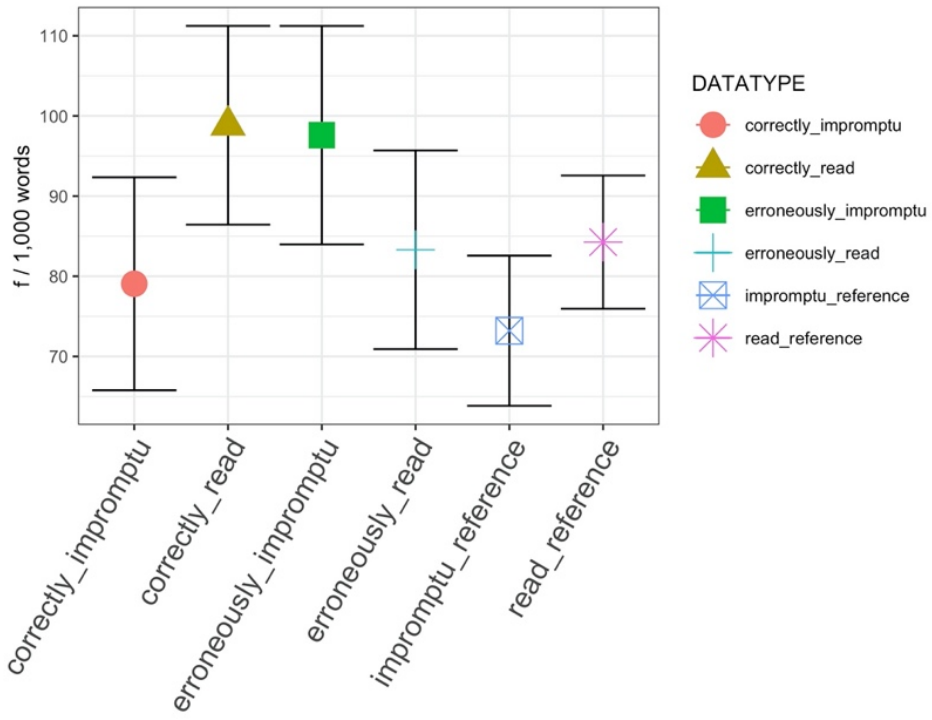


Figure 6: Normalized frequencies of DETNODE_det_NOUNHEAD dependency bigram.

read out fall in between the two correctly labelled datasets. This might be taken to suggest that they are in this respect more formal than the other impromptu texts.

5 Discussion and conclusion

Formality is oftentimes referred to in comparisons across registers, genres or varieties as an explanation of the differences in linguistic features observed between them. However, so far it has rarely been the focal point of corpus-based linguistic investigations in general, and in Translation Studies in particular. As pointed out in §2, the link between formality and the features that attest to it is usually established indirectly, as formality differences are not subject to independent evaluation. In this paper we have made an attempt to fill this gap by triangulating human judgements and specific linguistic features derived bottom-up from

a corpus. On the basis of the human-validated dataset of formality features, we used a corpus-based approach to examine formality differences of interpreted and non-interpreted texts.

We started by examining read out and impromptu speeches delivered at the European Parliament by native English speakers, and obtained a list of nine linguistic features contributing to text formality or informality. In an experiment involving human judgments, we observed that the distinction between read out and impromptu speeches is associated with a difference between more formal vs. more informal texts. On the basis of this evidence, we used the nine features in a model that classified interpreted texts as read out vs. impromptu (more vs. less formal). This analysis showed that interpreted texts were generally predicted as being read out, irrespective of the actual mode of delivery of their source text, pointing to a higher level of formality. Overuse of some of the features, however, pointed in the opposite direction, i.e. to informality in interpreted texts, even in cases where source texts were read out. In search of potential explanations for such results, we looked into the linguistic features that contributed to the most erroneous predictions of interpreted texts, or in other words the over- or underrepresented features in interpreted texts that increased or decreased the level of formality.

One of these features involves coordinated noun phrases, which occur more frequently in the read out texts examined here, and are overrepresented in some interpretations of impromptu texts and underrepresented in some interpretations of read out texts. Nouns, in general, are typically more frequent in formal texts (Heylighen & Dewaele 1999), as formal settings usually require clarity and precision, and nouns (binomials in particular) are more likely to increase precision than, e.g., context-dependent pronouns. Yet, nouns are cognitively more demanding than pronouns, as lexical access to content words is in general slower than access to function words (Segalowitz & Lane 2000), and interpreters need to carefully manage their cognitive load, which might be decisive in this context.

Looking at it from another angle, the outcomes reported here also tap into issues long investigated by translation and interpreting scholars. The greater use of nouns instead of pronouns might hint at interpreters' explicitating meaning (Blum-Kulka 1986). Hence, the use of coordinated noun phrases, here identified as a feature of formality, might in the case of interpreters be associated to factors like cognitive load and the need to disambiguate meaning.

Postverbal proper nouns acting as objects constitute another feature which was more frequently found in read out (formal) texts, and contributed to erroneous formality classifications of interpreted texts. A large proportion of the actual expressions hidden behind these dependency tags refer to the act of thanking

a specific person. In the context of the European Parliament, these formulae are used when a speaker is thanking another MEP or thanking the President for giving them the floor. Expressing thanks is a recurring act in this context and, as also pointed out by the respondents of our survey, formulae increase the level of formality of the text.

Frequencies of use of pronominal subjects also led to erroneous classifications of interpreted text. Both pronouns and verbs are typically more frequently used in informal texts, with pronouns being deictic words referring to immediate context (Heylighen & Dewaele 1999) and associated with personal involvement. It is worth noting that interpreted texts, even though produced simultaneously in the same setting as their source texts, are a product of mediation and transferring the message of the original speaker. It is plausible that both personal involvement and immediacy of context diminishes in language mediation, potentially leading to lower frequencies of pronominal subjects in a preverbal position in texts that otherwise bear more traits of informality.

Before concluding, a few limitations of the research design and method should be highlighted. First, the small size of the sample cannot be overlooked. This was mainly justified by the labour intensiveness of transcribing speeches, as well as the need to have part of them annotated by human subjects. Replication studies are therefore in order to test the results obtained here, based on larger and/or more varied datasets (e.g. in terms of text types), and ideally involving a higher number of respondents, e.g. by adopting crowdsourcing methods. While the statistical methods were selected with these limitations in mind, it is likely that richer featuresets (e.g. word trigrams) would have fared relatively better with larger datasets. On the other hand, the advantage of simpler, and arguably more abstract, featuresets such as POS dependency bigrams, is that they make study designs like this feasible. Second, it should be noticed that the use of POS dependency bigrams, though reaching a satisfactory level of classification accuracy, limits the scope of the investigation to syntactic (and partly lexical) phenomena only, thus excluding features pertaining in the level of discourse, which were mentioned by respondents as being equally important as lexis in determining formality. The third and last note of caution concerns the use of a dependency parser to extract model features. Parsers are usually trained on written data, while in this case we applied them to spoken data: further studies could investigate the impact of parser accuracy in study setups like the one adopted here.

In terms of applications, we think of interpreter training as the field on which our results have a more direct bearing. As shifts of formality might have an impact on the perception of the speech, it is vital that both interpreters and interpreter trainers are sensitized to this issue and the list of features associated with

formality differences in English could hopefully help in the development of adequate training aids. Hopefully, the approach demonstrated in this paper might also be instrumental in the development of interpreter training aids targeted at sensitizing future interpreters to formality shifts in genres other than parliamentary debates.

Appendix A

A.1 Editing guidelines for spoken texts, and text selection criteria

- Eliminate DYSFLUENCIES (e.g. "and **t-** tremendous concern"), EMPTY and FILLED PAUSES (e.g. "we have to **ehm** protect").
- Eliminate REPETITIONS, but only when these are in the context of other dysfluencies (e.g. "procedure in this House, ehm because **we have ma-** we have been able to make significant improvements").
- Keep "Thank you President" at the beginning.
- Add punctuation, especially commas, especially in the EPIC texts, where punctuation is not present (e.g. "And we have to ask why do they do it, and" => "And we have to ask: why do they do it? And")
- Select texts with around 150 words or more. Where needed, shorten texts to make them no longer than 200 words (for golden standard/train data), and 250 words (for test set).

A.2 Example of an edited text

Transcript text	Text after edits
thank you very thank you very much ehm President. can say that where I come from in Northern Ireland we have a very vibrant poultry industry. and t- tremendous concern has been expressed to me by that industry. and what has happened while is unfortunate what has happened in Asia I think we have to ehm protect ehm the European market because it's an extremely ehm large market ehm for the poultry industry. I am concerned about the length of time it took the authorities in Asia in letting us know wha- that the the the outbreak had taken place.	Thank you very much President. I can say that where I come from in Northern Ireland we have a very vibrant poultry industry. And tremendous concern has been expressed to me by that industry. What has happened in Asia, I think we have to protect the European market because it's an extremely large market for the poultry industry. I am concerned about the length of time it took the authorities in Asia in letting us know that the outbreak had taken place.

Appendix B Text of questionnaire

Dear Participant!
Language we encounter in our everyday lives varies in many ways - spoken language diverges from written language, different dialects differ from each other, language in the school text books is different from legal documents and so on.

One of the ways different uses of language diverge from each other is **formality** - certain texts seem more formal than others. In this study, we are interested in formality and would like to ask for your help in understanding it better. In what follows, you will be shown ten pairs of texts. We would like you to **quickly read the texts and simply indicate which of the texts in each pair you find more formal**. All the texts come from the European Parliament discussions. We encourage you to follow your first instinct in the decision making.

The whole questionnaire should take about 10 minutes. After the **ten questions**, we will ask whether there was something specific that governed your decision-making. You will also be asked a couple of very basic questions on your language background. Overall, the whole questionnaire is anonymous and nei-

ther we nor anyone else have access to any personal information of the participants.

References

Abu Sheikha, Fadi & Diana Inkpen. 2010. Automatic classification of documents by formality. In *Proceedings of the 6th International Conference of Natural Language Processing and Knowledge Engineering*, 1–5. Beijing: NLPKE-2010. DOI: /10.1109/NLPKE.2010.5587767.

Andrén, Mats, Johan M. Sanne & Per Linell. 2010. Striking the balance between formality and informality in safety-critical communication: Train traffic control calls. *Journal of Pragmatics* 42(1). 220–241. DOI: 10.1016/j.pragma.2009.05.022.

Atkinson, J. Maxwell. 1982. Understanding formality: The categorization and production of "formal" interaction. *British Journal of Sociology* 33(1). 86–117. DOI: /10.2307/589338.

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.

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/llc/fqi039.

Bartłomiejczyk, Magdalena. 2016. *Face threats in interpreting: A pragmatic study of plenary debates in the European Parliament.* Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Doctoral dissertation). https://wydawnictwo.us.edu.pl/sites/wydawnictwo.us.edu.pl/files/face_threats_in_interpreting_czw_st_e.pdf.

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

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37. DOI: 10.1515/cllt-2012-0002.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House & Shoshana Blum-Kulka (eds.), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, vol. 17, 17–36. Tübingen: Gunter Narr Verlag.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.

Conrad, Susan & Douglas Biber (eds.). 2001. *Variation in English: Multi-dimensional studiess*. Essex: Pearson.

De Sutter, Gert, Isabelle Delaere & Koen Plevoets. 2012. Lexical lectometry in corpus-based translation studies. In Michael P. Oakes & Meng Ji (eds.), vol. 51 (Studies in Corpus Linguistics 51), 325–345. DOI: 10.1075/scl.51.13sut.

Deshors, Sandra C. & Stefan Th. Gries. 2016. Profiling verb complementation constructions across New Englishes. *International Journal of Corpus Linguistics* 21(2). 192–218. DOI: 10.1075/ijcl.21.2.03des.

Fang, Alex Chengyu & Jing Cao. 2009. Adjective density as a text formality characteristic for automatic text classification: A study based on the British National Corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, 130–139. Hong Kong: City University of Hong Kong. https://www.aclweb.org/anthology/Y09-1015.

Ferraresi, Adriano & Silvia Bernardini. 2019. Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In Irene Doval & M. Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*, vol. 90 (Studies in Corpus Linguistics), 123–139. Amsterdam/Philadelphia: John Benjamins.

Graesser, Arthur C., Danielle S. McNamara, Zhiqang Cai, Mark Conley, Haiying Li & James Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal* 115(2). 210–229. DOI: 10.1086/678293.

Gries, Stefan Th & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1). 109–136.

Gries, Stefan Th. & Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms*, 35–54. New York: Springer.

Gries, Stefan Th. & Sandra Deshors. 2015. EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1). 130–159.

Gries, Stefan Th. & Sandra C. Deshors. 2020. There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal* 44(1). 69–96. DOI: /10.2478/icame-2020-0003.

Hale, Sandra. 1997. The treatment of register variation in court interpreting. *The Translator* 3(1). 39–54.

Heylighen, Francis & Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 7(3). 293–340. DOI: /10.1023/A:1019661126744.

Heylighen, Francis & Jean-marc Dewaele. 1999. *Formality of Language: Definition, measurement and behavioral determinants*. Internal report. Brussels, Belgium: Center "Leo Apostel", Free University of Brussels,

Ivaska, Ilmari & Silvia Bernardini. 2020. Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics* 43(1). 33–57. DOI: 10.1017/S0332586520000013.

Ivaska, Ilmari, Adriano Ferraresi & Silvia Bernardini. In print. Syntactic properties of constrained English: A corpus-driven approach. In Sylviane Granger & Marie-Aude Lefer (eds.), *Extending the scope of corpus-based translation studies* (Bloomsbury Advances in Translation). London: Bloomsbury.

Jarvis, Scott. 2013a. Capturing the diversity in lexical diversity. *Language Learning* 63. 87–106.

Jarvis, Scott. 2013b. Defining and measuring lexical diversity. In Scott Jarvis & Michael Daller (eds.), *Vocabulary knowledge: Human ratings and automated measures* (Studies in Bilingualism), 13–44. Amsterdam: John Benjamins.

Jarvis, Scott. 2017. Grounding lexical diversity in human judgments. *Language Testing* 34(4). 537–553.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu. (Doctoral dissertation). https://repozytorium.amu.edu.pl/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf.

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino & Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, 133–142. Brussels, Belgium: Association for Computational Linguistics.

Koppel, Moshe & Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1318–1326. Portland, Oregon: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1132.

Kruger, Haidee & Gert De Sutter. 2018. Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior* 1(2). 251–290. DOI: https://doi.org/10.1075/tcb.00011.kru.

Kruger, Haidee & Bertus van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide* 37(1). 26–57. DOI: 10.1075/eww.37.1.02kru.

Kruger, Haidee & Bertus van Rooy. 2018. Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide* 39(2). 214–242. DOI: 10.1075/eww.00011.kru.

Kursa, Miron B. & Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36(11). 1–13. DOI: 10.18637/jss.v036.i11.

Li, Haiying, Arthur C. Graesser, Mark Conley, Zhiqiang Cai, Phillip I. Pavlik jr & James W. Pennebaker. 2016. A new measure of text formality: An analysis of discourse of Mao Zedong. *Discourse Processes* 53(3). 205–232. DOI: 10.1080/0163853X.2015.1010191.

Mauranen, Anna. 2008. Universal tendencies in translation. In Gunilla M. Anderman & Margaret Rogers (eds.), *Incorporating corpora: The linguist and the translator*, 32–48. Bristol, Blue Ridge Summit: Multilingual Matters.

Moe, Marija Zlatnar. 2010. Register shifts in translations of popular fiction from English into Slovene. *Why translation studies matters* 88. 125.

Olohan, Maeve. 2003. How frequent are the contractions?: A study of contracted forms in the Translational English Corpus. *Target. International Journal of Translation Studies* 15(1). 59–89. DOI: 10.1075/target.15.1.04olo.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London; New York: Routledge.

Popescu, Marius. 2011. Studying translationese at the character level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 634–639. Hissar, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/R11-1091.

Pym, Anthony. 2005. Explaining explicitation. In Krisztina Károly & Ágota Fóris (eds.), *New trends in translation studies. in honour of Kinga Klaudy*, 29–34. Budapest: Akadémiai Kiadó.

Rabinovich, Ella, Sergiu Nisioi, Noam Ordan & Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1870–1881. Berlin: Association for Computational Linguistics.

Sandrelli, Annalisa, Claudio Bendazzoli & Mariachiara Russo. 2010. European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary results on lexical patterns in simultaneous interpreting. *International Journal of Translation* 22. 165–203.

Segalowitz, Sidney J. & Korri C. Lane. 2000. Lexical access of function versus content words. *Brain and language* 75(3). 376–389.

Shlesinger, Miriam. 1989. *Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum*. Tel Aviv University, Faculty of the Humanities, Department of Poetics & Comparative Literature. (MA thesis). DOI: 10.13140/RG.2.2.31471.69285.

Shlesinger, Miriam & Noam Ordan. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24(1). 43–60. DOI: 10.1075/target.24.1.04shl.

Toury, Gideon. 1995. *Descriptive translation studies and beyond* (Benjamins translation library 4). Amsterdam: Benjamins. DOI: 10.1075/btl.100.

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.

Wright, Marvin N. & Andreas Ziegler. 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1). 1–17. DOI: 10.18637/jss.v077.i01.

Wulff, Sfefanie & Stefan Th. Gries. 2019. Particle placement in learner English: Measuring effects of context, first language, and individual variation. *Language Learning* 69(4). 873–910. DOI: /10.1111/lang.12354.

Chapter 3

Fluency parameters in the Polish Interpreting Corpus (PINC)

Agnieszka Chmiel[a], Danijel Korzinek[b], Marta Kajzer-Wietrzny[a], Przemysław Janikowski[c], Dariusz Jakubowski[c] & Dominika Polakowska[a]

[a]Adam Mickiewicz University [b]Polish-Japanese Institute of Information Technology [c]University of Silesia

The following chapter introduces PINC — the Polish Interpreting Corpus, a Polish-English and English-Polish corpus of short European Parliament speeches and their interpretations. The uniqueness of PINC, apart from its language combination, consists in careful balancing of mode of delivery, in rich metadata, interpreter identification and availability of a strictly controlled subcorpus of retour interpretations.

The chapter also briefly presents custom-built tools used in the making of the corpus, especially for transcription, text-audio alignment at word level and interpreter identification.

To showcase PINC's potential for analysing various aspects of simultaneous interpreting, we examined fluency parameters, such as speaking rate and pauses, in the Polish-English subcorpus. We found that interpreting speed was modulated by the source text speaking and articulation rate and the target text compression rate. Target texts had fewer but longer silent pauses and more numerous and longer filled pauses. Together with shorter runs, understood as utterances uninterrupted by pauses, this suggests more fragmented delivery of interpretations. We also found interesting individual differences in compression rate with the majority of interpreters producing interpretations longer than the source texts.

1 Introduction

New empirical paradigms require constant development of tools that would allow us to investigate increasingly challenging research questions. This is partic-

Agnieszka Chmiel, Danijel Korzinek, Marta Kajzer-Wietrzny, Przemysław Janikowski, Dariusz Jakubowski & Dominika Polakowska. 2022. Fluency parameters in the Polish Interpreting Corpus (PINC). in Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 63–88. Berlin: Language Science Press. DOI: ??

ularly visible in the case of Corpus Interpreting Studies, where new incarnations of interpreting or intermodal corpora based on the European Parliament plenary debates have emerged every few years ever since 2005, when EPIC: European Parliament Interpreting Corpus (Monti et al. 2005) was announced. Despite the readiness of the corpus creators to collaborate and share their data (most corpora are available either on-line or from their owners upon request), all of them stand by their own preferred corpus tools and compilation procedures as these fit their research needs best. This is related to the fact that interpreting is not limited to the text and the linguistic aspects captured in transcripts do not reflect the full communication event. As the Corpus Interpreting Studies pioneer, Shlesinger (1998: 1) put it "[w]hile transcription, however laborious, can provide us with a representation of the interpreter's linguistic output, its failure to reflect the concomitant paralinguistic dimensions is a major drawback". Hence, so far, most interpreting corpora have been compiled with particular research objectives in mind. Such is also the case of PINC: The Polish Interpreting Corpus, which, at a later stage of the project, will be used to analyse activation and inhibition and thus needs intense annotation of such features as e.g. temporal details of individual words, pause length or word-level alignment. Many of these features will enable a peek into the process of interpreting, rather than being strictly product-oriented and the data obtained will inform the selection of stimuli for final-stage experimental procedures. This puts quite a heavy demand on strict balancing and control as well as the sheer size of the corpus.

This chapter presents this newly created Polish Interpreting Corpus and offers an example study that shows the potential of PINC in analysing various aspects of simultaneous interpreting. We have decided to concentrate on interpreter fluency, including speaking rate and pauses, and to look for characteristics in the source text that modulate fluency parameters in interpretations.

To the best of our knowledge, only three interpreting corpora have been created for the Polish-English language combination so far. Two of them (Dumara 2015, Bartłomiejczyk 2016) were analysed manually and with a narrow research focus, such as intrusive pronouns or face threats. The third is a Polish English small-scale subcorpus currently available as part of EPTIC (Department of Interpreting and Translation - Forlì Campus). We hope that PINC, thanks to its size and advanced analytical metadata (to be described below), will make it possible to tackle varied and numerous research questions.

2 PINC: a new member of the EPIC suite of corpora

2.1 Features

The Polish Interpreting Corpus (PINC) adds to an ever-growing family of interpreting or intermodal corpora derived from the European Parliament debates called by Bernardini et al. (2018) "the EPIC suite of corpora". As summarized by Bernardini et al. (2018: 22) "[t]he availability of interpretations and translations from and into a large number of languages, the ease of access to the videos (downloadable from the Internet), and the high professional standards of the interpreters" makes this source very promising for interpreting corpora, which is why EPIC suite is constantly growing. Next to EPIC: European Parliament Interpreting Corpus (Monti et al. 2005), TIC: Translation and Interpreting Corpus (Kajzer-Wietrzny 2012: 57), EPICG: European Parliament Interpreting Corpus – Ghent (Defrancq et al. 2015) and EPTIC: European Parliament Translation and Interpreting Corpus (Ferraresi & Bernardini 2019), PINC comprises a collection of recordings and transcripts of speeches delivered during the plenary sessions of the European Parliament, as well as their simultaneous interpretations. These were obtained from the Europarl website (Directorate-General for Communication).

Several aspects of the PINC compilation process have been modelled on the work of the creators of the other corpora of the EPIC suite (e.g. EPIC or EPTIC). Thus, the texts compiled in the PINC corpus follow the same topic classification as EPIC and EPTIC for ease of comparison; similar contextual metadata have been collected and transcription guidelines were, to a large extent, very much alike. Similarly to TIC, interpreters' voices have been distinguished from one another and individual codes assigned to each voice. The uniqueness of PINC consists in the specific language combination, careful balancing of mode of delivery, detailed interpreter voice identification, speech-to-text and sentence alignment of the whole corpus and in the specific tools employed to automatise parts of the compilation process. Some of these tools and features are described in more detail below.

2.1.1 Corpus size, speech length and mode of delivery

Since we are interested in a fully bi-directional analysis, PINC consists of four balanced subcorpora: Polish source texts (ST-PL), their interpretations into English (TT-EN), English source texts (ST-EN) and their interpretations into Polish (TT-PL). All of these were collected from the Europarl website from plenary sessions recordings of the European Parliament sittings taking place between Jan-

uary 2009 and September 2010. The reasons for selecting such a time frame were twofold. First, we wanted to use verbatim reports to facilitate the transcription process and later recordings are not accompanied with them. Second, we are planning to correlate corpus data for individual interpreters with other data, such as working memory spans, obtained from the same interpreters in the same time frame and analysed in other studies (Chmiel 2012; Chmiel 2016; Chmiel 2018).

The PINC corpus comprises texts ranging between 100 and 500 words, with the mean text length of 204 and the median text length of 183 words. Including a higher number of shorter speeches rather than fewer longer ones made it possible to achieve greater variation within the data, as a greater proportion of longer speeches in a corpus of the same size could have easily skewed the data. Thus, texts longer than 500 words have been excluded from the corpus altogether.

As in other corpora of the EPIC suite, speeches in PINC are annotated for mode of delivery. Following EPIC (Monti et al. 2005), most EP-based interpreting corpora use a three-way classification of mode of delivery: impromptu (for unscripted speeches), read (for scripted speeches) and mixed (semi-scripted speeches). In the course of compilation of PINC, a decision was made to include only the first two types of speeches in the corpus; hence speeches of varying degree of scriptedness are not part of the PINC corpus. The reason for excluding mixed speeches was that we found it difficult to indicate precise and objective criteria for assigning speeches to that category. Table 1 presents basic data about the number of speeches and tokens, as well as speech rate in each subcorpus of PINC. More information about ST and TT speaking rates will be provided in §3.1

2.1.2 Topics

As in the remaining corpora of the EPIC suite, specific topics of debates taking place at the European Parliament have been grouped into more general categories including Agriculture and Fisheries, Economics and Finance, Employment, Environment, Health, Justice, Politics, Procedure and Formalities, Science and Technology, Society and Culture. There are topics that dominate the EP agenda, such as Politics or Economics and Finance and those that are only occasionally discussed, e.g. Science and Technology so an even distribution of topics across such a corpus is always difficult to obtain. Moreover, MEPs from different countries are not equally active in all debates. A perfect balance of topics is impossible, but it is still vital to be able to control the impact of the topic in those empirical investigations that require it. In the data selection process we paid particular attention to achieving a relatively even distribution of read and impromptu speeches

Table 1: Basic data about four subcorpora of PINC

		Number of speeches	Number of tokens	Average speech rate (wpm)
Polish source texts	Impromptu	117	20769	127
	Read	115	19399	126
English interpretations	Impromptu	117	21627	127
	Read	115	19656	131
English source texts	Impromptu	115	25715	178
	Read	115	28374	165
Polish interpretations	Impromptu	115	17496	121
	Read	115	19153	110

across topics, although in the end it was not always possible. Also the distribution of the two modes of delivery across topics in the two source language subcorpora, i.e. Polish and English, differs (Figure 1).

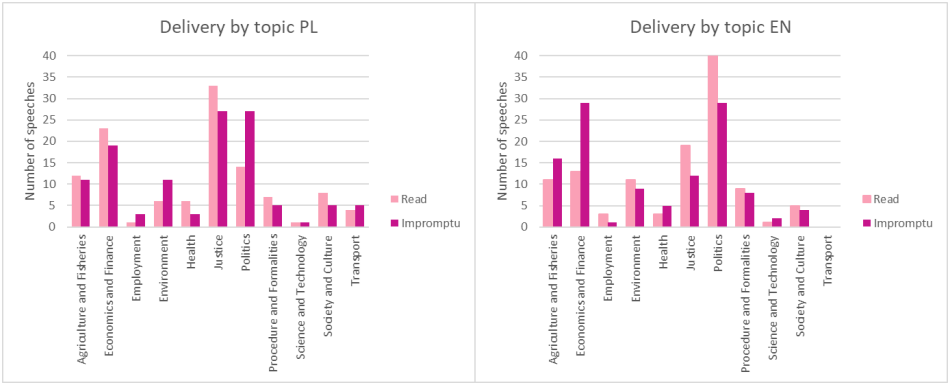


Figure 1: Topic coverage and mode of delivery in PINC subcorpora

It transpires from Figure 1 that, in the PINC dataset, speeches regarding Economics and Finance are more often delivered impromptu by native English speakers at the EP, while the Polish MEPs read them slightly more often. Even more striking differences regard the speeches on Politics, where most English speakers read texts out loud and the Polish ones predominantly spoke impromptu.

2.1.3 Speakers

The ST-EN subcorpus contains speeches of 65 unique speakers (20 female and 45 male), while the ST-PL includes 57 unique speakers (11 female and 46 male). With 230 ST-EN speeches and 232 ST-PL speeches this gives the average of 3.8 speeches per person (ranging from 1 to 19) in both subcorpora. Since PINC metadata includes precise speaker identification, we will control for the uneven number of speeches in our analyses, whenever possible. We took extra care to exclude any non-native speakers of either language and to only include MEPs.

Interestingly, the majority of speeches delivered by Polish female MEPs were read out while the majority of male MEPs spoke impromptu. This was also true for English-speaking MEPs, although the differences are not as pronounced (Figure 2).

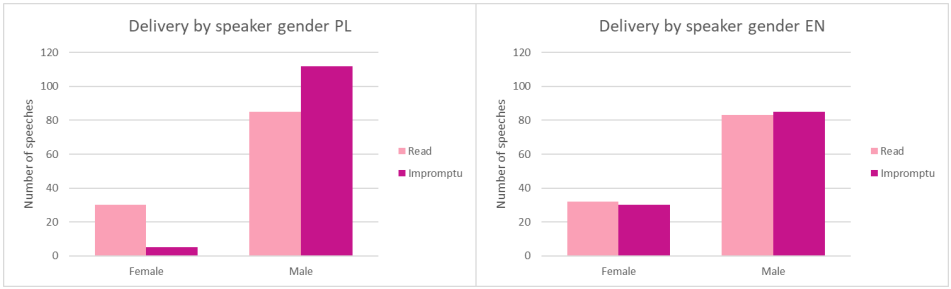


Figure 2: Speeches delivered by female and male MEPs in each mode

2.1.4 Interpreters

Interpreters are key in any interpreting corpus. Professionals working during the European Parliament plenary sessions are carefully selected in a process designed to guarantee top quality interpreting services at the EU institutions. The usual problem with EP data, however, is that the only detail allowing us to distinguish between them is their voice. As most interpreting corpora are compiled by interpreting scholars with no expertise in speaker identification, interpreter identity in the corpora of the EPIC suite is frequently disregarded. Yet, controlling for individual variation is desired in many empirical studies, hence PINC does include precise metadata on interpreter identity. This will greatly enhance the control of the individual variation in further analyses.

Both interpreting subcorpora of PINC consist of slightly more texts interpreted by females. There are altogether 39 different interpreters in the PINC corpus, all

of them Polish natives interpreting both into A (L1) and B language (L2). The TT-EN subcorpus contains texts interpreted by 21 different interpreters (10 female and 11 male) and the TT-PL subcorpus includes productions by 33 interpreters (23 female and 12 male). In most cases, interpreters interpreted both impromptu and read speeches in both directions (Figure 3), which makes it possible, for example, to investigate the same interpreter's interpreting output into two different languages.

As not many interpreters in the European Parliament have Polish as a C language, interpretations from Polish are frequently provided as retour interpretations by interpreters from the Polish booth (with Polish as A and English as B), or as relay interpretations when interpreters from other language booths use Polish-English retour as their pivot and the source input. We deliberately excluded any speeches interpreted via relay. As a result, the TT-EN subcorpus is a retour subcorpus and includes interpretations by the same interpreters who contributed to the TT-PL subcorpus. This offers an interesting opportunity for interlinguistic comparisons that are not between-groups but within-group. This differentiates PINC from other corpora, which include either interpretations into A languages only or which do not strictly control for the language status (A or B) of the interpreters in specific subcorpora.

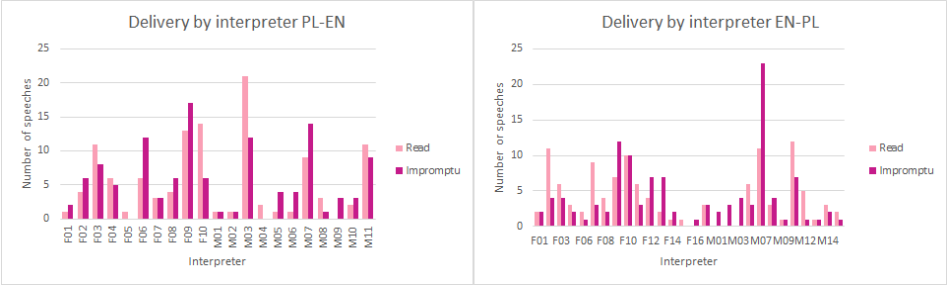


Figure 3: Read and impromptu speeches interpreted by individual interpreters (codes starting with capital F indicate female interpreters, codes starting with M indicate male interpreters)

2.2 Design

2.2.1 Interpreter identification

Identifying interpreters may have presented the greatest technical challenge in building PINC so far. Since the Europarl website provides no information about the individual interpreters and, as opposed to the original speakers, they are not

visually identifiable, in order to distinguish between the voices, we had to employ a three-stage procedure. In stage one, two human compilers who took part in collecting the corpus data (authors of this chapter) labelled each new interpreter in a spreadsheet. These were later proofed by another team member, especially where any doubts as to potential overlaps were expressed. In this manner a pool of potential interpreter voices was identified. In stage two this pool of identified voices was given for verification to an experienced conference interpreter who had worked with the interpreters included in the sample.

Independently of this strictly human-based procedure, in stage three, an automated attempt at interpreter identification was also made. It consisted in comparing the above pool of potential interpreter samples (enrollment data) to the recordings of all 476 interpretations (test data) within the Kaldi Speech Recognition Toolkit (Snyder et al. 2018) trained on large scale, open-source corpora of human voices (development data). The method relies on computing a multidimensional vector representation of an audio segment, known as the x-vector. This vector is computed both for the enrollment data and for all the test data. Next, a Probabilistic Linear Discriminant Analysis algorithm is used to compute a matrix of distances between each file and speaker, thus providing an easy method of assigning the most likely candidate for each file. Interestingly, the comparison of the human-made and automatic judgments yielded very satisfying results as only around 15% of stage-one interpreter judgments have been misassigned. A detailed description of the interpreter voice identification procedure is provided in Koržinek (2020b).

2.2.2 Transcription

As in most of the corpora from the EPIC suite (Bernardini et al. 2018), the source text subcorpora in PINC are based on verbatim reports, i.e. transcripts of the audio/video files of speeches downloaded from the EP website. The EP website offers relatively accurate renditions that had to be manually corrected to a small extent only in order to facilitate speech-text alignment. Unfortunately, the texts of interpretations available on the EP website are actually written translations of the original verbatim reports and thus depart heavily from what was said by the interpreters. Therefore, in the case of target text subcorpora we decided to use an automatic speech recognition system as input for later manual correction. We specifically used Google Cloud Speech as accessed through the WebMaus service (Kisler et al. 2017). To streamline the process of post-editing we set up a simple service based on the Corrector webapp (Koržinek 2019) consisting of a rich audio player (based on wavesurfer.js audio editor and controllable from the keyboard)

and a text field with basic change-tracking capabilities (Figure 4). Its cloud-based storage of results allowed for seamless cooperation between team members.

Figure 4: Corrector-webApp online environment

This application was used not only for correcting transcriptions, but also for manual endpointing, that is marking when the transcription starts and ends within the audio file (the pink areas in the waveform in Figure 4). Unfortunately, each audio recording begins and ends with a portion of speech that has a more dialogic and organisational character, such as the President giving the floor to a particular MEP whose speech is of primary interest in a given file. Thanks to endpointing the alignment tools described below only utilised the audio that perfectly matched the transcription.

In terms of principles our transcription was largely based on guidelines used in EPIC and EPTIC (Bernardini et al. 2018: 27), altered in order to meet the needs of PINC. One such need was the automatic speech-text alignment described below; another was the planned ST-TT alignment on the word level. This required as accurate marking of word-boundaries as possible, including unfinished, self-corrected and distorted words. In this respect we decided to introduce three special symbols: tildes <~> for truncated words, plus signs <+ +> marking the boundaries of filled pauses (the pluses were to surround an approximation of the actual sounds produced by the speaker, e.g. +ehm+) and square brackets <[ ]> to mark any external noises, such as applause, which could be picked up by the system and misinterpreted.

2.2.3 Speech-text alignment

While some corpora in the EPIC suite include only transcripts e.g. EPIC, others contain recordings that are time-aligned at various levels. Most language components of EPTIC are time-aligned with videos of the speeches at sentence-like-utterance level (Ferraresi & Bernardini 2019: 132) using a system of subtitles integrated into NoSkechengine online platform (Rychlý 2007). EPICG includes timestamps at event level aligned in Exmaralda (Schmidt & Wörner 2009). PINC has been automatically time-aligned to audio files of the speeches/interpretations. The word-level alignment was then manually corrected in yet another instance of computer-human interaction employed for best possible results. This time the starting point was automatic segmentation and alignment performed in the Kaldi toolkit (Povey et al. 2011) and based on a Gaussian Mixture-based acoustic model for which the endpointed transcriptions from the previous step were used as input along the audio recordings. Following that, two human aligners manually proofed and adjusted the output using the EMU-webApp (see Figure 5), which is an open-source browser-based labelling and correction tool that allows for a hassle-free cooperative annotation of audio files (Winkelmann & Raess 2014).

As a result all words, pauses and disfluencies are orthographically transcribed, time-stamped and available for analysis. A detailed description of the speech-text alignment in PINC is presented in Koržinek (2020a).

Further processing of the corpus (currently underway) involves POS tagging, text and video alignment, alignment of source texts and target texts on the utterance level and – most importantly – word level. This last alignment is especially crucial for the main objectives driving PINC creation. Apart from a plethora of corpus-driven research, PINC will first and foremost inform corpus-based studies on activation and inhibition as mechanisms of language control in interpret-

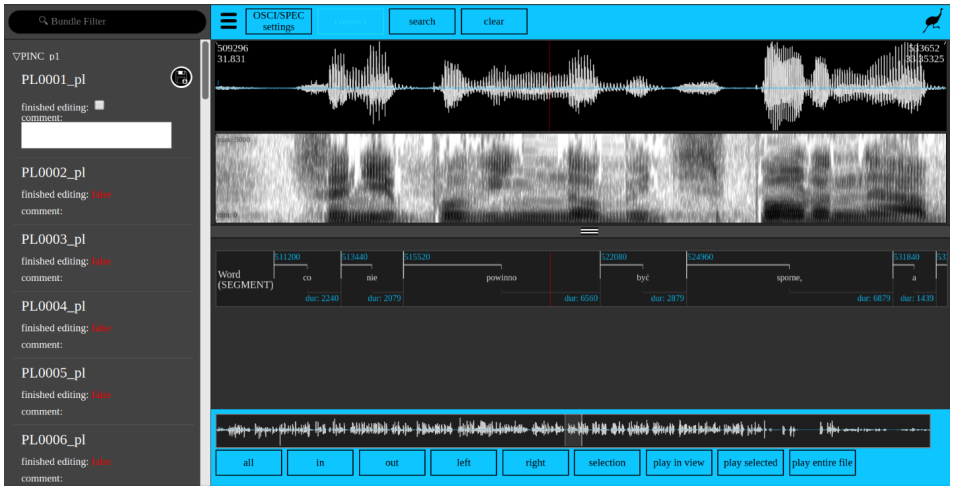


Figure 5: EMU-webApp online environment

ing. This is why precise timestamps are needed for specific words (cognates, homonyms, words with single and multiple translation equivalents), since we are interested, among other things, in the ear-voice span as a processing index of these words.

3 An example study: fluency parameters in interpreting

To show the potential of PINC, we present an example of a study that looks into interpreting fluency parameters, such as speaking rate and pauses. We compared source texts and their interpretations on a number of delivery parameters and tried to identify which factors modulate these parameters in interpreters' outputs. We also wanted to find out if interpreters speed up and compress their target text production when dealing with higher source text delivery rates. Thanks to interpreter voice identification in PINC metadata, we could explore individual differences and control for these differences in our analysis. We conducted the study on the Polish-English subcorpus, so the interpreting examined is performed into the interpreters' B language, i.e. the more demanding interpreting direction (Chang 2005; see also a review in Chmiel 2016).

3.1 Interpretation speed and its modulating factors

Speed of delivery is considered one of the most important input variables in interpreting, which has been shown to affect the quality of interpreting (Riccardi

2015), including omissions (Barghout et al. 2015) or the occurrence of filled pauses (Plevoets & Defrancq 2016). While the majority of studies focus on source text speed as an important factor that influences numerous aspects of interpreters' output, few studies have specifically focused on various factors that affect the target text speed. For instance, Han (2015) found that speech rate in interpreting has a strong correlation with perceived fluency. Below, we use PINC data to see what makes interpreters speed up their production. First, however, we analyse the corpus to compare ST and TT speeds on a number of measures and discuss our results in the context of other available data on comparable corpora.

As mentioned above, the average speaking speed in our Polish-English subcorpus was 126 WPM (SD=15, range: 88–166) for ST and 129 wpm (SD=18, range: 77–178) for TT. These values are considered as low speed of delivery by EPIC standards (Monti et al. 2005) and are lower than those reported for EPICG (158 wpm for ST and 142 wpm for TT) (Collard & Defrancq 2019). The ST speaking speed is also lower than 154 wpm from EPIC reported by Russo (2018) while the TT speed is comparable with the relevant data from the same study (130 wpm).

As languages may differ in word length, some researchers (Riccardi 2015; Seeber 2017; Tissi 2000) pinpoint that speaking speed may also be measured in syllables per minute. When measured this way, the PINC source texts are characterised by a significantly higher speaking rate (M=286 spm, SD=35) than target texts (M=199 spm, SD=27), $t(433)=30.26$, $p<.001$, which results from the fact that Polish words are on average longer (2.27 syllables per word in our corpus) than English ones (1.55 syllables per word in our corpus).

Another important measure of the speed of oral delivery is the articulation rate understood as the average speed of utterance without pauses (Christodoulides 2013; Riccardi 2015). The PINC source texts have a higher articulation rate (measured in syllables per minute) (M=330 spm, SD=36) than target texts (M=248 spm, SD=24), $t(404)=28.48$, $p<.001$.

We also measured the compression rate understood, following Russo (2018), as a relative difference in speech length, expressed in percent and measured according to the following formula: (total ST words – total TT words)*100/total ST words. If the compression rate is 0, the target text equals the source text in length. If it has negative value, the target text is compressed. If the value is positive, the target text is longer than the original. The mean compression rate for the whole subcorpus is 3.6%, which means that the interpretations are slightly longer than the originals. However, there is much variation among individual interpreters, and we can visualise that thanks to exact identification of interpreter voices in the corpus (Figure 6).

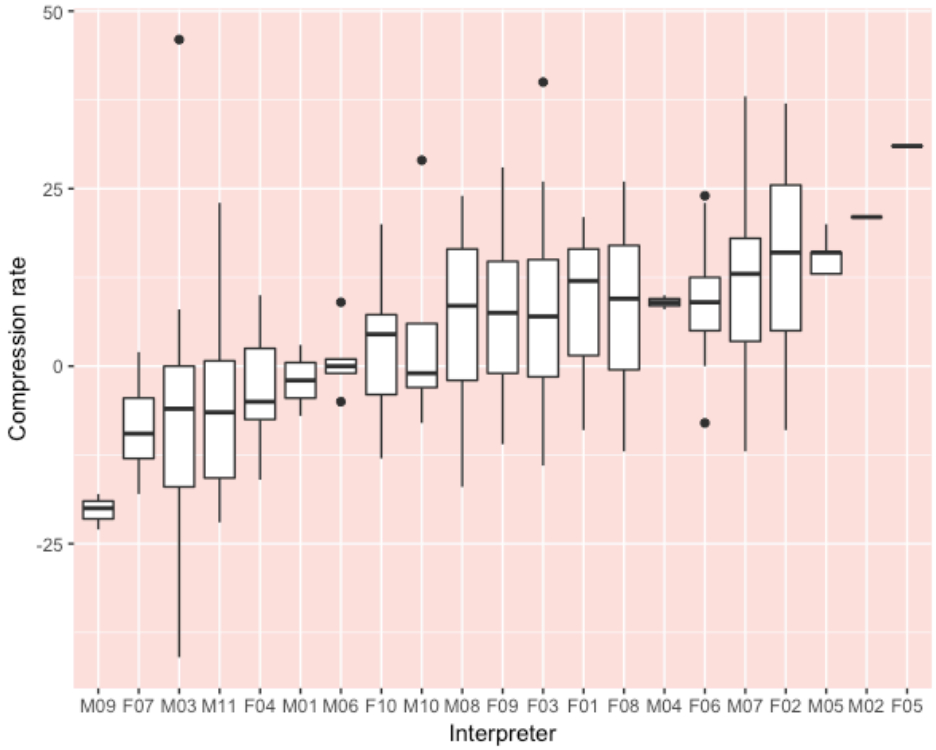


Figure 6: Individual variation of compression rates

There are six interpreters who consistently compress the source text while the majority of interpreters produce longer interpretations than originals, which is quite surprising and at a variance with Russo (2018) but might be triggered by two factors. First, PINC source texts are slower than those analysed by Russo: interpreters might not feel compelled to synthesize if there are no demanding temporal constraints. Second, this analysis pertains to interpretations into B language only and these, as such, might differ in production characteristics from interpretations into A language, for example in terms of opting for more descriptive formulations where precise one-to-one equivalents are not easily retrievable from the mental lexicon. Further comparisons are needed, and they will be possible when PINC, as planned, is extended to include a smaller subcorpus of the same language combination (PL-EN) with interpretations performed into A language.

In order to see whether interpreters speed up their delivery and compress more when processing a fast source text, we fitted three regression models. The data

show that source texts with higher word per minute values lead to interpretations with higher word per minute values (b=.49, SD=.07, t=6.96, p<.001) (Figure 7), with higher articulation rates (b=.54, SD=.10, t=5.35, p<.001) (Figure 8) and higher compression rates (b=-.40, SD=.05, t=-7.19, p<.001) (Figure 9).

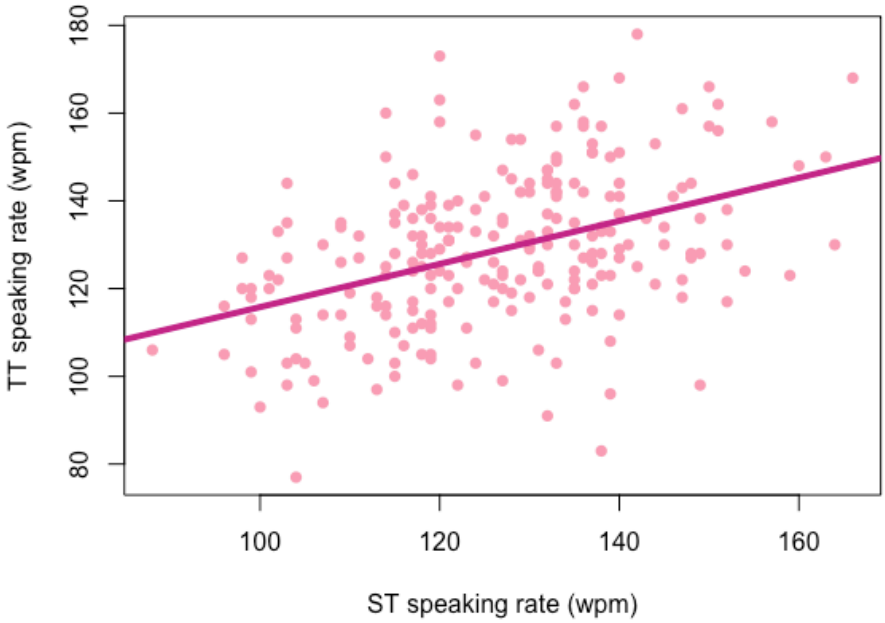


Figure 7: Mean source text speed (words per minute) plotted against target text speed (words per minute)

The results of our analysis regarding how interpreters modulate their output as a result of the source text speed are in line with those by Russo (2018), who also found that faster source texts lead to faster target texts and greater compression, and with those by Gerver (1969) and Barghout et al. (2015), who identified a similar relation between ST speed and TT compression. Slower speaking by interpreters as compared to source text speakers was previously confirmed by Russo (2018) and Christodoulides (2013). This might be explained by the fact that – due to compression – interpreters speak less and thus can slow down.

Taken together, these results show an established pattern of the source text speed affecting interpreters' production in terms of speed and compression. The

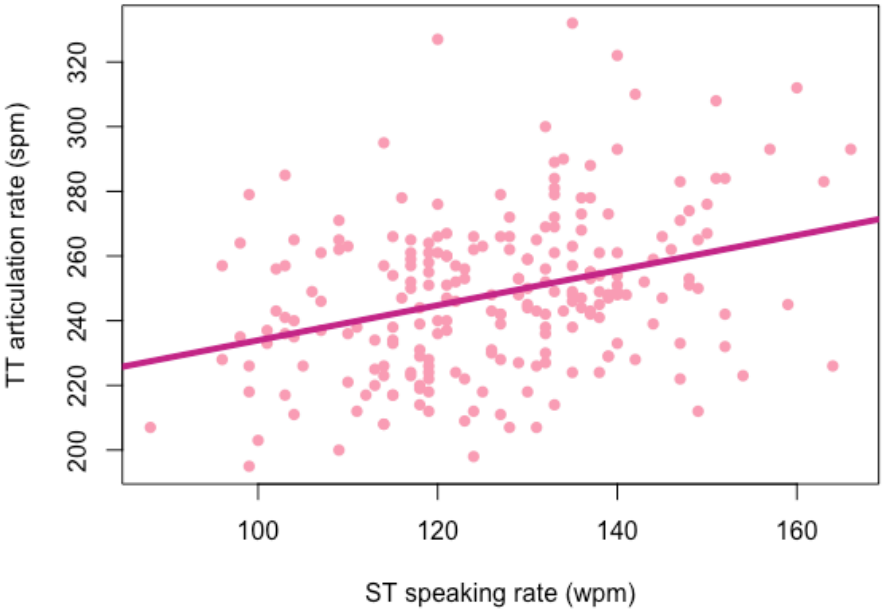


Figure 8: Mean source text speed (words per minute) plotted against target text articulation rate (syllables per minute)

novelty of PINC analysis is the option to better capture individual differences by using specific metadata regarding interpreter voice identification. For instance, in an additional analysis, we used the range of speaking rate of each speaker and interpreter as a variable of speaking rate variability. We excluded from this analysis those speakers and interpreters who only contributed one speech to the data set (as their variability was 0), which left us with data from 42 speakers and 20 interpreters to analyse. It turned out that the individual speaking rate variability of interpreters was much higher (M=40 wpm) than that of speakers (M=19.95 wpm), t(23.8)=4.60, p=.0001. This confirms the results obtained by Christodoulides (2013) on a much smaller corpus based on EP data.

3.2 Comparing other delivery parameters in source and target texts

Many studies compare fluency parameters of source texts and their interpretations to shed more light on the difference between non-mediated and interpreter-

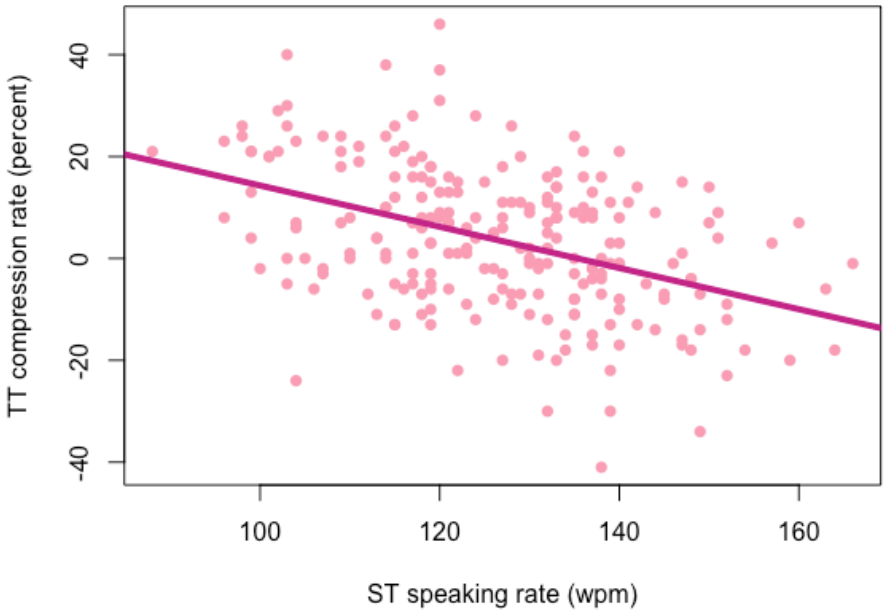


Figure 9: Mean source text speed (words per minute) plotted against compression rate (percent)

mediated texts (Ahrens 2005; Cecot 2001; Pöchhacker 1995; Tissi 2000; Wang & Li 2014). The emerging pattern of data resulting from these experimental studies is that the pausing pattern specific to interpreter-mediated texts includes fewer but longer pauses. Our analysis makes a contribution to the corpus data on ST-TT comparison. The novelty of our findings is that they are based solely on the subcorpus of retour interpretations (i.e. interpretations into the interpreter's B language). Below we compare source and target texts in the Polish-English subcorpus of PINC on a range of parameters other than those related to speed analysed above and pertaining mainly to pauses. We later compare our results to other findings based on other corpora.

The comparisons of PINC source and target texts are presented in Table 2. Duration was calculated from the onset of the first spoken word to the ending of the last word of each speech. Thus, it did not include silence periods before and after the utterance. This is quite important to remember since the TT duration does

not include the initial ear-voice span (EVS) and as such does not capture the dynamics of processing involved in interpretation. A detailed analysis of EVS will be the focus of another study. Further in Table 2, there are four parameters pertaining to silent and filled pauses – reflecting their number (normalised per minute of speech) and mean length. Various thresholds are used in the literature to identify silent pauses, ranging from 200ms (Chmiel et al. 2017; Collard & Defrancq 2019) to 300 ms (Wang & Li 2014). We identified a silent pause as a period of silence longer than 250 ms in line with the majority of studies (Cecot 2001; Han et al. 2020; Mead 2005; Pradas Macías 2006; Tissi 2000). A filled pause was identified as anything marked in transcription as +yyy+ or +eee+ or anything else between two plus signs. We applied no cut-off point for a filled pause following Plevoets & Defrancq (2016). A run was defined as a segment of speech uninterrupted by silent pauses, as applied by Han et al. (2020). Finally, speech proportion was calculated as a ratio of articulation time (i.e. not including pauses) to speech duration (Lee 1999).

Table 2: ST and TT delivery parameters compared

Parameter	ST mean	TT mean	t	p
Duration	1 min 22 s	1 min 24 s	-.28	=.78
Number of silent pauses per minute	12.06	10.36	5.51	p<.001*
Mean length of silent pauses (in ms)	487	626	-6.86	p<.001*
Number of filled pauses per minute	2.84	7.37	-11.10	p<.001*
Mean length of filled pauses (in ms)	613	722	-3.99	p<.001*
Mean length of runs (in syllables)	26.21	20.20	2.56	p=.011*
Speech proportion	0.87	0.80	11.38	p<.001*

Differences in all the parameters apart from duration turned out to be statistically significant. The comparison shows a familiar pattern: interpretations include fewer but longer silent pauses, which is in line with a study involving students by Tissi (2000), a small-scale study of A to B simultaneous interpreting involving professionals and trainees by Wang & Li (2014) and other studies

(Ahrens 2005; Christodoulides 2013; Collard & Defrancq 2019; Lee 1999; Pöchhacker 1995).

It is interesting to see that silent pauses in PINC are much shorter and less numerous than in a comparable corpus (EPICG) involving interpretations from the European Parliament and featuring different language pairs analysed by Collard & Defrancq (2019). In that study, there are almost 23 silent pauses in one minute of ST and 19 silent pauses in one minute of TT. The mean length is 10280 ms and 10580 ms for ST and TT, respectively. It seems that, as compared to PINC, silent pauses in EPICG are approximately twice as long and twice as numerous in all texts. We might speculate that this discrepancy is due to differences in speaking rates, which are much higher in EPICG. A comparison of PINC and EPICG data for filled pauses is also interesting. Despite differences in speaking rates, the numbers of filled pauses per minute match almost exactly across both corpora: 2.61 in EPICG and 2.84 in PINC for source texts and 7.52 in EPICG and 7.37 in PINC for target texts. Unfortunately, Collard and Defrancq do not include data for the mean length of filled pauses. A potential explanation for these results might be the different nature of both corpora. EPICG includes, to the best of our knowledge, only interpretations into the A language, while the subcorpus of PINC under analysis includes retour interpretations only (i.e. into the B language). Since production in one's B language is more difficult than into one's A language and since filled pauses, according to Setton (1999), reflect cognitive load related to formulation, interpretations into B should include more numerous filled pauses than interpretations into A. The reason PINC and EPICG match on this measure might be because the number of filled pauses in PINC is offset by its lower ST and TT speaking rate. This explanation is tentative and the prediction on the higher number of filled pauses present in retour interpretations as compared to interpretations into A will be tested on the Polish-English language pair once PINC is extended to include a subcorpus of PL-EN interpretations made into the interpreters' A language.

To the best of our knowledge, no study to date has compared the mean length of runs of source texts and target texts in a corpus study. Our data show that interpretations include shorter runs, or uninterrupted flows of utterance, than source texts. This might mean that interpreters work in shorter spurts and fragment their output due to processing constraints. Additionally, the speech proportion data are in line with Lee's (1999) results, showing that interpreters speak for a smaller proportion of time than speakers and use pauses for information processing.

In their study of perceived fluency of interpreting, Han et al. (2020) identified the following criteria as strongly associated with higher fluency: mean length of

runs, mean length of silent pauses, phonation time ratio (which is equivalent to speech proportion in the present study) and speech rate. All these criteria have lower values for interpreting than for the source texts in PINC. Although Han et al.'s (2020) data pertain to consecutive interpreting, we might tentatively assume that interpretations in our corpus could be perceived as less fluent than the source texts, although such a conjecture surely requires empirical verification.

3.3 Silent and filled pauses in interpretation and their modulating factors

Fluent delivery is an important criterion in interpreting (Pradas Macías 2006; Rennert 2010) and pauses are generally considered as an important element of fluency (Mead 2000). Pöchhacker (2004) considers silent and filled pauses as part of the dis-fluency phenomenon in interpreting related to the limited scope of planning involved in this type of oral production. Silent pauses are associated with problems with ST comprehension, lexical search for translation equivalents and production difficulties (Bartłomiejczyk 2006, Piccaluga et al. 2005, Tóth 2011). Although interpreters tend to follow the general pattern of pauses applied by the speaker, there are modifications to that pattern due to difficulties in processing (Cecot 2001; Goldman-Eisler 1972). Filled pauses can be interpreted as an indirect index of cognitive load and, similarly to silent pauses, can reflect processing difficulties. According to Setton (1999), while long silent pauses indicate high attention to input, long filled pauses reflect attention to formulation including speech planning and lexical access. Both silent and filled pauses turned out to be longer in interpretations than in source texts, thus testifying to extreme speech production conditions in interpreting.

We fitted a series of mixed effects linear models to evaluate the impact of ST fluency parameters on pauses in the target texts. Since PINC metadata include exact identification of interpreter voices, we could include interpreters as a random factor in each model. Fixed factors reflected the source text delivery characteristics, such as speed, number of silent pauses per minute, mean length of silent pauses, compression rate and delivery mode (whether the source text was read or delivered impromptu). We could not include the number and mean length of filled pauses due to the violation of the collinearity principle – both of these measures correlated moderately with the source text speaking rate. We used sliding contrasts for delivery mode and treatment contrasts for the remaining fixed factors. P values were obtained through Satterthwaite approximations. The number of silent pauses per minute in the interpretation was influenced by the ST speed (b=-.02, SE=.01, t=-3.03, p=.002), mode of delivery (b=-1.04, SE=.44,

t=-2.36, p=.02) and the number of silent pauses per minute in the source text (b=.15, SE=.07, t=2.23, p=.03). The faster the ST, the lower the number of silent pauses in TT. There are more silent pauses in interpreting when interpreting read out rather than impromptu speeches and when there are more silent pauses in ST. This last association is in line with Collard & Defrancq (2019). It seems that interpreters pause less when dealing with faster and read out source texts, but they pause more when the speakers pause more. Interestingly, the mean length of silent pauses in TT was not modulated by any factors, which is at a variance with Collard and Defrancq's study, where source text speaking rate did modulate the length but not the number of silent pauses in interpreting.

The number of filled pauses in TT was modulated by the mode of delivery (b=1.24, SE=.58, t=2.12, p=.03) and the compression rate (b=-9.67, SE=2.39, t=-4.05, p<.001). The data show that as compression increases and the interpretation becomes shorter, the number of filled pauses increases. There were also more filled pauses in read out speeches as compared to the impromptu ones. As postulated earlier, this might be related to the cognitive load triggered by increased reformulation involved in producing more compressed and structurally less complex target texts. The mean length of filled pauses in TT was modulated by the ST speed (b=-1.94, SE=.56, t=-3.45, p<.001) and the compression rate (b=-673.42, SE=150.58, t=-4.47, p<.001). These results show that the faster and the less compressed the source text, the lower the mean length of filled pauses in the interpretation. None of the predictors associated with the number of filled pauses in this study match those in EPICG (Collard & Defrancq 2019). In that study, ST speed influenced the number of filled pauses while in ours – it influenced the length of filled pauses. More research is needed to elucidate the phenomenon of filled pauses in interpreting.

Taken together, our data on factors modulating pauses are partially in line with Setton's (1999) general idea of the relationship between silent pauses and the focus on the ST input and between filled pauses and the focus on formulation. Only filled pauses were modulated by the compression rate – they became longer and more numerous as interpreters struggled to provide a more compressed, i.e. more reformulated version of the target text. Mode of ST delivery influenced the number of both silent and filled pauses. They were more numerous in interpretations of read out speeches. One may assume that silent pauses helped interpreters' comprehension of these speeches that are usually lexically denser and structurally more complex. Filled pauses, on the other hand, aided formulation, which was also more demanding as compared to impromptu speeches that are usually more similar in structure complexity to oral production involved in interpreting. However, these conjectures require further empirical support. The

study by Wang & Li (2014) constitutes an interesting attempt at providing detailed explanations of various categories of pauses thanks to a combination of experimental data with retrospective protocols. Alas, no differentiation between motivations for silent and filled pauses is made. This definitely is a promising research avenue worth pursuing in the future.

4 Conclusions

PINC offers excellent research material that is well-balanced, considering the external constraints. Issues of justice and politics predominate the topics of speeches, while as far as gender distribution is concerned, the majority of speakers are male and the majority of interpreters are female. This mirrors the European Parliament reality – male MEPs still dominate the chamber while interpreting is unceasingly a profession dominated by females (which is also true for experimental studies as gender balance is difficult to gain when recruiting study participants). The PINC creation workflow offers new tools and automation opportunities for future corpus developers.

Thanks to using similar categories of metadata (topics, mode of delivery), PINC will be easily comparable to other corpora from the EPIC suite, which should facilitate studies that involve various language combinations to control for language-pair-specific factors. Interestingly and due to the language regime and language profiles of interpreters in the European Parliament, PINC includes a strictly controlled Polish-English subcorpus of retour interpreting, an added value as compared to other existing corpora. In the future, it will also include a smaller subcorpus of Polish-English interpretations into A language. This offers a lot of potential for various novel comparisons in corpus-driven studies. We can compare interpretations by the same interpreters working into A (EN-PL) and B (PL-EN). We can also compare interpretations in the same direction (PL-EN) by two different groups of interpreters – native-speakers of English and interpreters with English as their B language.

Our initial exploratory corpus-driven study shows how important it is to apply various variables since not all of them are sensitive enough to capture differences. Our ST-PL and TT-EN corpora differed in speaking rate measured in syllables per minute but not in words per minute. Interestingly, the mean compression rate was slightly positive, meaning that target texts were actually longer than source texts. However, a detailed analysis of individual differences showed compression as an interpreter-specific feature. We found that interpreters speed up and compress their delivery more when the source text is delivered faster, showing an

expected pattern of results in line with previous studies. Our source and target texts differed also on a range of other fluency criteria, such as number and mean length of silent and filled pauses. We also applied another measure of fluency – mean length of runs (i.e. utterances uninterrupted by pauses) and found interpreters to produce more fragmented output due to processing constraints. Our findings show that interpreters produce more silent and filled pauses when interpreting a read-out text. More numerous silent pauses in the source text also increase the number of such pauses in the target text. Additionally, the number and length of filled pauses increase with increased compression rate, which seems to suggest that filled pauses could be a good index of production problems.

PINC has been created mainly to expand our knowledge about language control mechanisms (activation and inhibition) in interpreters on the basis of naturalistic data and to serve as a source of stimuli for future experimental studies. However, we hope that PINC, with its intended open access format, rich annotation and in-built interpreter identification will also help interpreting scholars find answers to many interesting corpus-driven research questions.

References

Ahrens, Barbara. 2005. Prosodic phenomena in simultaneous interpreting: A conceptual approach and its practical application. *Interpreting* 7(1). 51–76. DOI: 10.1075/intp.7.1.04ahr.

Barghout, Alma, Lucía Ruiz Rosendo & Mónica Varela García. 2015. The influence of speed on omissions in simultaneous interpretation: An experimental study. *Babel* 61(3). 305–334. DOI: 10.1075/babel.61.3.01bar.

Bartłomiejczyk, Magdalena. 2006. Strategies of simultaneous interpreting and directionality. *Interpreting* 8(2). 149–174. DOI: 10.1075/intp.8.2.03bar.

Bartłomiejczyk, Magdalena. 2016. *Face threats in interpreting: A pragmatic study of plenary debates in the European Parliament.* Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Doctoral dissertation). https://wydawnictwo.us.edu.pl/sites/wydawnictwo.us.edu.pl/files/face_threats_in_interpreting_czw_st_e.pdf.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 21–42. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-6199-8_2.

Cecot, Michela. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The interpreters' newsletter* 11. 63–85.

Chang, Chia-chien. 2005. *Directionality in Chinese/English simultaneous interpreting: Impact on performance and strategy use.* Austin: University of Texas. (Doctoral dissertation).

Chmiel, Agnieszka. 2012. Pamięć operacyjna tłumaczy konferencyjnych mierzona metodą RSPAN. In Maria Piotrowska (ed.), *Kompetencje tłumacza*, 137–154. Kraków: Tertium.

Chmiel, Agnieszka. 2016. Directionality and context effects in word translation tasks performed by conference interpreters. *Poznan Studies in Contemporary Linguistics* 52(2). 269–295. DOI: 10.1515/psicl-2016-0010.

Chmiel, Agnieszka. 2018. Meaning and words in the conference interpreter's mind: Effects of interpreter training and experience in a semantic priming study. *Translation, Cognition & Behavior* 1(1). 21–41. DOI: 10.1075/tcb.00002.chm.

Chmiel, Agnieszka, Agnieszka Szarkowska, Danijel Koržinek, Agnieszka Lijewska, Łukasz Dutka, Łukasz Brocki & Krzysztof Marasek. 2017. Ear–voice span and pauses in intra- and interlingual respeaking: An exploratory study into temporal aspects of the respeaking process. *Applied Psycholinguistics* 38(5). 1201–1227. DOI: 10.1017/S0142716417000108.

Christodoulides, George. 2013. Prosodic features of simultaneous interpreting. In Piet Mertens & Anne-Catherine Simon (eds.), *Proceedings of prosody-Discourse Interface Conference 2013 (IDP-2013)*, 33–37. Leuven: Unviersity of Leuven.

Collard, Camille & Bart Defrancq. 2019. Disfluencies in simultaneous interpreting a corpus-based study with special reference to sex. In Lore Vandevoorde, Joke Daems & Bart Defrancq (eds.), *New empirical perspectives on translation and interpreting*, 264–299. New York: Routledge. http://hdl.handle.net/1854/LU-8581686 (5 February, 2019).

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies*, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.

Dumara, Barbara. 2015. *How can interpreting corpora extend our knowledge on intrusive "we" in SI?* Poster presented at Corpus-based Interpreting Studies: The State of the Art. First Forlì International Workshop. Forli, Italy.

Ferraresi, Adriano & Silvia Bernardini. 2019. Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In Irene Doval & M. Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*, vol. 90 (Studies in Corpus Linguistics), 123–139. Amsterdam/Philadelphia: John Benjamins.

Gerver, David. 1969. The effects of source language presentation rate on the performance of simultaneous conference interpreters. In *Proceedings of the Second Louisville Conference on rate and/or frequency-controlled speech*, 162–184. Louisville: Center for Rate-Controlled Recordings, University of Louisville.

Goldman-Eisler, Frieda. 1972. Segmentation of input in simultaneous translation. *Journal of Psycholinguistic Research* 1(2). 127–140.

Han, Chao. 2015. (Para)linguistic correlates of perceived fluency in English-to-Chinese simultaneous interpretation. *International Journal of Comparative Literature and Translation Studies* 3(4). 32–37. DOI: 10.7575/aiac.ijclts.v.3n.4p.32.

Han, Chao, Sijia Chen, Rongbo Fu & Qin Fan. 2020. Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting* 22(2). 211–237. DOI: 10.1075/intp.00040.han.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu. (Doctoral dissertation). https://repozytorium.amu.edu.pl/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. DOI: 10.1016/j.csl.2017.01.005.

Koržinek, Danijel. 2019. *Corrector*. https://github.com/danijel3/Corrector.

Koržinek, Danijel. 2020a. *Automating word segmentation*. PINC Project. https://pincproject2020.wordpress.com/2020/04/08/automating-word-segmentation/.

Koržinek, Danijel. 2020b. *Speaker identification*. PINC Project. https://pincproject2020.wordpress.com/2020/04/28/speaker-identification/.

Lee, Tae-Hyung. 1999. Speech proportion and accuracy in simultaneous interpretation from English into Korean. *Meta: journal des traducteurs/Meta: Translators' Journal* 44(2). 260–267. DOI: 10.7202/003443ar.

Mead, Peter. 2000. Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10(200). 89–102.

Mead, Peter. 2005. Directionality and fluency: An experimental study of pausing in consecutive interpretation into English and Italian. *Communication and Cognition. Monographies* 38(1-2). 127–146.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli & Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta: Journal des traducteurs/Meta: Translators' Journal* 50(4). DOI: https://doi.org/10.7202/019850ar.

Piccaluga, Myriam, Jean-Luc Nespoulous & Bernard Harmegnies. 2005. Disfluencies as a window on cognitive processing. An analysis of silent pauses in simultaneous interpreting. In Jean Véronis & Estelle Campione (eds.), *Disfluency in spontaneous speech*, 151–155. ISCA.

Plevoets, Koen & Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* 11(2). 202–224. DOI: 10.1075/tis.11.2.04ple.

Pöchhacker, Franz. 1995. Slips and shifts in simultaneous interpreting. In Jorma Tommola (ed.), *Topics in interpreting research*, 73–90. Turku: University of Turku, Centre for Translation & Interpreting.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London: Routledge.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer & Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1–4. Waikoloa: IEEE Signal Processing Society.

Pradas Macías, Macarena. 2006. Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting* 8(1). 25–43. DOI: 10.1075/intp.8.1.03pra.

Rennert, Sylvi. 2010. The impact of fluency on the subjective assessment of interpreting quality. *The Interpreters' Newsletter* 15. 101–115.

Riccardi, Alessandra. 2015. Speech rate. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 1st edn., 397–399. London: Routledge.

Russo, Mariachiara. 2018. Speaking patterns and gender in the European Parliament Interpreting Corpus: A quantitative study as a premise for qualitative investigations. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, 115–131. Singapore: Springer Singapore.

Rychlỳ, Pavel. 2007. Manatee/Bonito-A Modular Corpus Manager. In Petr Sojka & Aleš Horák (eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, 65–70. Brno: Masarykova Univerzita.

Schmidt, Thomas & Kai Wörner. 2009. EXMARaLDA |Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19(4). 565–582.

Seeber, Kilian G. 2017. Interpreting at the European Institutions: Faster, higher, stronger. *CLINA: Revista Interdisciplinaria de Traducción, Interpretación y Comunicación Intercultural* 3(2). 73–90. DOI: 10.14201/clina2017327390.

Setton, Robin. 1999. *Simultaneous interpretation: A cognitive-pragmatic analysis* (Benjamins Translation Library 28). Amsterdam: John Benjamins. DOI: 10.1075/btl.28.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: Journal des traducteurs* 43(4). 486–493. DOI: 10.7202/004136ar.

Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey & Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. Calgary, AB: IEEE. DOI: 10.1109/ICASSP.2018.8461375.

Tissi, Benedetta. 2000. Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis. *The Interpreters' Newsletter* 10. 103–127.

Tóth, Andrea. 2011. Speech disfluencies in simultaneous interpreting: A mirror on cognitive processes. *SKASE Journal of Translation and Interpretation* 5(2). 23–31.

Wang, Binhua & Tao Li. 2014. An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives* 23(1). 124–142. DOI: 10.1080/0907676X.2014.948885.

Winkelmann, Raphael & Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4129–4133. Reykjavik: European Language Resources Association (ELRA).

Chapter 4

Migration in EP plenary sessions: Discursive strategies for the Other construction and political Self representation in Italian to Spanish interpreter-mediated texts

Ilaria Anghelli[a] & Laura Mori[b]

[a]Centro ricerche INAIL [b]Università degli Studi Internazionali di Roma - UNINT

This paper deals with transfer of meaning or lack of thereof in interpreting from Italian to Spanish of EP speeches delivered within the framework of Parliamentary debates during plenary sessions dealing with the phenomenon of migration. Political discourse on this topic tends to be characterised by polarity of in-group vs. out-group ideologies expressed through discursive strategies and ethnopragmatic devices that embody Self representation and the (negative/positive) construction of the Other. Our objective is to describe how migrants are linguistically represented (referential strategies), what qualities and traits are attributed to them (predicational strategies) and what are the argumentations and the forms of mitigation and intensification used to convey the political ideology toward the topic of migration. Therefore, a preliminary discourse analysis was conducted on the EP original speeches in Italian in order to pinpoint referential expressions used to designate social actors involved in international migration and the social phenomenon in itself together with the predicational strategies used to discuss on this and the organisation of the entire argumentation flow in accordance with speakers' political pro- or anti- immigration stance. In fact, within a political environment, speakers might manifest their ideology and attitudes through their pragmalinguistic behaviour, which plays a fundamental role in building his/her political Self. Thus, beyond the locutionary aim of any political statement, interpreters of political speeches are asked to render the perlocutive dimension of the political message enacted in the original. Therefore, our main goal will consist in evaluating, through a contrastive qualitative discourse analysis the mediation strategies adopted to preserve

(or conversely alter or even distort) politicians' intentionality and to detect cues of mitigation and/or intensification of the original pragmatic intent.

1 Introduction: EU political discourse on migration

In van Dijk's words "discourse analysis is not a method, but a broad, multidisciplinary field of study of the humanities and social sciences, a field that therefore should rather be called Discourse Studies" (van Dijk 2018: 227).

In political discourse politicians convey their political Self through linguistic encoding and through the enhancement or mitigation of their agency while being constantly in the process of constructing their self-image by using indexical signs and performative devices. In doing so, they express their own involvement and accountability regarding what they do, what they say and who they are. The construction of political Self tends to be affected by a polarity between in-group vs. out-group ideologies expressed through ethnopragmatic devices that embody self-representation and the (negative/positive) Other-presentation (see Duranti (2006)).

Political discourse may be intended as a genre, that is to say "a socially ratified way of using language in connection with a particular type of social activity" (Fairclough 1995: 14) and it may be interpreted by means of the Critical Discourse Analysis (CDA, henceforth). As a matter of fact, CDA provides a theoretical framework to deal with "the discursively enacted or legitimated structures and strategies of dominance and resistance in social relationship of *class*, *gender*, *ethnicity*, *race*, *sexual orientation*, *language*, *religion, age*, *nationality* or *world-region* (van Dijk 1995: 18, italics in original).

The aim of the CDA approach focuses on the analysis of the complex dialectical interplay of language and social practice in discourse and "much work is about underlying *ideologies* that play a role in the reproduction of or resistance against dominance or inequality" (van Dijk (1995: 18), italics in original). Therefore, CDA studies have developed successful methodological techniques to detect contextual cues of ideological perspectives and speakers' attitude toward ethical issues[1] as well as adequate heuristic tools when considering mediated discourse in institutional settings, such as EU ones, where implications related to multicultural and multilingual communication have a potentially high impact.

[1]For reference studies focusing on the CDA approach: **WodakVanDijk2000**, **VanDijk2001**, **VanDijk2006**, Wodak 1996, 2001, 2015, Reisigl & Wodak 2001, van Dijk 1995. Migration discourse analysis is discussed in **WodakVanDijk2000**, Rojo & van Dijk 1997, Wodak 2015.

This framework has also proved to be particularly useful in observing how interpreters handle conveying the evaluative components and metaphorical components encoded in original political texts (see **Boyd2016**) considered that the interpreter's degree of participation may vary with the mode of interpreting (simultaneous or consecutive), thus creating a difficulty in assigning a stable role to him/her, in terms of addressee or side-participant (Pöchhacker 2004).[2]

From this perspective, in this study we are focusing on the contextual frame of Parliamentary debates which are expected to demonstrate evidence of politicians' perspective toward a social phenomenon, such as migration, which drives opposing political ideologies. Immigration policy has a key-role within the EU political framework[3] and the necessity for a systematic international cooperation[4] to face this phenomenon has been highlighted by MEPs asking for reforms to manage migration pressure in light of the following EP standpoint:[5]

> the EU aims to set up a balanced approach to managing regular immigration and combating irregular immigration. Proper management of migration flows entails ensuring fair treatment of third-country nationals residing legally in Member States, enhancing measures to combat irregular immigration, including trafficking and smuggling, and promoting closer cooperation with non-member countries in all fields. It is the EU's aim to establish a uniform level of rights and obligations for regular immigrants, comparable to that for EU citizens. (European Parliament 2021: 1)[6]

In order to face irregular immigration the European Union has signed agreements for reciprocal cooperation between EU Member States and Third Countries and some relevant directives have been enacted,[7] such as the Council Directive 2001/55/EC "on minimum standards for giving temporary protection in

[2] See §2 for more considerations on the interpreter's role.

[3] The legal basis for control over borders, asylum and immigration may be found in articles 77, 78, 79 and 82 of the Treaty on the functioning of the European Union.

[4] In order to safeguard the area of freedom, security and justice Frontex, the European Border and Coast Guard Agency, is aimed at supporting at the external borders to guarantee free movement. Frontex has three strategic objectives: to reduce vulnerability of the external borders based on comprehensive situational awareness; to guarantee safe, secure and well-functioning EU borders, and to plan and maintain European Border and Coast Guard capabilities (https://frontex.europa.eu/about-frontex/foreword/). For more details see Di Giambattista et al. (2015: 17).

[5] In this regard see speeches by the following politicians: (SI7), (RC8), (SS5).

[6] https://www.europarl.europa.eu/RegData/etudes/fiches_techniques/2017/N54569/doc_en.pdf

[7] See Di Giambattista et al. (2015) for the legal background concerning EU policy on irregular migration, management and security of external borders, asylum and legal migration.

the event of a mass influx of displaced persons and on measures promoting a balance of efforts between Member States in receiving such persons and bearing the consequences thereof"[8], Directive 2008/115/EC of the European Parliament and of the Council "on common standards and procedures in Member States for returning illegally staying third-country nationals" and Directive 2009/52/EC of the European Parliament and of the Council "providing for minimum standards on sanctions and measures against employers of illegally staying third-country nationals."

In such a multilingual and multicultural context, European Parliamentary debates could enhance common discursive practices and, at the same time, reveal cross-speakers' differences related to the socio-political background of politicians belonging to different political parties. Having in mind this pragmatic complexity our research question concerns the solutions interpreters adopt to face the mediation of ethnopragmatically-oriented choices and discursive strategies on migration that reveal the politician's stance.

In order to answer this research question, in the following sections we are going to focus on the interpreter's role and on the challenges for interpreters in the EP setting (§2). Afterwards, we will present our methodological apparatus (§3) to discuss the most relevant results with bilingual examples (original Italian; interpreted Spanish) in §4 Conclusive remarks are reported in §5.

2 Mediation of EP speeches: advantages and disadvantages for interpreters and their role

During a plenary session, about 1000 interpreters are involved to cover all the EU official languages. As pointed out by Bartłomiejczyk (2016), plenary sessions constitute a more demanding setting than, for example, meetings of political groups or committee meetings, due to the quick succession of speakers and to the variety of languages spoken on the floor. Following the criteria for time allocation within the EP, a large political group may have up to five minutes, while a small group may only have one minute, thus affecting the delivery of speeches and the interpretations provided by official interpreters. In fact, as suggested by Vuorikoski (2004: 79):

[8]It is worth-mentioning that intertextual references to this Directive are frequent in speeches here under examination in §4: MM8, SA8, SI8, DS8, AP8, MS8, and FP12 (See Appendices A–BAnnex 1–2).

> this results in speeches that are written and recited at fast rate. Furthermore, this rule may also explain why the speeches tend to be extremely dense regarding their information content.

It would appear that the average speaking rate in a plenary session is 150 words per minute (Monti et al. 2005), while the optimal speed is around 95–120 words per minute (quoted in Bartłomiejczyk (2016: 52)). Therefore, it can be assumed that the majority of EP plenary speeches are faster than what is considered comfortable to interpret.

Moreover, Marzocchi (1998: 69) and Kent (2009: 63) remark that interpreters face problems related to the oral delivery of written texts, with the specific prosody related to reading aloud, the lesser redundancy, and other obstacles due to the syntactic and semantic complexity of planned, written speech, as well as the lack of fluency some MEPs may have in the *institutional lingua franca* they use (usually English). It must be said that these problems are caused by the unavailability of transcripts of speeches and the "linguistic behaviour" of most speakers in the session, despite the efforts made to limit the difficulties of interpreting, such as the distribution of leaflets to the MEP on how to communicate through interpretation.[9]

Finally, it seems that interpreters feel frustrated because of the lack of an effective debate and because communication among participants in the sitting is not the primary goal for speakers. In fact, according to Kent (2009: 57):

> Although described as debate the speeches given by Members during plenaries are mainly directed to consumption by home country audiences via the internet, television and radio rather than as engagement with colleagues who are in the room.

Nonetheless, there are some common linguistic features and fixed structures in most speeches that, thanks to their predictability and pragmatic inference, may turn out to be an advantage for the interpreter. First of all, these speeches belong

[9] Among the recommendations (here summarised), delivered by Vice-President Miguel Angel Martínez during his speech on 25 March 2009: "speak at a regular speed, and not too fast, speak in your mother tongue (if possible), avoid changing language when you speak, speaking is better than reading, but if there is no alternative to reading, make sure that interpreters have the text, clearly give references to documents, articulate clearly any figure that is mentioned, explain abbreviations that you use in what you say, remember that jokes are difficult to translate. Also, when you are chairing a meeting, wait a moment before giving the floor to the next speaker so that the interpreter can finish the speech and change to the appropriate channel" (Bartłomiejczyk 2016: 55).

to the argumentative text-type and they share highly ritualised conventions that help the interpreter and allow her/him to focus more on less predictable statements.[10] In this regard, Vuorikoski (2004) states that oral texts belonging to this genre are composed of:

- an introduction where deference to the previous speaker or greetings to the President, the Vice-president or other members are enounced;
- the main body where the speaker's stance is presented;
- final remarks.

During their speech, MPs may offer an important clue as to the opinions that they express belonging to one of the existing political groups and to the expected level of formality they usually adopt (Marzocchi 1998: 66).[11] This has an impact on interpreting in terms of the degree of planning of speeches, taking into account register shifts, rhetorical purposes and for handling with prosody.

In these communicative events, the interpreter assumes a fundamental social role,[12] that is intended as "a set of more or less normative behavioral expectations associated with a 'social position'" (Pöchhacker 2004: 147). This especially refers to legal interpreters and interpreters in healthcare, domains where cultural differences and unfamiliar contexts enhance their role as facilitators, intercultural experts and visible agents.[13] In this respect, Pöchhacker (2004: 149) reported two studies conducted by **Morris1989** and Shlesinger (1991) on simultaneous interpreting, showing how interpreters were responsible for omissions and stylistic changes, thus leading to a "sort of intrusiveness (as perceived by participants) or latitude (as perceived by the interpreters themselves)".

As far as the EP setting is concerned Beaton (2007) examined the impact of the simultaneous interpretation on ideology. This study provides examples of interpreter mediation and agency by defining him/her as "an additional subjective actor in heteroglot communication" (Beaton 2007: 271).

More recently, researchers investigated on the visibility or invisibility of the interpreter, focusing on how he/she plays an active role in the communicative

[10] See Gile's effort model (Gile 1995: 169–170) for a study on this topic.

[11] However, it has to be noted that "MEPs' identity is inherently hybrid and in view of the weakness of the existing whipping system in the European Parliament, they might as well give priority to national or regional interests over the interests of their own political group." (Beaton 2007: 105–108).

[12] See Anderson (2002) for a preliminary analysis on the role of the interpreter.

[13] As described in studies by **LasterTaylor1994**, **Barsky1996** and **Angelelli2001** respectively and summarized by Pöchhacker (2004: 147–149).

event. As a matter of fact, Beaton-Thome (2013) underlined the visible role of interpreters considering that they tend to select more neutral terms than those used in the original speakers, making the ideological stance less pronounced. On the other hand, there are also examples where the interpreter intensifies the speaker's ideological stance by explicitating something that was stated implicitly in the original.

On this regard Bartłomiejczyk (2016: 128) highlights that studies on conference interpreting (e.g. Diriker 2004, Monacelli 2009) reassessed the role of the interpreter by stating the impossibility to expect that the presence of an interpreter mediating between the speaker and the hearer(s) will not have any influence on facework in the interaction (Bartłomiejczyk 2016: 128).

3 Methodology

In our paper, we focus on migration discourse related to the Parliamentary speeches that are "distinguished by genre-specific linguistic forms and/or structures and are closely linked to specific social and institutional contexts" (Fairclough 2006: 32). EP speeches may be considered as belonging to a specific sub-genre[14] since Parliaments represent peculiar *loci* for evaluating social use of language and discursive strategies aimed at persuading, negotiating, building opinions in relation with the reference political party:

> the discourse of parliament results in (or is the final stage of a process which results in) concrete action in the outside world, establishing regulations as to what must, may and may not be done in a given society. (Bayley 2004: 12)

The discursive interaction within EP debates complies with a prototypical frame acknowledged by any member of the EP: a "context model" (see van Dijk 2003) shared in accordance with the specific setting (location, time), participants (and inter-personal relations), activities and actions in which MPs are engaged as political and institutional actors willing to affirm their political Self. More specifically, in EP plenary sessions, the President of the European Parliament chairing the session assisted by the 14 vice-presidents opens the sitting with a speech on the current topic. During a Parliamentary debate, any speaker plays a communicative role (by expressing his/her own opinions or acting as the spokesperson

[14] Parliamentary debates have started to be investigated in the literature by a number of scholars, but for the most part studies focused on mainly on political national Parliaments rather than at supranational level. For more details see Ilie (2015: 5–6).

of his/her party), an interaction role (opponent, enemy or ally) or a social role (based on the group, class, and the ethnicity identified with).[15]

3.1 Research goal

Over the last few years some empirical studies have focused on the multilingual functioning of the European Parliament (EP) generating insights into the interpreters' role (such as Bartłomiejczyk 2016, Beaton-Thome 2013, Kučiš & Majhenič 2018). From a pragmatic perspective it is interesting to consider the filtering effect that could cause misinterpretation of the speaker's illocutionary force, intention and attitude. In such a context, in fact, simultaneous interpreting has to comply with interactive patterns featuring plenary debates as well as with the consumption by home country audiences (see Kent (2009: 57)). In this way, it is relevant to consider both audiences, colleagues in presentia and external public in absentia and different speakers' pragmatic intentions and their degree of engagement toward these two targets.

Our research goal is to evaluate the implications of oral mediation into Spanish of EP migration discourses of Italian politicians as far as how migrants are linguistically represented in EP discourses by focusing on referential strategies, what qualities and traits are attributed to them (by means of predicational strategies) and what are the argumentations and the forms of mitigation and intensification used to convey speakers' political ideology.

Moreover, it must also be borne in mind that political discourse about migration may be seen as a social practice thanks to which speakers act ethnopragmatically in order to build their public image in relation to a socially sensitive topic.

3.2 The collection of data

The dataset was collected from a corpus based on 60 speeches delivered in European Parliament debates during plenary sessions about migration-related issues by twenty-five MPs (15,311 tokens) and their interpretations into Spanish (16,997 tokens).[16] Speeches were selected[17] in accordance with the following external

[15]This could be detected by examining the shift in the use of allocutives: first personal singular vs. first person plural. Similarly, the opposition in the linguistic representation of ingroupness can be analysed through deictics in the distribution between Us and Them.

[16]Speeches are available on the following website: https://www.europarl.europa.eu/plenary/en/debates-video.html#sidesForm

[17]The selection of criteria complies with the research design outlined in Anghelli's (2019) MA thesis from which this study derives.

variables:

- time: in a given timespan corresponding to the 8th parliamentary term (2009–2014);[18]
- topic: the semantic field of migration[19] to lead a topic-oriented research;
- speakers' political profile: speakers belonging to ALDE (Alliance of Liberals and Democrats for Europe Party); EPP (European People's Party); S& D (Progressive Alliance of Socialists and Democrats) and EFD (Europe of Freedom and Democracy).

The European Parliament website only provides audio materials of the interpretations but not the transcription of the interpretations that was carried out in order to lead this current study.[20] In Annex 1 and 2 more details concerning speeches dealt with in §4 are reported.

3.3 The theoretical framework

Within the CDA theoretical framework, useful methodological approaches were developed to detect contextual cues of ideological perspectives (see **VanDijk2006**, **VanDijk2015**, Wodak 1996, 2001, 1996, Reisigl & Wodak 2001, van Dijk 1995) and discursive strategies of positive self- and negative Other construction. This can be seen in studies specifically devoted to dealing with the field of action[21] of migration such as **WodakVanDijk2000** on the discursive strategies used by politicians from six Western European countries (Austria, France, Germany, Great

[18] The selection of texts, collected before the end of the 9th parliamentary term, was limited to speeches in Italian where the interpreted versions were available. Basically, the speeches analysed were delivered in a three-year span 2009–2010–2011 (see Annex 1).

[19] Italian key-words used to filter the corpus selection are the following nouns and adjectives (in singular and plural forms): *migrazion** ("migration/s"), *immigrazion** ("immigration/s"); *fluss* migrator** ("migratory flow/s"), *emigrazion** ("emigration/s"), *migrant** (migrant/s), *emigrant** (emigrant/s), *immigrat** ("immigrant/s"), *profug*/rifugiat** ("refugee/s"), *richiedent* asilo* ("asylum seeker/s"), *clandestin** ("illegal immigrant/s"), *stranier** ("foreigner/s").

[20] Speeches here analysed are comprised in Section IA within the Corpus MULPOLDIS (Multilingual Multimodal Political Discourse) developed by the Corpus Linguistics Centre at the Università degli Studi internazionali di Roma (https://www.unint.eu/it/ricerca/centri-di-ricerca/centro-di-ricerca-linguistica-su-corpora-clc/1357-corpus-mulpoldis)

[21] "Fields of action" may be understood as segments of the respective societal 'reality', which contribute to constituting and shaping the 'frame' of discourse. The spatio-metaphorical distinction among different fields of action can be interpreted as a distinction among different functions or socially institutionalised aims of discursive practice" (Reisigl & Wodak 2001: 36).

Britain, Holland, Italy and Spain) to refer to migrants and to the phenomenon of immigration.[22] Reisigl & Wodak (2001) pinpointed discursive strategies used to define social actors and predicate on them by conveying an overt or implicit evaluation on speaker's attitude toward a given social category or phenomenon. As a matter of fact, findings in the analysis of migration discourse "allow[s] to conclude that much discourse about migrants and immigration seems to bear several almost universal features, throughout Europe and beyond, which can be explained by social theories about 'Othering' and the discursive construction of 'the stranger' and 'fear of the stranger' [...]" (Wodak 2015: 8).

In order to interpret our data, we decided to apply the analytical categories of the *Discourse hystorical approach* (DHA, in **ReisiglWodak2009**) aimed at analysing discursive strategies for the Other-representation and the discursive construction of migration in concrete text extracts in Italian mediated into Spanish (examples in §4). In particular, we are referring to five heuristic questions considered salient to DHA in Wodak's categorisation (Wodak 2015: 8):

- How are persons, objects, phenomena/events, processes, and actions named and referred to linguistically?
- What characteristics, qualities and features are attributed to social actors, objects, phenomena/events and processes?
- What arguments are employed in the discourse in question?
- From what perspective are these nominations, attributions and arguments expressed?
- Are the respective utterances articulated overtly, are they intensified or mitigated?

Therefore, our analysis on original speeches was focused on: a) referential strategies, b) predicational strategies, c) argumentative strategies.

The referential strategies are used either to include, suppress, specify, generic-ise, spersonalise or deny the Other date back to van Leeuwen's (1995) categorisation. To realise these strategies various forms of labelling are used to name social actors and characterise them with respect to inclusion/exclusion in social events and in terms of the way they may be personally or impersonally represented and classified specifically or generically (see Fairclough (2003: 145–146)).

These strategies are considered together with predicational ones:

[22] See also Rojo & van Dijk (1997) and Beaton-Thome (2013).

> Predication is the very basic process and result of linguistically assigning qualities to persons, animals, objects, events, actions and social phenomena. Through predication, persons, things, events and practices are specified and characterised with respect to quality, quantity, space, time and so on. Predications are linguistically more or less evaluative (deprecatory or appreciative), explicit or implicit and – like reference and argumentation – specific or vague/evasive. (Reisigl & Wodak 2001: 54)

Studies conducted on predicational strategies (such as **Wodak2000**, **VanDijk2002**, Wodak 2001, Reisigl & Wodak 2001) have mainly considered speeches concerning antisemitism, racism, nationalism or discrimination based on gender, race, religion where the Us/Them opposition emerged clearly in the pragmalinguistic representation of the Outgroup as opposed to the Ingroup.[23] Predications are developed by means of *topoi* (such as that of numbers[24]) or metaphors and "extended metaphors"[25] that support the argumentative strategies through which the migration discourse is based on.

In order to build argumentations and counter-argumentations referring to given social groups (Reisigl & Wodak 2001: 45), in political discourse the use of *topoi* is particularly exploited not only to discuss on a given topic but as a productive strategy to represent the commonsense reasoning typical for specific issues. Among other most frequently adopted argumentative strategies, we can cite metaphors used "when it is necessary to simplify complex issues, and to present them in vivid and potential emotional terms" (Semino 2008: 124) and the reporting personal experiences related to the speaker's private Self to direct public opinion by increasing the degree of legitimation of what is being said and, consequently, his/her reliability also by shedding a negative light on opposing past actions and on the political group through emphasis on the transformation between then and now.

The above-mentioned criteria were applied to lead our research by combining quantitative corpus-based analyses and qualitative ones as follows:

[23] In this regard, we can cite van Dijk's (1998) theoretical concept of "ideological square" through which he encapsulates the polarisation manifested in discourse by lexical choice and other linguistic features as far as the representation of Self and Others, Us and Them are concerned.

[24] See also the study on an anti-immigration leaflet conducted by Semino (2008: 118–124), where the use of numbers is consistent with the negative representation of migrants that emerges from the text.

[25] According to Semino (2008: 25) an extended metaphor - such as football metaphor widely used in Italian political speeches - is considered "as a particular type of cluster, where several metaphorical expressions belonging to the same semantic field or evoking the same source domain are used in proximity to one another in relation to the same element, or to elements of the same target domain".

- corpus-based analysis to identify referential strategies by looking for topic-related words (in terms of frequency) and qualitative analysis of selected examples and their renderings into Spanish;
- corpus-based analysis to highlight predicational strategies through concordancing and their comparison with mediated strategies adopted into Spanish;
- qualitative analysis of relevant *topoi* and metaphors for the construction of migrants and the representation of migration through politicians' argumentations.

This methodological approach allowed us to identify respectively referential, predicational and argumentative strategies adopted to express politicians' ideologies on "ethnic topics" through discursive practices mainly based on the polarisation between the Ingroup and the Outgroup.

These results were, then, assumed as a starting point to analyse contrastively the solutions used in mediated texts in order to focus the way, and to what extent, Italian politicians' discursive strategies interpreted into Spanish are conceived to convey pragmatic equivalence.

4 Discussion

The categorisation of discursive strategies in accordance with the DHA was applied to original EP speeches in Italian in order to evaluate to what extent they are re-codified during the mediation process into Spanish.[26] In the following subsections, the mediation strategies adopted to comply with the fulfilment of pragmatic equivalence in terms of referential strategies (§4.1), predicational strategies (§4.2), argumentation strategies (§4.3) are discussed.

4.1 Mediation of referential strategies

Strategies used to name the social actors involved were examined as far as specification or generalisation and reference to age, race, gender, origin, and so on are concerned. In Figure 1 it is possible to observe the distribution of referential strategies used in original discourses in Italian by referring to migrants as social actors (Figure 1) or to the social phenomenon of migration (Figure 2) and their interpretations in Spanish through: a) word-for-word translation (in blue), b) synonyms or re-elaborations of the originals (in grey), c) omissions (in orange).

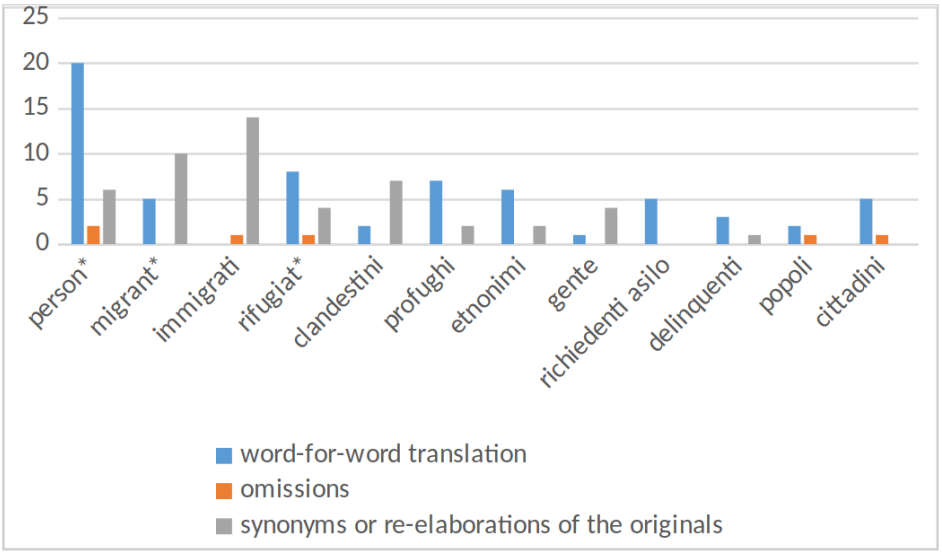


Figure 1: The mediation of referential strategies (social actors)

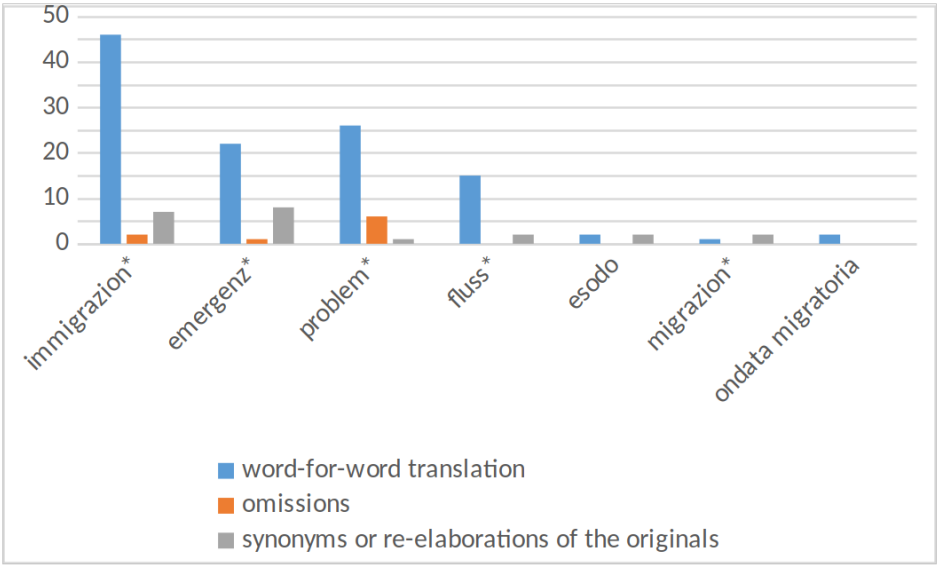


Figure 2: The mediation of referential strategies (phenomenon of migration).

Figure 1 and 2 show that word-for-word translation (a) is the most common strategy interpreters prefer to use, both to represent social actors and the phenomenon in itself. More specifically, this strategy allows the interpreter to convey both semantic and pragmatic equivalence, for example by maintaining the reference to the legal status of social actors (by translating *profughi*/*rifugiati* into *refugiados*/ *prófugos* and *richiedenti asilo* into *solicitantes de asilo*), the inclusion and the rights of migrants (by rendering *persone, cittadini* and *popoli* into *personas, ciudadanos* and *pueblos*), and the reference to criminality and delinquency by using criminonyms (*delinquenti* translated into *criminales*). Table 1 reports the percentage of use for the word-for word strategy for the following items (*refugees*, *asylum seekers*, *person*, *citizens*, *people*, *criminals*) and information on the communicative events (see Annex 1 and 2 for more details).

Table 1: Frequency of the word-for-word strategy

Items (number of occurrences)	Percentage of word-for-word strategy	Speakers[27]	Topics
Ciudadanos (6)	83%	SA (3), MP (2)	7, 8
Criminales (4)	75%	MBO (3)	8, 10
Personas (28)	71%	SA (3), RA (2), CM (2), DS, SI, AP (2), MBO, RC (3), SC (2), FP (3)	1, 3, 5, 7, 8
Prófugos (9)	78%	CF, SI (2), MBO, MB (2), SA	7, 8, 12
Pueblos (3)	67%	RC, MM	7, 8
Refugiados (13)	64%	SA, CC, CF, SI (4), DS	1, 3, 5, 7, 8
Solicitantes de asilo (5)	100%	RB, MM, SS, RA (2)	1, 8

Numbers into brackets indicate the occurrences per speakers for each item, in case it is more than one.

When referring to the phenomenon of migration, interpreters tend to adopt a word-for-word translation in order to maintain the same linguistic nuances of Italian items. They mainly use lexemes conveying the difficulty to manage the phenomenon by translating *emergenza* and *problema*[28] into *emergencia* and *prob-*

[26]See Anghelli (2019) for detailed results carried out on the original discourses.

[28]For a study on the use of "emergency" and "problem" on the Italian press, see Orrù (2017).

lema. They also refer to immigration as water-course or flood[29] by translating *flusso* and *ondata migratoria* into *flujo* and *oleada/ola migratoria/ de inmigración* as well as they make use of stereotyped metaphor of migration as an exodus: *esodo* > *éxodo*). Sometimes interpreters opt for more neutral solutions such as *inmigración* and *migración* when in original texts analogous referential strategies were adopted (*immigrazione* and *migrazione*).

In Table 2, the percentage of the word-for-word translation for the following items (*emergency*, *problem*, *flow*, *wave of migration*, *exodus*, *migration*, and *immigration*) is reported.

Table 2: Frequency of the word-for-word strategy

Items (number of occurrences)	**Percentage of word-for-word strategy**	**Speakers[30]**	**Topics**
Emergencia (32)	71%	SS (2), FP (3), MM (3), MB, PP, CF (2) SA, MBO (2), AP (3), CC, BM (2), RBA	6, 7, 8, 9
Éxodo (4)	50%	SS, GP	6, 12
Flujo (17)	88%	RB, SI (6), GLV, RBA, MM (2), GP, FP, RA, CF	1, 2, 3, 4, 6, 7, 8, 9, 10, 12
Migración (3)	33%	SI	8
Immigración (55)	84%	RA (9), CM (3), DS (5), RB (2), AC, MB (4), MM (6), PP, FP (3), SA (3), SI (3), RBA, CC, MBO (3), GP	1, 2, 3, 4, 7, 8,10, 11, 12
Oleada migratoria (2)	100%	PP, MB	7, 8
Problema (33)	79%	RA, AC (2), AP, MM (15), SI, PP (2), RC (2), FP, SA	1, 2, 5, 7, 8, 10

[30] Numbers into brackets indicate the occurrences per speakers for each item, in case it is more than one.

The second strategy (b) is generally adopted by interpreters when a word-for-word-translation is not possible to refer to social actors or to the whole phe-

[29] For a study on this topic, see Reisigl & Wodak (2001).

nomenon. In these cases, synonyms, other grammatical categories or syntactical variations of the Italian forms are selected. With respect to social actors, the most frequent example concerns the re-elaboration of the Italian item *clandestino*, rendered in Spanish[31] through the adjective category (*clandestino*). For this reason, interpreters opted for the syntactic structure [N + Adj], such as *emigrantes clandestinos*, embarcaciones clandestinas and *flujos clandestinos* (in MB8 exclusively).[32]

There are also cases where interpreters make some grammatical variations, namely the number category by using plural nouns to stress the generalisation strategy: three times the word *rifugiato* was translated into *refugiados* (SI3, SI5, AP5).[33]

Some syntactic changes that do not affect the meaning and the speakers' intent may be observed in the coupling of an adjective with a more neutral noun (1) or, oppositely, in the omission of the adjective when interpreters use a less neutral noun (2), as in the examples below:

(1) *Somali* (SA1) and *Eritrei* (DS1) were translated respectively into *habitantes de Somalia* and *habitantes de Eritrea.*

(2) *persone detenute* was translated into *reclusos* (SI2) and *persone tutelate* into *protegidos* (AP5).

When referring to migration as a phenomenon, we found an extended interchange of synonyms to refer to migration as a "problem" in order to maintain the original communicative intent and to present it as an emergency: from Italian *emergenza* to Spanish urgencia/*urgencia humanitaria* (MM7), *reto* (SS7), *situación urgente* (FP7) and *problema* (MB7).

In some cases interpreters preferred referring directly to social actors (3) or using a water metaphor (4) to refer to the phenomenon,[34] rather than a word-to-word translation:

[31]In two discourses (MB8, MBO8), the Italian noun *clandestino* has been rendered into Spanish as a noun, thus resulting in a calque.

[32]In the mediated text MBO11 the NP *inmigrantes ilegales* is adopted. It has to be noted that this is the translation proposed for any occurrence of the Italian noun clandestino in the official written translations of these speeches. (https://www.europarl.europa.eu/plenary/es/debates-video.html#sidesForm)

[33]We cannot avoid to remark that in one case rifugiati (SA7) was translated as delincuentes with a severe semantically inadequacy by using a criminonym rather than the reference to a specific legal status.

[34]For an extensive list of contribution on water metaphors, see Taylor (2020: 12).

(3) migrazione > inmigrantes (MBO8)

(4) immigrazione" > oleada migratoria (GLV8)
migrazione > flujo (MBO8)

From our research perspective, both renderings affect the pragmatic original meaning by using a personification (3) or the water metaphor in place of the neutral solution adopted in the original (4). In this last regard, Taylor (2020: 3) explains that "metaphors by their very nature are not neutral" and, referring to the water metaphor in particular, **Marlow2015** suggests that it can be used to enhance the perception of immigrants as a source of threat.

The fact that interpreters seem to be focused on the semantic content without paying enough attention or being enough aware of the pragmatic effect is even more evident in the following examples where the original referential strategies are rendered through semantic intensification (5), imperfect rendering (6) and gender-specification (7):

(5) *problema* is translated into *tragedia* (MB012), thus intensifying the original meaning by adding a semantic component.

(6) *persone* and *migranti* are translated as *asilados* (AP5) and *solicitantes de asilo* (SS6) both these choices refer to the legal status that was not at all taken into consideration in the original version.

(7) *uomini* used to refer to human beings is rendered into mediated texts by means of *hombres y mujeres* (FP8a), thus making it explicit the reference to both genders.

The third strategy investigated (c) concerns the use of omissions, as referred to by **Wadensjoe1998**, in order to detect cases of "zero rendition" and "reduced rendition". As a matter of fact, a remarkably imperfect pragmatic correspondence in mediated discourse into Spanish emerged in instances of omissions, affecting the original pragmatic content and, consequently, the speaker's pragmatic (and political) intent. This failure in recoding speaker's intentionality could depend on the speed of speeches delivered by Italian speakers – considering that they are generally read aloud and pre-planned – and on the high information density due to the shortness of this textual sub-genre. Moreover, the difficulty in giving a complete pragmatic correspondence is caused by the nature of the context itself, where interpreters have to deal with a quick succession of speakers (see §2).

The occurrences in our corpus, exemplified in (8) and (9) show that omissions usually have an impact on the rendition of the interpreter, especially when the

pragmatic intent of the speaker is not transmitted. The following example reported in (8)[35] clearly shows the mismatch between the original and the mediated version: in the Italian statement *la nostra gente* underlines the sense of ingroupness as opposed to the Outgroup (*gli immigrati*). The "reduced rendition"[36] in the mediated version blurs this communicative purpose:

(8) a. MB2: Tutto il resto è retorica buonista che non aiuta né l'integrazione degli immigrati né tantomeno la nostra gente.
b. Negarlo no sirve para nada.

We may also find examples of original relevant expressions left untranslated, identified as cases of zero renditions, where the interpreter avoids to convey the idea of migration as a problem (9) and, at the same time, he/she refer to efforts accomplished to face it rather than mentioning the necessity for future endeavors in accordance with the original speech:

(9) a. DS7: Noi sappiamo che **il problema** è italiano, ma è anche europeo. Occorre un considerevole sforzo finanziario perché avvenga questo e avvenga questo in un quadro di politiche coordinate, (…)
b. **(Ø)** Se han desplegado grandes esfuerzos también por parte de la Unión Europea para que nos podamos dotar de un marco de políticas coordinadas (…)

4.2 Mediation of predicational strategies

One of the selected predicational strategies used by Italian speakers to codify their political message concerns the addition of some modifiers to nouns referring to the phenomenon itself by using either NP collocates [N+Adj] or compounds such as:

- *immigrazione clandestina* (RA1, DS1, RB1, MB2, RA3, RA4, MM7, DS7, SA8);
- immigrazione illegale (SA1, MB11, MB11);
- immigrazione irregolare (RB1).
- problema dell'immigrazione (MM10, RA1, SA8);

[35] The Italian speech from which it was extracted comprises 150 tokens uttered in just over a minute.

[36] A "reduced rendition" includes less explicitly expressed information than the preceding "original" utterance (**Wadensjö1998**)

- emergenza immigrazione (FP12, DS7).

In mediated discourses in Spanish, interpreters manage to express the original global meaning through word-for-word translation: the collocate *immigrazione clandestina* is translated into *inmigración clandestina*, opting for formal adherence.[37]

In one case, a lack of pragmatic equivalence is reported: *problema dell'immigrazione* is translated into *tema de la inmigración* (MM10), with the consequent loss of evaluation concerning immigration as a problem.

In Table 3 other Noun Phrases are used to predicate on the topic:

Table 3: Noun Phrases describing the social phenomenon.

Speakers	NPs (original version)	NPs (interpreted versions)
AC1	non siamo in una situazione normale	no están en una situación normal
CF7	a. questi rivolgimenti di dimensione potenzialmente epocale	a. (este) acontecimiento de dimensión enorme
	b. flusso straordinario di immigrati	b. flujo (Ø) de inmigrantes
MB7	uno stravolgimento storico	un cambio histórico
MB8	un'emergenza senza precedenti	una emergencia sin precedentes
MBO8	a. questa emergenza epocale	a.esta situación de emergencia (Ø)
	b. emergenza anche umanitari) di carattere straordinario	b. situación de emergencia humanitaria (Ø) también
MM4	fenomeno epocale	fenómeno (Ø)
MM7	sommovimento epocale	movimiento que marca época
SA8	eventi straordinari	hechos extraordinarios
S18	a. pressione migratoria eccezionale	a. presión migratoria excepcional
	b.una situazione eccezionale	b. esa situación excepcional

As we can see, perfectly equivalent examples of word-for-word translation are attested (SI8 a–b, SA8, AC1, MB8). As for re-elaboration (MM7, CF7 a, MB7),

[37]This occurs in seven times out of nine (RA1, DS1, DS1, RB1, MB1, RA3, DS2), despite the fact that *inmigración ilegal* is the solution adopted in the official written translations where this is used exclusively.

these cases affect the nouns within NPs more than modifiers, and the pragmatic result is a mitigation of the original meaning and, consequently, of the speaker's attitude toward the phenomenon. In case of omissions (MBO8 a–b, MM4, CF7 b), they concern modifiers and they can be categorised as "skipping omissions", that is "the omission of a single lexical item such as a qualifier or a short phrase which appears to be skipped over by the T and which is of minor consequence" (**Barik1975**). Certainly, omissions cannot be considered a relevant mistake with regard to the general meaning. However, from a pragmatic point of view, they can provoke a loss of the emphasis as far as the dimension of the phenomenon is concerned. This may also be observed in (10) where the interpreter omits the final portion of the sentence, thus affecting the semantic content and the speaker's intent: the immigration not being a resource during troublesome times.

(10) a. MB2: Una strategia davvero utile che deve basarsi su alcuni punti fermi: lotta all'immigrazione clandestina lungo la frontiera sud, condivisione tra tutti gli Stati europei degli oneri a contrasto della clandestinità, politica di accordi con paesi terzi, soprattutto il riconoscimento **che l'immigrazione nel momento della crisi non è una risorsa**.

b. Hay una serie de puntos que tendrían que ser determinantes: limitar la inmigración clandestina en el Sur, llegar a acuerdos con países terceros en materia de inmigración **(Ø)**.

Another common predicational strategy in migration discourse concerns the use of definite or approximate numbers to quantify the extent of the phenomenon especially when referring to migrants' arrivals: "the very first attribute applied to immigrants coming to the country is in terms of their numbers" (**VanDijk2002**). The relevance of the phenomenon is expressed either through the use of the *topos* of numbers[38] (11) or by means of general quantifications[39] (12):

(11) a. RC8: Fra questi c'erano **4.500 minori**, ragazzini di 12–13 anni, che condividevano quella condizione, una condizione di disagio che condividevano anche gli abitanti di Lampedusa

b. En fin, situaciones desastrosas. **500 menores**, niños de doce, trece años que vivían en estas condiciones. (Ø)

[38]Occurrences of the *topos* of numbers are reported in the following speeches: SI5, MB7, MB7, SI7, CF7, MBO10, MBO10, MBO10, MBO10, DS10, MB012, MM8, RC7, RC8, RC8, FP8 b, FP8 b, FP8 b, SA1, RA1, RC8, NR4, RB1, and MB8.

[39]General quantifications characterise the following speeches: CF7, MB8, DS1, MB7, SA1.

From the interpreter's side, numbers can be very difficult elements to translate, as example 11 shows.[40] In this regard, Pearl (1999: 19–20) states:

> The trouble with asemantic elements is that they are not part of a semantically linked chain, but just so many unconnected or loose links which cannot be inferred or anticipated from the speech flow. […] When the interpreters find the semantic flow interrupted by figures and are forced to abandon their lag, their attempts to grapple with them, as often as not unsuccessful, to take up a disproportionate amount of their time and attention, with the net result that not only are the figures themselves garbled, mangled or omitted, but the surrounding semantic material also suffers damage or omission in the confusion.

In our collection of texts, the number of arrivals was represented through figures or more general quantifications (such as "dozens of" and "many") and it has been emphasised through the recourse of temporal deixis in order to highlight the short time span in which disembarkations occurred (12) and to the repetition of the number to enhance its scope (13):[41]

(12) a. NR4: La sua minuziosa contabilità è arrivata a 4.200 vittime, **18** delle quali **lo scorso marzo**: una vera ecatombe

b. Parece ser que hay 4.200 víctimas en, sobre todo **en Marzo de este año** ha habido **muchísimas víctimas**.

(13) a. RC7: Invece facciamo diventare un'emergenza umanitaria il fatto che una delle più grandi potenze mondiali, l'Italia, abbia a fronteggiare il problema di come accogliere **5.000, 5.000 persone**.

b. En cambio, lo que hacemos es convertir en una urgencia humanitaria el hecho de que una de las mayores potencias mundiales, que es Italia, tiene que hacer frente a como acoger a (Ø) **5.000 personas**.

The above-mentioned examples clearly show that interpreters do not always succeed in translating numbers. Sometimes they may opt for a substitution with approximative quantifications,[42] as in the second part of the sentence in (12),

[40] It is the only case in the corpus where the interpreter reproduces a wrong number rather than using a more general quantification.

[41] It must be noted that in the Spanish version the epistemic modality is mitigated: the interpreter uses *parece ser* ("it seems to be") rather than *è* ("is").

[42] Approximative quantifications instead of figures can be found in the following mediated-texts: MB8 and NR4.

where number 18 is replaced by the general quantification *muchísimas*. In this way, the rendition lacks precision, and it affects the pragmatic effect since the number refers to individuals and the speaker had expressly adopted a personification strategy.

Moreover, in example (13), the interpreter omits the repetition of the number:[43] in this case, the omission does not alter the semantic content, but it affects the communicative equivalence since the speaker's intent was to emphasise the phenomenon size through the repetition.

Another predicational strategy is realised through predicates and it is implicitly connected with the *topos* of threat in order to highlight the importance of reducing – or even stopping - the arrivals or the need to take measures to face the emergency. In all cases, interpreters manage to transmit the original meaning, through a word-for-word translation or by using synonymic expressions, as we can see from the following examples:

(14) a. far fronte a /affrontare/ fronteggiare > hacer frente[44]
b. bloccare > bloquear (RC7)
c. frenare > frenar (PP7)
d. porre un limite > limitar (RBA8)
e. governare > governar (GP12)
f. prevenire > prevenir (GLV8)
g. contenere > contener (RA1)
h. fermare > poner frenos (PP7)

Related to the *topos* of threat some verbs, such as *invadere* (to invade) and the metaphoric *prendere di mira* (to target) can be considered since they aim to emphasise negative evaluation of these arrivals:

(15) a. MB8: L'Italia **è** da settimane **presa di mira** da centinaia di barconi di clandestini. Lampedusa **è stata invasa** da decine di migliaia di nordafricani che l'isola, mai e poi mai, potrebbe accogliere.
b. Lampedusa **ha sido invadida** por centenas de miles de norteafricanos que la isla no puede acoger. Italia, desde hace semanas, **tiene a** muchas embarcaciones clandestinas.

[43] Omissions can be found in mediations of RC7, MBO10 and MB7.

[44] In MP8, MM12, FP12, FP7, MM7, PP7, FP8 a, FP8 a, SA8, CC8, FP12, MB7, PP7, PP7, RA7, SS7, MB8, MBO8, BM8, MBO10, BM8, DS2, RC7.

As we can see, the interpreter did not manage to transmit the same intent by translating *invadere* into *tener a*, which is a sort of mitigation.

In (16) a reduced rendering of the referential expression is adopted since the Outgroup is identified with migrants (rather than illegal migrants) and the different morphosyntactic structure (from passive to active) blurs the role of the Italian government in the mediated-text.

(16) a. MB8: I clandestini **devono essere spediti** a casa loro.
b. Los inmigrantes **tienen que volver** a sus países.

Finally, from a semantic point of view, there are verbs expressing a prolonged action over time, such as *continuare* (or *permanere*),[45] used to declare the daily problem local people have to face and the key role Italy plays by giving shelter to them. Examples in (17) show the way interpreters succeed in maintaining a perfect equivalence by selecting predicational strategies that show speakers' defending and siding with the Outgroup:

(17) a. MS8: (…) **continuano** gli sbarchi, la gente muore in mare e si affolla in condizioni disumane sulle coste italiane e maltesi (…)
b. (…) **siguen produciendose** desembarcos, la gente muere en las costas italianas y maltesas (…)
c. RBA8: L'Italia **continua** a fare la sua parte nell'accoglienza a questi disperati.
d. Italia **sigue** desempeñando su papel en la acogida que se da a estos desesperados.

4.3 Mediation of argumentative strategies

The preliminary pragmalinguistic analysis carried out on original interventions allowed us to identify the main argumentation strategies adopted by Italian politicians with the objective of focusing on the way interpreters manage to take them into consideration in their mediated texts.

Usually, speakers exploit argumentative strategies to create a positive representation of themselves (and therefore to discredit other political ideologies) by either showing empathy towards migrants or representing them in a negative way, explaining the reason why migration has to be contrasted.

This may be observed when referring to personal experiences and to concrete examples aimed at giving a sort of empirical proof and more accountability to what is being said:

[45] In MB8, MB8, RBA8, RBA8, MS8, SI10, SA1, RBA8, SA1, RBA8.

(18) a. S17: **(…) Io sono siciliano** e sono stato, **contrariamente a molti altri che hanno qui parlato**, nel centro di identificazione ed espulsione. Quel centro era un centro che è servito in quel momento ma che, dopo l'accordo sul trattato di amicizia, cessava di avere alcuna validità.

b. **Yo soy Siciliano, (Ø)** por cierto, existía un centro de acogida que se utilizaba a la sazón, pero tras el acuerdo de amistad entre Libia e Italia ya no servía de nada ese centro de acogida.

In the example (18), the speaker aimed to stigmatise other politicians' behaviour by referring to his own experience, being of Sicilian origin. It is to be noted that, in this speech, the interpreter does not create an equivalent version, due to the omission of this important contrast.

Among other argumentative schemes used by speakers in their political speeches: (a) the pro-migration argumentation; (b) the anti-migration argumentation.

The former (a) highlights a positive attitude towards migrants who are not referred to through a binary opposition (Them *vs* Us; Outgroup *vs* Ingroup), while the latter (b) enhances the ingroupness to the local community, thus strengthening ideological polarisation (§3).

The first category comprises the *topoi* of history, humanitarianism and "Italy as a country of migration", the latter consists of *topoi* of fear, disadvantage and the *topos* of burden sharing.[46]

Being standard arguments shared by a wide range of speeches belonging to the migration domain, interpreters managed to transmit the pragmatic meaning in all occurrences. For instance, the following example shows the rendering of the *topos* of "history as a teacher" into Spanish:

(19) a. DS1: Vogliamo sapere, signor Presidente, se la Commissione intende intervenire sulla legislazione italiana, verificare l'accordo italo-libico. (…). **Non possiamo consentire vent'anni dopo la caduta del Muro di Berlino ad alcuni governi di alzarne di nuovi.**

b. Queremos saber, señor Presidente, si la Comisión intervendrá en la legislación italiana, comprobando ese acuerdo italo-libio. **No podemos permitir que, después de la caída de Berlín, veinte años después, algunos Gobiernos vuelvan a levantar otros muros**.

Similarly, in (20) we see that the interpreter succeeds in translating the overall meaning of the original text topos of burden sharing:

[46]See also example (9) where the speaker refers to an Italian problem as being a European one as well.

(20) a. DS8: (…) **Domani, in Parlamento faremo la nostra parte. Però occorre che anche gli altri facciano la loro**, che i governi siano molto meno egoisti, che la solidarietà che serve per attivare una politica europea, beh, ci veda promotori. In questo, il suo lavoro naturalmente è al centro di questo sforzo, perché senza i governi l'Europa sarà più debole.

b. **Mañana en el Parlamento, vamos a hacer lo que nos toca hacer pero los demás también tienen que hacerlo**. Los gobiernos tienen que ser mucho menos egoístas, la solidariedad necesaria para aplicar una política europea. Pues, tenemos que ser nosotros los promotores de ella. Y su trabajo está en el centro de esos esfuerzos, porque, sin los gobiernos, Europa será más débil.

In the following extract (21) the interpreter successfully conveyed the global meaning though she/he failed to express the "involvement strategy"[47] by omitting the adverb *purtroppo* (unfortunately) which plays the function of pragmatic clue:

(21) a. SA1: **Purtroppo**, il Mediterraneo è ormai diventato un cimitero a cielo aperto e il governo Berlusconi, quindi il governo italiano, ha adottato un accordo con la Libia che consente **purtroppo** il respingimento non solo di migranti, ma consente anche il respingimento dei rifugiati che provengono da paesi dove sono in atto persecuzioni, guerre civili, come la Somalia e l'Eritrea, e nega a questi poveri disgraziati il diritto di chiedere asilo, violando così non solo tutte le norme internazionali, ma viola soprattutto la Convenzione di Ginevra.

b. **(Ø)** El Mar Mediterraneo se ha convertido en un cementerio a cielo abierto y el Gobierno Berlusconi, el Gobierno de Italia ha aprobado un acuerdo con Libia que **(Ø)** permite dar no acogida a los inmigrantes y rifugiados que vienen de países donde hay guerra civil como Somalia, Eritrea y niega a esos pobres desgraciados el derecho de asilo, violando así no solo todas las normas internacionales, sino que procede a una violación del Convenio de Ginebra.

In (21), the interpreter did not translate the adverb *purtroppo,* probably not considered to be relevant, though the Italian speaker used it twice to stress her own

[47] "Strategies of involvement (see **Tannen1989**) aim both at expressing the speakers' inner states, attitudes and feelings or degrees of emotional interest and engagement and at emotionally and cognitively engaging the hearers in the discourse" (Reisigl & Wodak 2001: 81).

opinion and attitude toward the topic. Differently the reference to an authoritative source (*Geneva Convention*) to legitimate a statement and speaker's personal opinion was rendered in the Spanish mediated-text.

In the following example (22) it is worth to note the use of metaphors to build political argumentation: metaphors that aim at emphasising and clarifying the propositional content are often used.[48] In this case the reference to football metaphor as the most popular sport[49] in Italy, gives rise to

> a new type of political discourse in the Italian context, one that abandons the traditional obscurity and replaces it with vivid and relatively simple references to domains that are likely to be accessible to a wide audience. (Semino & Masci 1996: 266)

(22) a. MM9: C'è **una partita da giocare**, quindi. Ciò che mi colpisce è che mi sembra che in alcune circostanze **i giocatori della partita** degli ideali rinunciano a **giocare la partita**. (…) Ora, come fa la squadra degli ideali a **vincere la partita** se **i nostri giocatori** rinunciano a tirare in porta magari perché pensano che **il portiere** è troppo bravo? (…) E allora, mi permetto fare questa osservazione: chi sono **i giocatori dell'attacco**? Sono le Istituzioni europee: Parlamento, Commissione e anche lei, signor Presidente Van Rompuy. (…). Vi chiedo allora semplicemente: **siete i giocatori del nostro attacco**, **passatevi la palla**, **giocate all'attacco**, **fate goal** e come si dice in questo tipo di partite, fateci sognare.

b. Hay, por tanto, **una partida**. Lo que me surprende es que, en algunas circumstancias, parece que **los jugadores del partido** de los ideales renuncian a **jugar el partido**. (…) Entonces, ¿Cómo puede el equipo de los ideales **ganar el partido** si **sus jugadores** se niegan a **tirar a gol** porque quizá piensan que **el portero** es demasiado bueno? (…) Por ello me permito hacer este comentario. ¿Quiénes son **los jugadores al ataque**? Las instituciones europeas, el Parlamento, la Comisión y también usted, señor Presidente Van Rompuy. (…) Solamente le pregunto: ¿**Son jugadores de nuestro ataque**? **Pues pásense la pelota**, **juguen al ataque**, **ganen el partido**, hágannos soñar, **marquen un gol**.

[48] See Taylor (2020). For a list of the functions of metaphors in discourse, see Semino (2008: 30–32).

[49] In this sense, according to Semino (2008: 98): "since cultures and countries differ in their sporting preferences, different sport metaphors tend to dominate in different languages and countries."

As for the interpreted version, in this case the interpreter uses the same football metaphor in Spanish[50] and, since football is a popular sport in Spain, the pragmatic effect is totally maintained.

In (23) an example the equivalent pragmatic effect of the original metaphor used in Italian is obtained through the use of figurative language:

(23) a. SA8: Pur in presenza dell'articolo 80 del trattato sul funzionamento dell'UE e del principio dell'equa ripartizione della solidarietà, ogni paese di fatto **tira acqua al proprio mulino** e l'atteggiamento della Francia al confine con l'Italia è inammissibile nell'attuale quadro europeo.

b. Aunque tenemos el artículo 80 del tratado de funcionamento de la Unión Europea y de un reparto equitativo de la solidariedad, cada país **intenta barrer hacia sus puertas** y, de hecho, la postura de Francia en las fronteras con Italia es inadmisible.

In fact, in the interpreted version of example (23), the fixed multi-word expression *barrer para casa* ("to look after number one") is used with an equivalent pragmatic meaning though the interpreter increases the creativity (see Semino 2008: 21) by substituting *casa* ("house") with *puertas* ("doors").

The metaphor used in (24) is rendered through a paraphrase that preserves the semantic equivalence, though not the pragmatic meaning. In fact, the Italian metaphor referring to the gruyere cheese (*groviera*) and its holes (*buchi*) is used to give the idea of vulnerability of Romania and Bulgaria. In this case, the same idea is rendered without using a metaphor by the interpreter, who simply refers to these countries as places used to enter in the European Union, thus causing a loss in terms of political rhetoric. Besides this, it has to be remarked upon that the metaphor used to refer to Bulgaria and Romania in the original speech was extended to the whole European Union in the mediated text, thus producing a distorted meaning.

(24) a. MBO11: Suggerisco la necessità di una sospensione della procedura di entrata di Bulgaria e Romania nel sistema di Schengen sulla base del principio di precauzione, anche in vista della prevedibile enorme pressione dalle frontiere esterne verso questi due paesi che **diventano la groviera, i buchi di groviera del sistema** dell'Unione europea rispetto all'entrata dei clandestini.

[50] See Turrini (2004) for an explanation of equivalent metaphors, as well as the use of different images or paraphrase in the interpreted versions of texts.

b. Por eso **pido** la suspensión del procedimiento de adhesión de Bulgaria y Rumania en Schengen, basándose en el principio de precaución y también a la vista de la presión enorme previsible que provendrá de estas fronteras exteriores hacia la Unión Europea que se **van a convertir en los lugares por donde pasen** todas esas personas que son victimas de la trata de seres humanos.

From our perspective what is even more remarkable in (24) is the avoidance of an equivalent for the Italian verb *suggerire* (*suggerisco*) which is not used to build the speaker's argumentation in Spanish. This provokes an imperfect equivalence in terms of "degrees of strength" because this verb has a lower illocutionary force compared with the Spanish verb *pedir (pido)*. In this way, the mitigation intended to be used by the Italian speaker is not rendered into the Spanish version.

5 Conclusive remarks

By adopting the analytical categories of the DHA developed within the framework of Critical Discourse Analysis our study focused on the mediation of discursive strategies related to migration/migrants and of ethnopragmatic devices that embody self-representation and the (negative/positive) construction of the Other.

Analysis of migration discourse was conducted on a corpus of EP original speeches in Italian by MEPs during plenary sessions in order to pinpoint referential expressions used to designate social actors and the social phenomenon of international migration in itself. Predicational strategies used to discuss the selected referents and the way speakers organise their entire argumentation flow, in accordance with their political pro- or anti- immigration stance, were also examined. In fact, within a political environment, speakers might manifest their ideology and attitudes through their pragmalinguistic behaviour, which plays a fundamental role in building his/her political Self. Thus, beyond the locutionary aim of any political statement, interpreters of political speeches are asked to render the perlocutive dimension of the political message enacted in the original.

This preliminary analysis on 60 speeches by twenty-five politicians allowed us to uncover the most potentially interesting statements due to their ethnopragmatic value and to detect possible shifts introduced when interpreting into Spanish. Our pragmatically-oriented approach allowed us to pinpoint the main strategies used to convey speakers' political message when discussing the social phenomenon of migration. In particular, it was possible to identify either

inclusive or exclusive discursive strategies (namely referential strategies, predicational strategies and argumentative strategies) aimed at the Other construction through or a positive/negative description of the phenomenon itself and of migrants as social actors.

Our corpus was conceived in order to balance the effect of a potentially relevant variable, such as the speaker's political group in selecting peculiar discursive strategies to reach a pragmatic intent converging toward two ideology-induced poles: neutralising attitude vs. stigmatising attitude. In order to fully understand pragmalinguistic cues it was necessary take into consideration the national political party of MEP's and its ideological orientation and political rhetoric on the Italian socio-cultural background.

Our data on the original speeches show that there seems to be a distinction as far as referential strategies are concerned. Namely, criminonyms ("delinquents" or "illegal migrants"), rather than neutral referential expressions, are used in the 93% of cases by EFD members belonging to the Italian "Lega Nord" party. Reference to the social condition as permanent ("immigrated") is sligthly more attested (64%) within EPP (from "Lega Nord" and "Fratelli d'Italia") while the frequency of the variant "migrant" is perfectly balanced and it doesn't seem to be correlated with the speaker's political orientation.

In order to name the social phenomenon and predicate on it the water metaphor is mainly used (73%) by EPP and EFP exponents (from "Fratelli d'Italia" and "Lega Nord"). Other predicational strategies are used to support the anti-migration argumentation: the topos of threat is overtedly adopoted in examples 15 and 16 by EFD ("Lega Nord"). Differently the pro-migration argumentation develops with the topos of numbers to emphasise the personification in example 12 by ALDE ("Italia dei Valori")[51] and examples 11 and 13 by S& D ("Partito Democratico"), through reference to the Italian welcoming tradition in the second example reported in 17) and to the topos of history in example 19, both by S& D members ("Partito Democatico"), as well as to the topos of humanitarism in example 21 by ALDE ("Italia dei Valori").

By considering the European Parliament-Nation Parliament corresponce there emerges a more definite mapping that mirrors opposing ideologies that drive the politicians' discursive strategies on this topic. This preliminary consideration clarifies the fundamental importance of the pragmatic awareness interpreters' and for any other involved in the process of recodifying political messages in terms of representation of the political Self.

[51]This political affiliation refers to the time span of speeches here considered.

Our main goal was to contrastively analyse the interpreter-mediation into Spanish of the most relevant discursive strategies (namely referential, predicational and argumentative ones) in mediated migration discourse. Data presented in §4 resulted from our qualitative analysis aimed at describing if and how interpreters maintain, or fail to reproduce, the pragmatic equivalence of the source political intent. As a matter of fact, interpretations into Spanish were analysed in accordance with an ethnopragmatic approach by focusing on the way interpreters are handling to transfer not only the politician's message but also his/her attitude toward a specific topic (migration), complying with her/his political ideology, in a given situational context (see Bülow-Møller (2003)).

Our results illustrate different strategies applied by interpreters to achieve their mediation task:

- word-for-word translation, whenever possible, to ensure semantic and pragmatic equivalence;
- synonymy or re-elaboration of the original strategies in order to obtain the same perlocutive effect;
- omissions through zero renditions or reduced renditions.

Attention was also paid to the entire argumentation flow developed in each oral text by focusing on the mediation of common anti-migration or pro-migration *topoi*, the exploitation of aesthetic devices, such as metaphors, and figurative language referring to water and football play, together with the practice of reporting speakers' own experiences, thus introducing their private domain for the construction of their political Self.

In general, although our qualitative analysis put in evidence the high-degree of semantic correspondence between original discourse and interpreter-mediated one, it was possible to detect cases of partial/full loss of pragmatic equivalence during the interpreting process, thus affecting the pragmatic encoding of the speaker's perspective and the rendering of his/her intentionality as far as a specific topic (migration) is concerned.

From our research perspective, the most interesting examples are those where it is clear that interpreters focused on the semantic content without paying enough attention[52] to (or not being enough aware of) the value of speakers' pragmalinguistic choices or opting for mitigation strategies. In both cases, this might affect

[52]It has also to be noted that this failure in recoding speaker's intentionality could depend on the speed of speeches uttered by Italian speakers - considered that they generally read them - and on the high information density due to the shortness of this genre.

the representation of the political Self[53] and, in this specific context, the rendering of his/her attitude toward ethical issues. This emerged especially as far as the mitigation of pragmatic cues is concerned by:

- neutralising the identification strategy (*topos* of numbers in examples (12)–(13));
- omission of the Ingroup-Outgroup dynamics (example (8));
- blurring of the performative agency as for the political Self representation (example (18));
- change in terms of represented agency (passive-active diathesis) and indexical value expressed by the original verb *spedire via qualcuno* (to drive out) in example (16);
- deletion of orientation markers such as purtroppo (unfortunately) which complies with the politician's involvement strategy (example (21));
- lack of aesthetic agency devices (example (24)).

Intensification cues of the original pragmatic purpose because of the "intrusion" of the interpreter's Self are also reported, thus provoking a pragmatic shift:

- metaphorisation by means of water metaphors contributing to the representation of the situation as particularly serious and dangerous[54] (example (4));
- addition of evaluative components (example (5));
- gender-fairness of referential expressions[55] (example (7)).

[53] These results are aligned with some trends observed by Coppola (2019) in her analysis of agency in interpreter-mediated speeches during bilateral (Italian-German) institutional encounters.

[54] The domain of flooding (and natural disasters in general) is typical of anti-immigration, racist or xenophobic discourses (Semino 2008: 88).

[55] The interpreter's choice toward a gender-inclusive solution - through the explicitation of both genders (men and women) - is not self-evident since within the national and supranational institutions the debate is still going on. The issue is multi-faceted especially and it deserves a special attention when mediating oral or written texts. For some considerations on this topic from a multilingual perspective see Cavagnoli & Mori (2019). For considerations on the use of gender-neutral language by the European Parliament refer to European Parliament (2018).

In conclusion our study highlights the relevance of the pragmatic dimension when mediating oral discourse, providing evidence of the way speakers are using their speech not only to deliver a content but (rather) to perform linguistically their political Self and ideologies especially when social sensitive topics - such as migration - are discussed. A caveat must be borne in mind: the dataset we are dealing with is limited and we cannot have access to information regarding the interpreters and institutional constraints on their output.

Cross-pollination of this research area with Interpreting Studies is undoubtedly valuable and, as far as ethnopragmatically-oriented analysis of agency and indexicality are concerned, research directions may be manifold. Furthermore, applicative implications are also envisaged for interpretation training (see Boyd & Monacelli (2010)) where the knowledge of theoretical models in the field of political discourse analysis could enhance students' meta-pragmatic awareness.

Acknowledgments

This paper arises from the joint collaboration between the authors. For scientific purposes sections are authored as follows: Ilaria Anghelli (Sections 2, 4); Laura Mori Sections (1, 3, 5). Authors are grateful to Michael S. Boyd for fruitful discussion on the topic and they are indebted to anonymous reviewers for their insightful observations.

Appendix A Topics

Topic 1: Immigration, the role of Frontex and cooperation among Member States (debate) (15-11-2009)

Topic 2: Stockholm Action Plan (debate) (18-05-2010)

Topic 3: European Refugee Fund for the period 2008 to 2013 (amendment of Decision No 573/2007/EC) (18-05-2010)

Topic 4: Union for the Mediterranean (20-05-2010)

Topic 5: Cost of examining asylum seekers' applications in Member States (19-01-2011)

Topic 6: State of European asylum system, after the recent decision of the European Court of Human Rights (15-02-2011)

Topic 7: Immediate EU measures in support of Italy and other Member States affected by exceptional migratory flows (15-02-2011)

Topic 8: EU response to the migration flows in North Africa and the Southern Mediterranean, in particular, in Lampedusa - Migration flows arising from instability: scope and role of EU foreign policy (04-04-2011)

Topic 9: Conclusions of the European Council meeting (24–25 March 011) (05 -04-2011)

Topic 10: Migration flows and asylum and their impact on Schengen (10-05-2011)

Topic 11: Application of the provisions of the Schengen acquis relating to the Schengen Information System in Bulgaria and Romania (07-06-2011)

Topic 12: Preparations for the European Council meeting (24 June 2011) (continuation of debate (22-06-2011)

Appendix B Speaker and speech context

Code	Name	Political party	Topic
AC	Antonio Cancian	EPP	1
AP	Alfredo Pallone	EPP	8
BM	Barbara Matera	EPP	8
CC	Carlo Casini	EPP	8
CF	Carlo Fidanza	EPP	7
DS	Davide Maria Sassoli	S& D	1, 7, 8
FP	Fiorello Provera	EFD	7, 8 (x2), 12
GLV	Giovanni La Via	EPP	8
GP	Gianni Pittella	S& D	12
MB	Mara Bizzotto	EFD	2, 7, 8, 11
MBO	Mario Borghezio	EFD	7, 8, 11, 12
MM	Mario Mauro	EPP	4, 7, 8, 9, 10, 12
MP	Mario Pirillo	EPP	8
MS	Marco Scurria	EPP	8
NR	Niccolò Rinaldi	ALDE	4
PP	Pier Antonio Panzeri	S& D	7
RA	Roberta Angelilli	EPP	1, 3, 4
RB	Rita Borsellino	S& D	1
RBA	Raffaele Baldassare	EPP	8
RC	Rosario Crocetta	S& D	7, 8
SA	Sonia Alfano	ALDE	1,8
SC	Silvia Costa	S& D	8
SI	Salvatore Iacolino	EPP	2, 3, 5, 7, 8
SS	Sergio Paolo Francesco Silvestris	EPP	6, 7, 8
RB	Rita Borsellino	S& D	1
RBA	Raffaele Baldassare	EPP	8

References

Anderson, R. Bruce W. 2002. Perspectives on the role of interpreter. In Franz Pochhäcker & Miriam Shlesinger (eds.), *The interpreting studies reader*, 209–217. London/New York: Routledge.

Anghelli, Ilaria. 2019. *Il tema della migrazione in seduta plenaria al Parlamento europeo: Studio delle strategie discorsive di un corpus di discorsi in italiano e interpretati in spagnolo*. Roma: Università degli Studi Internazionali di Roma. (MA thesis).

Bartłomiejczyk, Magdalena. 2016. *Face threats in interpreting: A pragmatic study of plenary debates in the European Parliament*. Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Doctoral dissertation). https://wydawnictwo.us.edu.pl/sites/wydawnictwo.us.edu.pl/files/face_threats_in_interpreting_czw_st_e.pdf.

Bayley, Paul. 2004. Introduction: The whys and wherefores of analysing parliamentary discourse. In Paul Bayley (ed.), *Cross-cultural perspectives on parliamentary discourse*, 1–44. DOI: 10.1075/dapsac.10.01bay.

Beaton, Morven. 2007. Interpreted ideologies in institutional discourse: The case of the European Parliament. *The Translator* 13(2). 271–296.

Beaton-Thome, Morven. 2013. What's in a word? Your 'enemy combatant' is my 'refugee'. The role of simultaneous interpreters in negotiating the lexis of Guantánamo in the European Parliament. *Journal of Language and Politics* 12(3). 378–399. DOI: https://doi.org/10.1075/jlp.12.3.04bea.

Boyd, Michael S. & Claudia Monacelli. 2010. Politics,(con) text and genre: Applying CDA and DHA to interpreter training. *The Interpreters' Newsletter* 15. 51–70.

Bülow-Møller, Anne Marie. 2003. Second-hand emotion: Interpreting attitudes. *The Interpreters' Newsletter* 12. 1–36.

Cavagnoli, Stefania & Laura Mori. 2019. *Gender in legislative languages: From EU to national law in English, French, German, Italian and Spanish*, vol. 144. Berlin: Frank & Timme GmbH.

Coppola, Claudia. 2019. *Analisi etnopragmatica dell'agentività in discorsi istituzionali di politici italiani e della sua interpretazione in tedesco*. Roma: Università degli Studi Internazionali di Roma - UNINT. (MA thesis).

Di Giambattista, Lorella, Viviana Di Felice, Luca Briasco & Letizia Formosa. 2015. *Le politiche dell'Unione europea in materia di controlli alle frontiere, asilo e immigrazione: Normativa di riferimento e prospettive future*. Tech. rep. 215. Servizio Studi del Senato della Repubblica. http://www.senato.it/service/PDF/PDFServer/BGT/00917232.pdf.

Diriker, Ebru. 2004. *De-/Re-contextualizing conference interpreting*. Amsterdam: Benjamins.

Duranti, Alessandro. 2006. Narrating the political self in a campaign for US Congress. *Language in Society* 35(4). 467–497.

European Parliament. 2018. *Gender-neutral language: In the European Parliament*. Tech. rep. https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf.

European Parliament. 2021. *Fact sheets on the European Union*. https://www.europarl.europa.eu/factsheets/en/home (30 September, 2021).

Fairclough, Norman. 1995. *Critical discourse analysis: Papers in the critical study of language*. London: Longman.

Fairclough, Norman. 2003. *Analysing discourse: Textual analysis for social research*. London/New York: Routledge.

Fairclough, Norman. 2006. Genres in political discourse. In Keith Brown (ed.), 2nd edn., 32–38. Boston, MA: Elsevier. DOI: 10.1016/B0-08-044854-2/00719-7.

Gile, Daniel. 1995. *Regards sur la recherche en interprétation de conférence*. Lille: Presses Universitaires de Lille.

Ilie, Cornelia. 2015. Parliamentary discourse. In Karen Tracy (ed.), *Parliamentary discourse*, 1–15. Hoboken, NJ: John Wiley & Sons.

Kent, Stephanie Jo. 2009. A discourse of danger and loss. Interpreters on interpreting for the European Parliament. In Sandra Hale, Uldis Ozolins & Ludmila Stern (eds.), *The critical link 5: Quality in interpreting: A shared responsibility*, vol. 5, 55–70. Amsterdam & Philadelphia: Benjamins.

Kučiš, Vlasta & Simona Majhenič. 2018. Cultural and stress-related manifestations of political controversial language in the European Parliament from the view of interpreters. *Babel* 64(1). 33–62.

Marzocchi, Carlo. 1998. The case for an institution-specific component in interpreting research. *The Interpreters' Newsletter* 8. 51–74.

Monacelli, Claudia. 2009. *Self-preservation in simultaneous interpreting. Surviving the role*. Amsterdam: Benjamins.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli & Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta: Journal des traducteurs/Meta: Translators' Journal* 50(4). DOI: https://doi.org/10.7202/019850ar.

Orrù, Paolo. 2017. *Il discorso sulle migrazioni nell'Italia contemporanea: Un'analisi linguistico-discorsiva sulla stampa (2000-2010)*. Milano: Franco Angeli.

Pearl, Stephen. 1999. The other three eighths & the four f's. Finiteness, fallibility, freedom of speech and fair competition in the simultaneous interpretation environment. *The Interpreters' Newsletter* 9. 3–28.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London: Routledge.

Reisigl, Martin & Ruth Wodak. 2001. *Discourse and discrimination: Rhetorics of racism and antisemitism*. London: Routledge.

Rojo, Luisa Martín & Teun A. van Dijk. 1997. "There was a Problem, and it was Solved!": Legitimating the Expulsion ofIllegal'Migrants in Spanish Parliamentary Discourse. *Discourse & Society* 8(4). 523–566.

Semino, Elena. 2008. *Metaphor in discourse*. Cambridge: Cambridge University Press Cambridge.

Semino, Elena & Michela Masci. 1996. Politics is football: Metaphor in the discourse of Silvio Berlusconi in Italy. *Discourse & Society* 7(2). 243–269.

Shlesinger, Miriam. 1991. Interpreter latitude vs. due process. Simultaneous and consecutive interpretation in multilingual trials. In Sonja Tirkkonen-Condit (ed.), *Empirical research in translation and intercultural studies*, 147–55. Tübingen: Gunter Narr.

Taylor, Charlotte. 2020. Representing the Windrush generation: Metaphor in discourses then and now. *Critical Discourse Studies* 17(1). 1–21. DOI: 10.1080/17405904.2018.1554535.

Turrini, Cinzia. 2004. Metafora e dintorni: L'interpretazione simultanea del linguaggio non letterale al Parlamento europeo. In Gabriele Bersani Berselli, Gabriele Mack & Daniela Zorzi (eds.), *Linguistica e interpretazione*, 125–146. Bologna: CLUEB.

van Dijk, Teun A. 1995. Aims of critical discourse analysis. *Japanese Discourse* 1. 17–27.

van Dijk, Teun A. 1998. *Ideology: A multidisciplinary approach*. London: SAGE.

van Dijk, Teun A. 2003. Knowledge in parliamentary debates. *Journal of Language and Politics* 2(1). 93–129. DOI: 10.1075/jlp.2.1.06dij.

van Dijk, Teun A. 2018. Discourse and migration. In Ricard Zapata-Barrero & Evren Yalaz (eds.), *Qualitative Research in European migration studies* (IMISCOE Research Series), 227–245. Cham: Springer. DOI: 10.1007/978-3-319-76861-8.

van Leeuwen, Theo. 1995. The representation of social actors. In Carmen Rosa Caldas-Coulthard & Malcolm Coulthard (eds.), *Texts and practices*, 32–70. London: Routledge.

Vuorikoski, Anna-Riita. 2004. *A voice of its citizens or a modern tower of Babel? The quality of interpreting as a function of political rhetoric in the European Parliament*. Tampere: Tampere University Press. (Doctoral dissertation).

Wodak, Ruth. 1996. The genesis of racist discourse in Austria since 1989. In Carmen Rosa Caldas-Coulthard & Malcolm Coulthard (eds.), *Texts and practices*, 107–127. London: Routledge.

Wodak, Ruth. 2001. What CDA is About – a Summary of its History, Important Concepts and Its Developments. In Ruth Wodak & Michael Meyer (eds.),

Methods of critical discourse analysis, 1–13. London: SAGE. DOI: 10.4135/9780857028020.n1.

Wodak, Ruth. 2015. The discursive construction of strangers: Analyzing discourses about migrants and migration from a discourse-historical perspective. *Migration and Citizenship. Newsletter of the American Political Science Association* 3. 6–10. https://eprints.lancs.ac.uk/id/eprint/73299/ (30 September, 2021).

Chapter 5

Using the Gravitational Pull Hypothesis to explain patterns in interpreting and translation: The case of concatenated nouns in mediated European Parliament discourse

Marie-Aude Lefer[a] & Gert De Sutter[b]
[a]Université catholique de Louvain [b]Ghent University

In this chapter, we present a corpus study of the French rendition of English concatenated nouns, such as climate change, comparing two modes of interlingual mediation at the European Parliament, namely simultaneous interpreting and written translation. Using parallel corpus data extracted from the European Parliament Translation and Interpreting Corpus, we examine how frequently English concatenated nouns are rendered with semantically equivalent items in the two mediation modes, and which factors stimulate the use of these equivalent (vs non-equivalent) renditions. Alongside the complexity and lexicalization of English concatenated nouns, we consider several frequency-related variables inspired by Halverson's (2017) cognitive linguistic model of translation, the gravitational pull hypothesis. The model posits three cognitive sources of translation effects: gravitational pull (source salience), connectivity (cross-linguistic link strength) and magnetism (target salience). The results show that there are far fewer semantically equivalent renditions in interpreting than in translation. In addition, the regression analysis provides strong evidence that connectivity and magnetism play a crucial role in the selection of semantically equivalent vs non-equivalent renditions in interpretations and translations, alongside the length of source concatenated nouns, with stronger effects in interpreting. By contrast, source-language variables related to gravitational pull and lexicalization do not seem to influence renditions in French. The study brings to the fore key commonalities between translation and interpreting and shows that the three cognitive sources in Halverson's gravitational pull model can be successfully disentangled in a multifactorial research design.

Marie-Aude Lefer & Gert De Sutter. 2022. Using the Gravitational Pull Hypothesis to explain patterns in interpreting and translation: The case of concatenated nouns in mediated European Parliament discourse. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 127–153. Berlin: Language Science Press. DOI: ??

1 Introduction

The last 20 years have seen the application of a wide array of corpus-based and corpus-driven techniques to increasingly large amounts of translated text, in many languages. Corpus-based translation studies (CBTS) has produced numerous descriptions of translation-related phenomena, ranging from translation procedures for specific linguistic items and structures (e.g. culture-specific lexis) to typical features of translated text (e.g. increased explicitness). In recent years, in the wake of Shlesinger's pioneering work in corpus-based interpreting studies (CIS) (Shlesinger 1998), CBTS has progressively branched out to include intermodal studies, where different mediated language varieties are compared (typically, written translation and simultaneous interpreting; cf. Bernardini et al. 2016). This type of intermodal research has been further promoted by **Kotze2020**'s (**Kotze2020**) constrained-language framework, which aims to identify the commonalities between language varieties where constraints of different kinds play an above-average role (see e.g. Kajzer-Wietrzny & Ivaska 2020). The key constraint dimension along which translation and interpreting differ is the 'register/modality' dimension, as translation and interpreting represent written and spoken language production respectively. What they have in common is that they both rely on a preexisting text (the source text or speech) and involve bilingual language processing, where two languages are simultaneously activated, one as the source, the other as the target.

The present study adds to the growing body of intermodal corpus research by examining the French renditions of English noun concatenations (i.e. sequences of at least two nouns, such as *food prices*) in two modes of interlingual mediation commonly practiced at the European Parliament (EP), namely simultaneous interpreting of the speeches delivered during EP plenary sessions and written translation of the official verbatim reports of these speeches. The reason for examining concatenated nouns is that they have been described as difficult, error-prone items in both modalities. Intermodal research on this topic, however, is still scarce. In the present study, we aim to assess the impact of a large set of frequency-, complexity- and lexicalization-related variables on the use of semantically equivalent (vs non-equivalent) renditions in translation and interpreting, drawing both on insights from previous empirical research on noun sequences and on Halverson's (2003, 2007, 2010, 2017) cognitive linguistic model of translation, the gravitational pull hypothesis. The model posits three cognitive sources of translational effects: source language salience (*gravitational pull*), target language salience (*magnetism*) and cross-linguistic link strength (*connec-*

tivity), where salience is operationalized as, among other things, frequency of use. To date, the model has been tested on a handful of linguistic items, such as morphemes and individual lexemes (**Hareide2016**; Vandevoorde 2020; Marco 2021), but it has rarely been used to study items above the word level. However, we believe that the model holds great potential for the study of structures such as concatenated nouns, since psycholinguistic research has shown the crucial role played by frequency in processing and producing these items (cf. Baayen et al. 2010). In addition, to the best of our knowledge the model has not been applied to interpreting, nor has it been used in robust multifactorial research designs such as the one we propose here.

The chapter is structured as follows. §2 presents the phenomenon under scrutiny here, English concatenated nouns and their French equivalents, introduces Halverson's gravitational pull model and shows how it can be used to inform the intermodal study of concatenated nouns. In §3, we describe the corpus data used, the data extraction and coding procedures adopted and the multivariate statistics applied to the dataset at hand. The results of the analysis are presented and discussed in §4 The chapter ends with concluding remarks and suggestions for future research.

2 Background

2.1 Concatenated nouns in English-French translation and interpreting

The notion of 'concatenated noun' is a blanket term for two main types of noun sequence: (1) established (i.e. institutionalized and lexicalized) compounds and multiword terms (e.g. *car insurance*, *food chain*) and (2) non-institutionalized, nonce formations, which are created ad hoc (e.g. *kinship child*, *poultry and pig establishment*) (cf. Bauer 1983: 45–50; Hohenhaus 2005; note that there is no watertight borderline between the two categories, see Bauer 1998). English noun concatenations encompass several structures (also called *patterns* or *schemas*), including N+N (e.g. *pork products*), [N+N]+N (e.g. *trade defence instruments*), [A+N]+N (e.g. *national unity government*) and A+[N+N] (e.g. *small market share*).

Three main aspects of concatenated nouns in English and French are worth considering in contexts of bilingual language production: the complex semantics of English concatenated nouns, and English-French cross-linguistic differences in pattern productivity and word order. First, English N+N sequences are semantically versatile, i.e. they convey a large variety of semantic relations. Classic taxonomies typically range from circa 10 to 50 semantic relations, thereby display-

ing various degrees of granularity (Fernández-Domínguez 2020: 82). For instance, Levi (1978: 75–118; quoted in Fernández-Domínguez 2020) lists nine semantic relations found in English N+N sequences: CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM and ABOUT (*chocolate éclair* 'an éclair which *has* chocolate', for example, illustrates the HAVE relation). In addition, some N+N concatenations display semantic indeterminacy, i.e. they cannot easily be disambiguated or interpreted, even when their co-text is taken into consideration. In a corpus-based study of more than 500 N+N compounds, Fernández-Domínguez (2020) finds that a third of the items under scrutiny can be attributed a second reading on top of the most obvious, primary reading (e.g. *army plan*: ABOUT 'a plan which is *about* the army' vs IN 'a plan which is prepared/implemented *in* the army'). Second, the core N+N pattern, which is attested in both English and French (e.g. *coin cuisine* lit. 'corner kitchen'), is much more productive in English (Paillard 2000: 49–51; Arnaud & Renner 2014). As a result, many English N+N sequences need to be rendered in French by means of other patterns, such as N+A (e.g. *trade agreement* > *accord commercial*) or N+prep+N (e.g. *security wall* > *mur de sécurité*), with cases where the two patterns are found to alternate (e.g. *fishing stocks*: *stocks halieutiques*$_{\text{N+A}}$ vs *stocks de poissons*$_{\text{N+PREP+N}}$). Finally, as regards constituent order, English N+N sequences are typically right-headed (e.g. *timber products*), while their French equivalents, whatever the pattern, are mostly left-headed (e.g. *produits du bois*). This aspect of concatenated nouns has been examined in compound acquisition research, where English-French bilingual children have been shown to produce N+N novel compounds in reversed order (i.e. left-headed in English and right-headed in French), under the influence of crosslinguistic transfer (cf. Nicoladis 2002; see also De Cat et al. 2015).

The contrastive literature on English-to-French translation mentions two major types of translation difficulty (see e.g. Chuquet & Paillard 1987). The first is that, in addition to obvious shifts in word order, English concatenated nouns often require explicitation of the semantic relation that holds between the constituents of the sequence (which, in English, is not overtly expressed), for instance through the insertion of prepositions (e.g. *adoption law* > *loi sur l'adoption*, *foreign policy objectives* > *objectifs en matière de politique étrangère*). Explicitation of the semantic relation between head and modifier(s) in the concatenation is no easy task, in view of the above-mentioned semantic versatility of the English N+N pattern. The second difficulty frequently mentioned in the contrastive literature is that the underlying structure of some of the longer English sequences is potentially ambiguous and hence difficult to interpret and translate. This is often the case when an adjective or a noun premodifies an N+N sequence (e.g. *modern history section*: [*modern history*] *section* vs *modern* [*history section*]). This causes

acute problems in learner translation. In their error analysis of English-to-French student specialized translations, Kübler et al. (2022) find numerous translation errors triggered by English noun phrases whose structure is ambiguous (e.g. *stable solution complexation* > **complexation stable de solution* instead of *complexation en solution stable*).

Similar difficulties are also discussed in the field of interpreting, where it is stressed that the interpretation of English concatenated nouns is effortful because they are informationally very dense and require major syntactic changes (i.e. reordering) in French and other Romance languages (see e.g. Gile 1995 on proper name compounds). Noun concatenations have been investigated empirically in CIS. Relying on the *European Parliament Interpreting Corpus*, Ghiselli (2018) analyzes Italian interpretations of English complex noun phrases (phrases where nominal heads are premodified by several items, be they nouns, adjectives, numbers or participles). She finds that only 55% of English complex noun phrases are rendered successfully in Italian. In her dataset, incomplete or wrong renditions are particularly prominent when source speech delivery is fast (180+ words per minute), pointing to the effect of time constraints on interpreters' renditions of complex noun phrases. In their French-Dutch study based on the *European Parliament Interpreting Corpus Ghent* and the parliamentary debate subcorpus of the *Corpus Gesproken Nederlands*, Defrancq & Plevoets (2018) investigate intra-word filled pauses, including intra-compound pauses, in Dutch interpreted from French and in non-mediated (original) Dutch. The authors provide tentative evidence that "compounds are an important factor in the increase of cognitive load during interpretation" (ibid. 57). As acknowledged by the authors themselves, however, the data sample analyzed is small (the analysis is based on 18 occurrences of intra-compound filled pauses in interpreted Dutch).

In a pilot study based on the *European Parliament Translation and Interpreting Corpus* (Ferraresi & Bernardini 2019), Lefer & De Clerck (2021) find that English concatenated nouns are interpreted with French semantically equivalent renditions in only half of the cases, the other half being made up of incomplete and wrong renditions, as well as omissions. Although based on a small dataset, their qualitative analysis of the disfluencies typically found in the vicinity of incomplete or wrong renditions suggests that three types of N+N sequence are particularly vulnerable in interpreting: ad hoc (i.e. non-lexicalized) sequences, long sequences (made up of 3 constituents or more) and rare (i.e. infrequent) sequences. These preliminary findings point to the potential role of the lexicalization, length and frequency of English concatenated nouns in shaping the use of semantically (non-)equivalent renditions in French. Although admittedly very tentative, this

ties in with the ample evidence provided by psycholinguistic studies on compound processing, where it is shown that compound and constituent length and frequency all play a decisive role in lexical access (see e.g. Baayen et al. 2010). To date, however, length and frequency have not been examined concomitantly in robust multifactorial research designs in corpus-based translation and interpreting studies devoted to concatenated nouns (or compounds in general). This is what we intend to do in the present study, relying on Halverson's gravitational pull model to inform cognitively motivated frequency analyses.

2.2 Applying Halverson's gravitational pull model to the study of concatenated nouns

Combining insights from cognitive grammar, psycholinguistic approaches to bilingualism and second language acquisition research, Halverson (2017) posits three cognitive sources of translational effects (patterns of under- and overrepresentation, source-language interference, normalization, etc.): (1) source language salience (*gravitational pull*), (2) target language salience (*magnetism*) and (3) cross-linguistic link strength (*connectivity*), where salience is operationalized as frequency of use and ease of recall. Gravitational pull is described as "a cognitive force that makes it difficult for the translator to escape the cognitive pull of highly salient representational elements in the source language" (ibid. 14). This force can cause interference in translation. Magnetism is a force that affects the cognitive search for a target language item, whereby "the translator is more likely to be drawn to a target language item with high salience/frequency" (ibid.). Connectivity is "the nature and strength of links between elements in a bilingual's two languages" (ibid.). Halverson's hypothesis is that "the more established (entrenched) a link is, the more likely it will be activated and used in translation, and vice versa" (ibid. 15).

In her 2017 study, Halverson takes as a test case the English polysemous verb *get* and two of its Norwegian equivalents, triangulating monolingual and parallel corpus data, elicitation data and keystroke logs. While her results provide initial support for the posited cognitive forces, some of the predicted overrepresentation patterns are not found in the corpus and keystroke data examined. Vandevoorde (2020) uses the model as a post-hoc interpretative framework in her corpus-based study of the Dutch inchoative verbs *beginnen* 'begin' and *starten* 'start' in Dutch translated from English and French and in non-translated Dutch. Vandevoorde shows that the model can be used to explain some of the patterns observed in her data, but she acknowledges that in some specific cases, several

cognitive forces overlap (e.g. gravitational pull and magnetism), making it difficult to disentangle their cumulative effects.

In addition to polysemous verbs, the gravitational pull model has also been tested on unique items, i.e. linguistic items that "lack straightforward linguistic counterparts in other languages" (Tirkkonen-Condit 2004: 177). As pointed out by Tirkkonen-Condit, unique items are not necessarily untranslatable, rather, "they are simply not similarly manifested (e.g. lexicalized) in other languages" (ibid.). Typically, in this context, translated texts from two source languages are compared: one source language where a given phenomenon is not attested, the other where it is. For example, **Hareide2016** has examined the Spanish gerund in texts translated from English (a language with progressive and non-finite adverbial phrases) and Norwegian (a language that has no gerund), providing strong support for the gravitational pull model. A similar approach is taken in Marco & Oster (2018), which deals with diminutive suffixes in Catalan translated from German (which has productive diminutive suffixes) and English (which has no productive diminutive suffix), and in Marco (2021), devoted to modal verbs expressing obligation and necessity in Catalan translated from English and from French.

Remarkably, few empirical studies have examined items and structures above the morpheme or word level (cf. Halverson 2017: 40) or used multifactorial statistical testing to account for the relative strengths of the three cognitive forces at play. Also notable is the fact that the model has attracted little attention in corpus-based interpreting research to date. In this chapter, we set out to go some way towards remedying these gaps and further exploiting the full potential of Halverson's cognitive model by applying it to a structure situated above the word level (concatenated nouns), in two types of interlingual mediation (written translation and simultaneous interpreting), using multivariate statistics (regression analysis). In doing so, we also aim to extend previous translation and interpreting research on complex noun phrases and nominal compounds by relying on several corpus frequency counts that function as operationalizations of the three cognitive forces included in Halverson's model, namely *gravitational pull* (frequency of the concatenation and its constituents in the source language, here English), *connectivity* (cross-linguistic correspondence of the English source concatenation and its French translation or interpretation) and *magnetism* (frequency of the rendition in the target language, here French). The frequency variables drawn from the gravitational pull model will be considered alongside other factors that have been shown to influence compound processing, namely length and lexicalization, with a view to singling out the factors that condition the use of semantically equivalent (vs non-equivalent) renditions in French. We expect similar trends to

emerge in the two modalities, namely that lexicalization, short length and high frequency (reflecting strong gravitational pull, strong connectivity and/or strong magnetism) will go hand in hand with semantically equivalent renditions. However, we expect the effects of these variables to be more visible in interpreting, in view of the fact that "[b]ecause interpreting affords only limited opportunity for restatement or corrections, it can be seen as the practitioner's default version, with written translation representing a more polished rendition" (Shlesinger & Malkiel 2005: 185). Contrary to interpreting, written translation is an offline activity, often involving the use of resource tools, self-revision and editorial intervention. In other words, our prediction is that ad hoc concatenations, long concatenations, and concatenations that display low gravitational pull, low connectivity and/or whose equivalents display low magnetism will trigger incomplete renditions, wrong renditions and omissions more frequently in interpreting than in translation.

3 Data and methodology

3.1 Corpus data used

In this study, we made use of corpus data extracted from the *European Parliament Translation and Interpreting Corpus* (EPTIC; Bernardini et al. 2016; Ferraresi & Bernardini 2019)[1]. EPTIC is a multilingual intermodal corpus developed at the University of Bologna in collaboration with other European universities, among them UCLouvain in the case of the English-French language pair. The corpus comprises four components, two spoken, two written: transcripts of speeches delivered at the EP and transcripts of their simultaneous interpretations; verbatim reports of the same speeches and their official written translations. Transcriptions are performed on the basis of the videos of the plenary sessions made available online by the EP, adhering to detailed transcription conventions specifically developed for EPTIC. Verbatim reports and their official translations are derived from the EP website, where EP proceedings are archived and available to the public. One of the unique features of the corpus is that the source speeches (spoken component) and the source verbatim reports (written component) are almost identical, which makes it possible to study the interpretations and translations of practically the same input. The corpus, whose compilation is still ongoing at the time of writing, is made available to the research community through the NoSketch Engine platform (Rychlý 2007), the open source version of Sketch Engine (Kilgarriff et al. 2014). It is sentence-aligned and POS-tagged.

[1] https://corpora.dipintra.it/eptic/

In this study, we relied on 106 speeches delivered in English by Members of the European Parliament, commissioners and guests, and their French simultaneous interpretations by highly skilled professionals who are all native speakers of French. We also used the verbatim reports of these speeches and their French translations. No information is available on the translators who produced the translations included in EPTIC, but it can be assumed that they are also highly skilled professionals translating into their native language. The spoken and written components of the subcorpus used in the study each total ca 60,000 tokens (see Table 1).[2]

Table 1: Size of the English-to-French EPTIC subcorpus used in the study (in tokens)

	English sources	French targets	Total
Spoken component	29,457	28,317	57,774
Written component	28,068	31,897	59,965

To code the EPTIC dataset with reference frequencies in English, French and English-French translation (see §3.2), we used the Europarl corpus as a reference corpus (Koehn 2005). Europarl is a multilingual parallel corpus that comprises the EP verbatim reports produced between 1996 and 2011 (the year that translation of the reports was discontinued at the EP). Europarl is here taken to be representative of EP discourse as a whole, monolingually (EP discourse in English and EP discourse in French) and bilingually (EP discourse in the English-French pair). We used version 7 (Europarl7), available on Sketch Engine. The English version of Europarl7 totals 53+ million tokens. It is a mix of verbatim reports of speeches originally delivered in English (by native or non-native speakers of English) and speeches originally delivered in other languages and subsequently translated into English. The French version of Europarl7 contains 59+ million tokens. Like the English version of the corpus, it is comprised of verbatim reports of speeches originally delivered in French (by native speakers of French, with few exceptions) alongside speeches delivered in other languages and translated into French (in some cases with English as a pivot language). The English and French versions of the Europarl corpus, which are sentence-aligned, was also

[2]The source speeches included in the English-to-French EPTIC subcorpus used in the study were read-out (44% of the subcorpus), impromptu (34%) or mixed (22%), with an average speed of delivery of 161 words per minute. They were given by both native and non-native speakers of English (corresponding to 60% and 40% of the subcorpus, respectively).

used in the present study as a parallel corpus. It is important to stress, however, that it was used here as a *non-directional* parallel corpus, in the sense that we disregarded translation directions and use of English as a pivot (cf. Lefer 2020: 259). In other words, the full English-French parallel Europarl used as a reference corpus in the study includes texts in original English translated into French, texts in original French translated into English and texts produced in other languages and translated into both English and French.

3.2 Data extraction and coding

English concatenated nouns used in EPTIC source speeches and verbatim reports were automatically extracted on the basis of a CQL query aimed at identifying all sequences of at least two common nouns. Irrelevant occurrences were then manually removed, i.e. POS-tagging errors (e.g. *the consequences of* ***printing*** **money** too cheaply), contiguous nouns that are not concatenated (e.g. *all the* ***remarks*** **people** have made) and strings containing titles (e.g. *madam chairman*).

The resulting dataset contains 853 occurrences, equally distributed among the spoken and written components of the subcorpus used[3] , which were then manually matched with their renditions in interpreted and translated outputs, relying on EPTIC sentence alignment. All occurrences were coded for the response variable 'semantically equivalent rendition' vs 'semantically non-equivalent rendition'. Transfer of meaning is central to this distinction. The 'semantically equivalent rendition' category was attributed to outputs where the propositional content conveyed by the source concatenation was also found in the interpretation or translation, as in *euro crisis* > *crise de l'euro* and *tax evasion* > *évasion fiscale*. This category also includes (rare) cases where the semantic relation that holds between the constituents of the source concatenation is explicitated in the output (e.g. *rural development policy* > *politique dans le domaine du développement rural*) (cf. **Wadensjoe1998** on expanded renditions). The 'semantically non-equivalent rendition' category, by contrast, subsumes three types of rendition: (1) incomplete renditions, where part of the propositional content found expressed in the source concatenation is left out in the output (e.g. *adoption process* > *processus* 'process', European fishing industry > industrie européenne 'European industry'); (2) wrong renditions, where the propositional content of the rendition is not semantically equivalent to that of the original (cases of misinterpretation, incorrect meaning, etc.) (e.g. *export figures* > importations 'import', partner country > pays d'origine 'country of origin') (**AmatoMack2011**); and (3) omissions, when source concatenations are entirely omitted in the output.

[3] All but nine occurrences occur in both source speeches and corresponding verbatim reports.

The data were also coded for the following explanatory variables: speech-text id (unique id attributed to pairs of source speeches and corresponding verbatim reports), modality (translation or interpreting) and ten gravitational-pull-, magnetism-, connectivity-, complexity- and lexicalization-related variables, which are all described and illustrated below.

3.2.1 Gravitational-pull-related variables

The gravitational pull of source concatenated nouns (i.e. their salience in English EP discourse) was operationalized by means of two corpus frequency variables: (i) their overall frequency and (ii) the average frequency of their individual constituents. Relative frequencies per million words were computed on the basis of the full English version of Europarl7 (53+ million tokens). The reason for operationalizing gravitational pull both as concatenation frequency and constituent frequency is that, as mentioned in §2.2, psycholinguistic research has shown that nominal compounds are accessed both as wholes and via their component parts (Baayen et al. 2010; Gagné 2011). Examples of noun concatenations with particularly strong gravitational pull in EP discourse (at the level of the whole concatenation) include *labour market*, *action plan* and *climate change*, while weak gravitational pull items are, for instance, *birth country*, *legality assurance system* and *tuna processing facilities*. Some of the items that display a weak concatenation-based gravitational pull exert a rather strong pull at the level of their individual constituents, such as *information measures*, *development needs*, *security situation* and *market construction*, which shows the usefulness of including different frequency operationalizations of gravitational pull when dealing with items above the word level.

3.2.2 Magnetism-related variable

The magnetism (i.e. salience) of the French renditions found in interpreted and translated outputs was operationalized as their overall frequency (normalized per million words) in the French version of Europarl7 (59+ million tokens). Renditions with strong magnetism in French EP discourse include, for instance, *sécurité alimentaire*, *proposition de resolution* and *états membres*. Examples of weak-magnetism renditions are *accords en matière de pêche*, *droit familial* and *coût de l'énergie*.

3.2.3 Connectivity-related variables

The strength of the cross-linguistic link between a given source noun concatenation and its rendition in translation or interpreting (i.e. connectivity) was operationalized on the basis of both bilingual lexicographic/terminographic and parallel corpus data. First, we coded whether the source noun concatenation and its rendition were recorded as equivalents in English-French bilingual entries (i) in the Interactive Terminology for Europe (IATE) database and (ii) in the subscription-based Oxford English-French bilingual dictionary. We chose to rely on the Oxford bilingual dictionary because contrary to other online English-French dictionaries, the two sides of the dictionary can be queried simultaneously, with direct access to main entries and subentries. Three values were used to code these two connectivity-related lexicographic variables: YES (when the concatenation and its rendition were listed as equivalents in IATE or the Oxford bilingual dictionary, e.g. *interest rate* > *taux d'intérêt*), NO (when they were not, e.g. *dioxin scare* > *alerte à la dioxine*) and PARTIAL (for longer concatenations, when part of the source concatenation and part of its rendition were recorded as equivalents in a bilingual entry in IATE or Oxford, e.g. *draconian maternity leave* > *congé de maternité draconien*). In addition, we relied on a frequency-based variable, taking advantage of the parallel nature of the Europarl corpus. For each pair of source concatenation and corresponding rendition in either interpreting or translation, we computed a Pointwise Mutual Information (PMI) score on the basis of (i) the frequency of the source noun concatenation in Europarl7-English, (ii) the frequency of its rendition in Europarl7-French and (iii) the frequency of their cross-linguistic correspondence in the English-French parallel version of Europarl7. Specifically, we used the following formula, where p = probability, s = source (English noun concatenation), t = target (French rendition), s-t = source-target correspondence: $\log (p(s\text{-}t)^3 / p(s) * p(t))$ (cf. Role & Nadif 2011). The higher the PMI^3 score, the stronger the connectivity. For example, the pair *health services-services de santé* displays a stronger cross-linguistic link in EP discourse (PMI^3 = 2.77) than the pair *health services-services en matière de santé* (PMI^3 = -6.93). The main advantage of PMI^3, compared with other corpus-based measures of correspondence (such as Altenberg's (1999) mutual translatability), is that it does not give excessive scores to pairs that involve low-frequency items (Role & Nadif 2011). These low-frequency pairs are in fact quite numerous in the dataset at hand (some English-French pairs from our EPTIC dataset occur only one or twice in the whole Europarl corpus).

3.2.4 Complexity-related variables

We coded the length of the source concatenations in terms of the number of constituents they contain, distinguishing between concatenations made up of two words and those made up of three or more words. To account for potentially complex (and hence cognitively demanding) co-text, we also coded whether the noun concatenation under scrutiny was embedded in a larger noun phrase, as in *the carbon footprint of Brazilian beef* or *part of our contribution to global food security*. The main reason for including these two complexity-related variables is that Halverson's (2017) model in its current form, being primarily aimed at translated text, does not cater for some of the cognitive constraints inherent in online tasks such as simultaneous interpreting (e.g. time constraints, memory load). In the case of concatenated nouns, we expect long noun sequences and sequences embedded in larger noun phrases to be responsible for increases in cognitive load in interpreting (cf. Defrancq & Plevoets 2018), and hence to be potential triggers for non-equivalent renditions.

3.2.5 Lexicalization-related variables

Finally, in line with Lefer & De Clerck's (2021) observations, and because we examined concatenated nouns irrespective of their status as syntactic constructions (noun phrases) or lexical units (nominal compounds), we also coded lexicalization. Practically speaking, it was operationalized as attestedness in the Oxford English Dictionary (OED) and in IATE, whether as a main entry or subentry (cf. Hilpert 2019). For each lexicographic variable, we distinguished between lexicalized concatenations (e.g. *energy efficiency*, *food chain*, *free trade zone*, *road map*, listed in OED and IATE), partially lexicalized concatenations (e.g. *excessive price volatility*, with *price volatility* listed in IATE) and ad hoc concatenations (e.g. *transportation corridor*, *pork product*, which are not recorded in these two resources).

Table 2 provides an overview of the explanatory variables used in the present study.

3.3 Statistical testing

Preliminary tests on the frequency variables discussed in §3.2, which are all numerical, showed that their distribution was skewed. They were therefore log-transformed. We measured the simultaneous effect of the explanatory variables on our response variable, namely the use of a semantically equivalent vs non-equivalent rendition, by means of a generalized linear mixed-effects model (glmm),

Table 2: Overview of the explanatory variables used in the study

	Variable	Description
modality	modality	Simultaneous interpreting vs written translation
gravitational pull	freq_concat	Relative frequency of the source noun concatenation in English EP discourse (per million words)
	freq_constit	Average relative frequency of the individual constituents of the source noun concatenation in English EP discourse (per million words); calculated by adding up the lempos frequencies of constituents and dividing the sum by the total number of constituents in the concatenation
magnetism	freq_rendition	Relative frequency of the rendition of the source noun concatenation in French EP discourse (per million words)
connectivity	connect_PMI	Pointwise mutual information (PMI^3) of the source noun concatenation and its rendition in English-French EP discourse (per million words)
	connect_IATE	Cross-linguistic link between the source noun concatenation and its rendition as recorded in an entry or subentry of the Interactive Terminology for Europe (IATE) database: yes, partial, no
	connect_bil_dic	Cross-linguistic link between the source noun concatenation and its rendition as recorded in an entry or subentry in the online Oxford English-French bilingual dictionary: yes, partial, no
complexity	source_length	Length of the source noun concatenation, measured as the number of words it contains: 2 words vs 3 or more words
	NP_embedding	Embeddedness of the source noun concatenation in a larger noun phrase (whether as head or as postmodifier): yes, no
lexicalization	lex_OED	Attestedness of the source noun concatenation in the online Oxford English Dictionary: yes, partial, no
	lex_IATE	Attestedness of the source noun concatenation in the Interactive Terminology for Europe (IATE) database (irrespective of a potential cross-linguistic link with a French equivalent): yes, partial, no

using RStudio 1.1.383 (R Core Team 2018). The regression model we used makes it possible to determine whether the test variables have a statistically significant effect on the response variable, what the effect of each variable is and what the overall performance of the model is in terms of descriptive and predictive adequacy.

4 Results and discussion

Figure 1 shows that semantically non-equivalent renditions account for 26% of the EPTIC dataset (n=224/853), while the remaining 74% are equivalent renditions (n=629/853). As shown in Figure 2, however, the distribution of the two types of rendition is markedly different in the two mediation modes: while translators produce semantically equivalent renditions in an overwhelming 96% of cases (n=407/422), the proportion drops to a mere 52% in simultaneous interpreting (n=222/431). This intermodal difference is statistically significant ($\chi^2(1)=220.04$, $p < 2.2e\text{-}16$). This figure is very similar to the proportion of successful renditions reported in Ghiselli's (2018) analysis of Italian interpretations of English complex noun phrases (55%) and provides additional evidence that English concatenated nouns are vulnerable in simultaneous interpreting into Romance languages (cf. Gile 1995), leading as they do to substantial numbers of incomplete and wrong renditions.

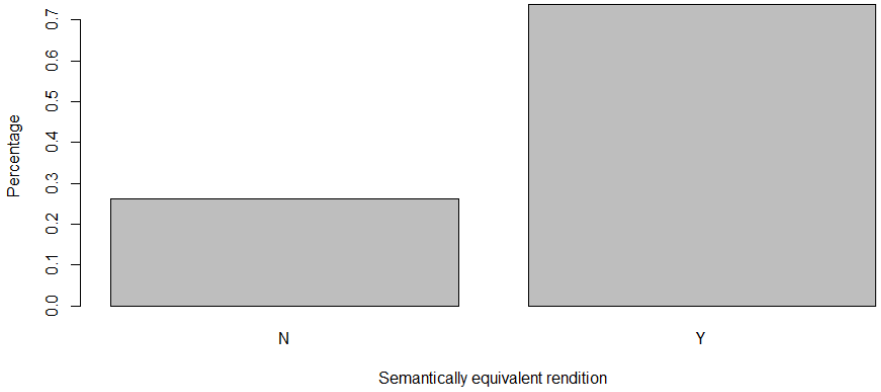


Figure 1: General distribution of semantically equivalent vs non-equivalent renditions of English concatenated nouns (n=853).

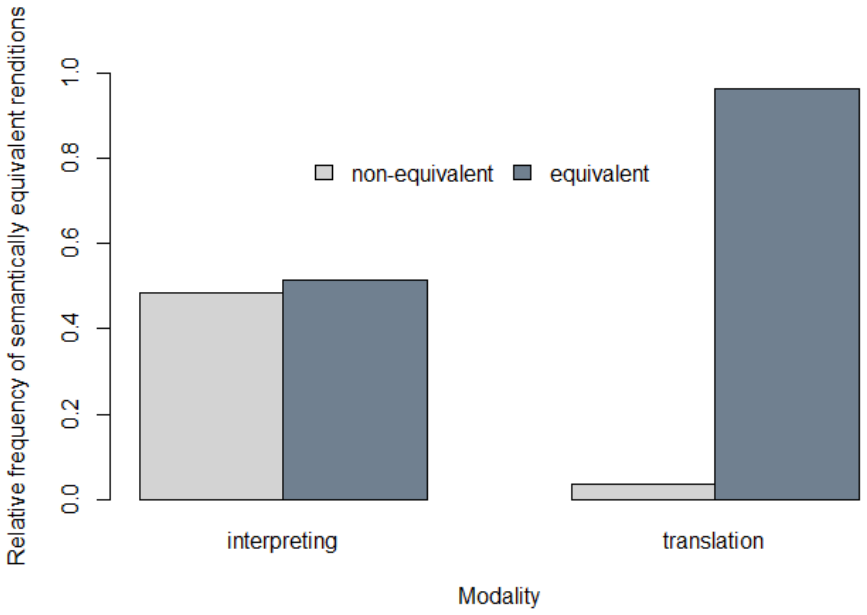


Figure 2: Association of modality (interpreting vs translation) and use of semantically equivalent vs non-equivalent renditions (n=853).

In the remainder of this section, we take a multifactorial approach to the EPTIC dataset at hand with a view to assessing how the variables under scrutiny simultaneously condition the use of semantically equivalent renditions (vs non-equivalent renditions) in the two mediation modes. We ran a glmm-model, using (non-)equivalent rendition as response variable. Modality and the ten above-mentioned gravitational-pull-, magnetism-, connectivity-, complexity- and lexicalization-related variables were used as fixed effects, with speech/text id as random effect to accommodate variation across individual speeches. We adopted a stepwise procedure, starting from a null model containing only the random intercepts and then incrementally adding fixed effects which significantly reduced the Akaike Information Criterion (AIC) value of the model. Next to the main effect of each of the fixed factors, we also checked whether a model with two-way interactions containing modality significantly reduced the AIC value. We avoided overfitting by adopting the rule of thumb that the number of regressors multiplied by 20 should not be higher than the least frequent level of the response variable (cf. Harrell 2015: 72).

The significant fixed effects emerging from the glmm are shown in Figure 3 (the full model is given in Appendix A). This model, which contains two main effects and two interaction effects, outperforms an intercept-only model significantly ($\chi^2(8) = 483.1$, $p < 2.2e\text{-}16$). The marginal R^2 value is 0.69, the conditional R^2 value is 0.72 and the c-score is 0.94. These indicate that the model performs very well in explaining and predicting the variation at hand.

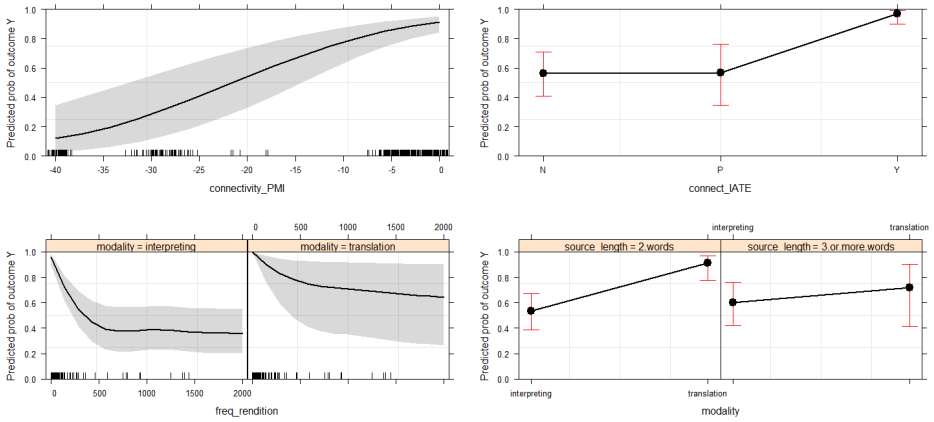


Figure 3: Effect plots of a generalized linear mixed-effects model with semantically equivalent vs non-equivalent rendition as response variable, connectivity_PMI, connectivity_IATE, freq_rendition, and source_length as fixed effects and speech/text id as random effect (n=853).

Three main trends emerge from Figure 3. First, we see that the probability of using a semantically equivalent rendition increases with connectivity (as measured by two variables: the corpus frequency-based PMI^3 score and inclusion of a given source-target pair in an IATE bilingual entry). Chances of using a semantically equivalent rendition almost reach 100% when PMI^3 scores are at their highest (strong cross-linguistic link of the source-target pair in parallel Europarl) and when the source concatenation and its rendition are recorded as cross-linguistic equivalents in IATE. This is in line with one of the basic tenets of Halverson's cognitive model that "the more established (entrenched) a link is, the more likely it will be activated and used in translation" (2017: 15) and, we should add, in interpreting too. Second, we find that the probability of using a semantically equivalent rendition decreases as magnetism increases (as measured here by the frequency of the rendition in French EP discourse). This shows that in the case of concatenated nouns, when translators and, even more, interpreters are "drawn

to a target language item with high salience/frequency" (ibid. 14), they are actually drawn to renditions that are not equivalent to the source concatenated nouns (incomplete renditions or wrong renditions). Examples of such cases from interpreting include *partner country* > *pays d'origine* 'country of origin', *young university graduate* > *étudiant* 'student', *dioxin contamination* > *pollution* 'pollution', *league table* > *liste* 'list'. Third, the glmm shows that the probability of using a semantically equivalent rendition decreases when source noun concatenations are longer (three or more constituents). Importantly, the effects of the latter two variables (magnetism and noun concatenation length) are significantly stronger in interpreting than in translation, as indicated by the two-way interactions with modality in the glmm.

In view of the above observations, we can say that our initial predictions are only partly borne out. Only two of the three cognitive forces from Halverson's (2017) model are found to shape translators' and interpreters' use of semantically equivalent vs non-equivalent renditions, i.e. connectivity (i.e. cross-linguistic link strength, in two of its guises, one based on parallel corpus frequencies, the other on terminographic data) and magnetism (i.e. target language salience, here operationalized as frequency of use in EP discourse). While strong connectivity goes hand in hand with the use of semantically equivalent renditions, strong magnetism pulls in the opposite direction, as it draws translators, and more strikingly, interpreters, to the use of non-equivalent renditions (be they incomplete or wrong). Gravitational pull (i.e. source language salience), contrary to our expectations, does not seem to impact on the use of semantically equivalent vs non-equivalent renditions in our data. In addition, the results of the regression analysis confirm the crucial role played by the length of source concatenated nouns, a complexity-related variable. This shows that source-language-related variables have an effect on content transfer (or lack thereof) in the target language, though not an effect that is directly related to frequency or salience in the source language. Overall, we find that the same factors condition the use of semantically equivalent vs non-equivalent renditions similarly in simultaneous interpreting and written translation, with two predictors (magnetism and source concatenation length) having a significantly stronger effect in interpreting. This is in line with our initial expectation that the same factors condition renditions across mediation modes, but that their effect should be more visible in interpreting, given its very specific constraints (time, memory load, etc.). Note, finally, that lexicalization in the source language does not appear to be a driving force here.

5 Conclusion

This chapter represents the first attempt at addressing the full complexity of Halverson's (2017) gravitational pull model to account for the rendition of a linguistic phenomenon above the word level, concatenated nouns, across two modes of interlingual mediation, simultaneous interpreting and written translation, using robust multifactorial statistics that make it possible to assess simultaneously the effect of the three cognitive forces in play. A key finding emerging from the regression analysis in that regard is that both connectivity and magnetism exert a strong influence on translators' and interpreters' use of semantically equivalent renditions of English concatenated nouns. While highly entrenched cross-linguistic links draw translators and interpreters alike to semantically equivalent renditions, the opposite force is observed in the case of magnetism (target language salience), with strong magnetism leading translators and, more particularly, interpreters to the use of semantically non-equivalent renditions, such as incomplete renditions and wrong renditions. We have proposed an elaborate research design to operationalize gravitational pull, magnetism and connectivity in both translation and interpreting, relying on a large reference corpus and on the pointwise mutual information score to derive cognitively motivated corpus frequency variables. We have also complemented the predictors inspired by Halverson's gravitational pull hypothesis with complexity- and lexicalization-related predictors, so as to better account for the specific features of the linguistic phenomenon at hand, concatenated nouns, and, more generally, interpreting.

One striking result of the present investigation is the lack of a source-language-induced pull effect, which raises the following question: is the rendition of more complex linguistic structures such as noun concatenations not affected by such a mechanism (an outcome which should be interpreted along cognitive-linguistic lines) or do the research topic and research design we have adopted simply prevent a pull effect from emerging (a methodological reason)? It must be acknowledged that our research topic and our research design are both quite different from those adopted in previous GPH research (e.g. Halverson 2017, Marco 2021), and that this may have impacted the results we obtained. The linguistic phenomenon under scrutiny here has no direct formal equivalent in the target language: as mentioned above, English noun concatenations are right-headed, whereas their French equivalents are left-headed, and very often also require the insertion of a preposition or the transposition of the modifier noun into an adjective. This makes a *formal pull effect*, such as would cause the structure of the source construction to shine through in the target text, highly unlikely (at least

in professional translation and interpreting). In addition, even if translators and interpreters occasionally rendered this English structure in an (ungrammatical) word-for-word fashion, our current research design would not capture it, since the central variable in the study focuses on semantic equivalence, and not on the formal features of the renditions. With the benefit of hindsight, these two aspects may explain why a gravitational pull effect was not very likely to occur in the present study.

Admittedly, gravitational pull effects do not occur only through specific (unconventional or ungrammatical) constructions that are formally similar to source-text constructions, but can also emerge at a more aggregate level, namely when a certain linguistic phenomenon in translated language is over- or underused in comparison with non-translated language as a result of a higher or lower frequency of the equivalent representation in the source language (Halverson 2017, for instance, has studied this type of gravitational pull at a semantic level). Coming back to the present study, it is possible, for example, that noun concatenations in translated or interpreted French are used significantly more often in comparison with original French under the influence of the high frequency of noun concatenations in the English source language. Once again, however, our research design does not make it possible to detect that type of pull effect, since we adopted a *parallel-corpus* design (not a *comparable-corpus* design, as in previous studies) in which the translation of *individual* source-language items was analyzed (and not just aggregate patterns of over- and underrepresentation).

In other words, the question remains whether the topic and research design we adopted in this study were suited to picking up on gravitational pull effects. One could argue, of course, that highly salient noun concatenations can affect the semantically equivalent rendering in French positively, but it is hard to think of such an effect that works independently of a connectivity or a magnetism effect: for translators and interpreters, highly salient noun concatenations in the source language will unavoidably also have a high connectivity effect, i.e. the more frequent a construction in the source language, the likelier a translator or interpreter is to have encountered this construction before and hence the likelier she is to have a routinized translation solution at her disposal.

These considerations raise important questions as to how the theoretical model developed by Halverson can be tested in a variety of empirical research designs. It is important to stress, however, that the gravitational pull model, which aims to be a comprehensive cognitive-linguistic model that can be used to explain and predict translational choices, should not be restricted to studies on over- and underuse of particular linguistic phenomena based on comparable corpora (even though the model originated from that type of research), but that it can also be

used to account for local translation choices, above the word level, such as the ones studied in this chapter (Halverson, personal communication).

Although we believe that the present study has gone some way towards showing how the gravitational pull model can be tested empirically in all its complexity, thereby paving the way for further elaboration of the model, it can be complemented in several ways. First, the operationalizations of the three cognitive forces included in the gravitational pull model can be refined. For gravitational pull (source language salience), another variable worth considering is the productivity of nouns in the semantic relations in which they are frequently used, whether as heads or as modifiers (cf. Krott et al. 2009; Fernández-Domínguez 2020). In our dataset, for instance, we noticed that some nouns are particularly productive in EP concatenated nouns, either as heads (e.g. *cattle products*, *construction products*, *pork products*, *timber products*) or premodifiers (e.g. *trade agreement*, *trade benefit*, *trade flow*, *trade partner*). Corpus-derived operationalizations of magnetism (target language salience) could be refined along the same lines, also taking into consideration the magnetism exerted by competing equivalents in the target language. Connectivity (cross-linguistic link strength) also deserves closer attention. In particular, variables indexing the connectivity of individual constituents also need to be taken into consideration, ideally distinguishing between senses for polysemous nouns (e.g. *plant* in *plant species* vs *tuna processing plant*; see Schäfer & Bell 2020) and between cognate vs non-cognate equivalents (cf. Shlesinger & Malkiel 2005).

The length of source concatenated nouns (taken as a proxy for complexity) also emerges as a driving force behind translators' and – more crucially – interpreters' renditions. Care should therefore be taken to examine other length-related variables (e.g. constituent length, in characters) together with variables related to the temporal and cognitive constraints inherent in simultaneous interpreting, relying on cognitive-load-related parameters often investigated in CIS. These include, at the level of the speech, delivery rate, use of numbers, lexical density, syntactic complexity and formulaicity (see e.g. Plevoets & Defrancq 2018). From the perspective of the constrained-language framework (**Kotze2020**), it would also be interesting to consider the native vs non-native status of the speakers ('proficiency' constraint dimension), together with directionality (translation/interpreting into the native vs non-native language; 'language activation' constraint). Likewise, the sociocultural and technological factors that typically constrain written translation should, ideally, also be taken on board to better account for intermodal commonalities and differences. Finally, it should be borne in mind that corpus data ultimately need to be complemented with other data types, such as elicitation data, as "corpus data gives us only indirect evidence of cognitive linguistic

structure" (Halverson 2017: 22).

We hope that the intermodal research design proposed in this study, together with the avenues for future work we have outlined above, will lead to more systematic and more refined explorations of the gravitational pull model in empirical translation and interpreting studies and, in the longer run, to a better understanding of the commonalities and differences that typify mediated language varieties.

Acknowledgements

We wish to thank the two anonymous reviewers for their insightful feedback on an earlier version of the chapter. Their comments and suggestions helped us to broaden the scope of our initial analyses and improve our research design in many ways. We also wish to thank Sandra Halverson for fascinating discussions on the GPH and how it can be applied to the parallel corpus study of concatenated nouns. Her insights proved an invaluable help in better interpreting the findings of the present study and their implications for the GPH. Any remaining shortcomings are of course our own. Thanks are also due to the EPTIC project directors at the University of Bologna for their coordination and constant support, and to the UCLouvain students who contributed to the collection, transcription and alignment of the EPTIC subcorpus used here: Salomé Debève, Athina Danhier, Marie De Clerck, Émilie Degueldre, Tiffany Scohy, Sophie Steil, Fiona Thewissen, Florent Thirion, Antoine Van Gompel and Coraline Zizi. We also thank Salomé Debève for extracting the corpus data and coding part of it.

Appendix A Generalized linear mixed model

A.1 Random effects

Groups	Name	Variance	Std. Dev.
text_id	(Intercept)	0.4187	0.6471

Number of obs: 853, groups: text_id, 89

A.2 Fixed effects

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.862675	0.260998	3.305	0.000949 ***
modalitytranslation	3.069382	0.539696	5.687	1.29e-08 ***
connectivity_PMI	0.108200	0.016236	6.664	2.66e-11 ***
log(freq_rendition+1e-06)	-0.173665	0.036375	-4.774	1.80e-06 ***
connect_IATEP	0.009811	0.341616	0.029	0.977089
connect_IATEY	3.290843	0.642208	5.124	2.99e-07 ***
modalitytranslation: log(freq_rendition+1e-06)	-0.249602	0.101689	-2.455	0.014106 *
modalityinterpreting: source_length3.or.more.words	0.283585	0.349198	0.812	0.416731
modalitytranslation: source_length3.or.more.words	-1.405105	0.693004	-2.028	0.042606 *

References

Altenberg, Bengt. 1999. Adverbial connectors in. English and Swedish: Semantic and lexical correspondences. *Language and Computers* 26. 249–268.

Arnaud, Pierre J. L. & Vincent Renner. 2014. English and French [NN] $_{N}$$ lexical units: A categorial, morphological and semantic comparison. *Word Structure* 7(1). 1–28. DOI: 10.3366/word.2014.0054.

Baayen, R. Harald, Victor Kuperman & Raymond Bertram. 2010. Frequency effects in compound processing. In Sergio Scalise & Irene Vogel (eds.), *Cross-disciplinary issues in compounding* (Current Issues in Linguistic Theory 311), 257–270. Amsterdam/Philadelphia: Benjamins. DOI: 10.1075/cilt.311.20baa.

Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge university press.

Bauer, Laurie. 1998. When is a sequence of two nouns a compound in English? *English Language and Linguistics* 2(1). 65–86. DOI: 10.1017/S1360674300000691.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC |Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Chuquet, Hélène & Michel Paillard. 1987. *Approche linguistique des problèmes de traduction anglais-français*. Paris: Editions Ophrys.

De Cat, Cecile, Ekaterini Klepousniotou & R. Harald Baayen. 2015. Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.00077.

Defrancq, Bart & Koen Plevoets. 2018. Over-uh-load, filled pauses in compounds as a signal of cognitive load. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 43–64. Singapore: Springer.

Fernández-Domínguez, Jesús. 2020. Remarks on the semantics and paradigmaticity of NN compounds. *The Mental Lexicon* 15(1). 79–100. DOI: 10.1075/ml.00015.fer.

Ferraresi, Adriano & Silvia Bernardini. 2019. Building EPTIC: A many-sided, multi-purpose corpus of EU parliament proceedings. In Irene Doval & M. Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*, vol. 90 (Studies in Corpus Linguistics), 123–139. Amsterdam/Philadelphia: John Benjamins.

Gagné, Christina L. 2011. *Psycholinguistic perspectives*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199695720.013.0013.

Ghiselli, Serena. 2018. The translation challenges of premodified noun phrases in simultaneous interpreting from English into Italian. In Claudio Bendazzoli, Mariachiara Russo & Bart Defrancq (eds.), *inTRAlinea special issue: New findings in corpus-based interpreting studies*. https://www.intralinea.org/specials/article/2322.

Gile, Daniel. 1995. *Regards sur la recherche en interprétation de conférence*. Lille: Presses Universitaires de Lille.

Halverson, Sandra L. 2003. The cognitive basis of translation universals. *Target. International Journal of Translation Studies* 15(2). 197–241.

Halverson, Sandra L. 2007. Investigating gravitational pull in translation: The case of the English progressive construction. In Riitta Jääskeläinen, Tiina Puurtinen & Hilkka Stotesbury (eds.) (Publications of the Savonlinna School of Translation Studies 5), 175–195. Joensuu: University of Joensuu, Savonlinna School of Translation Studies Joensuu University Library.

Halverson, Sandra L. 2010. Cognitive translation studies: Developments in theory and method. In Gregory M. Shreve & Erik Angelone (eds.), *Translation and cognition* (American Translators Association Scholarly Monograph Series 15), 349–369. DOI: 10.1075/ata.xv.18hal.

Halverson, Sandra L. 2017. Gravitational pull in translation: Testing a revised model. In Gert De Sutter & Isabelle Lefer Marie-Aude und Delaere (eds.), *Empirical translation studies* (Trends in Linguistics. Studies and Monographs [TiLSM] 300), 9–45. DOI: 10.1515/9783110459586-002.

Harrell, Frank E., Jr. 2015. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.

Hilpert, Martin. 2019. Lexicalization in morphology. In *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780199384655.013.622.

Hohenhaus, Peter. 2005. Lexicalization and institutionalization. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 353–373. Dordrecht: Springer. DOI: 10.1007/1-4020-3596-9_15.

Kajzer-Wietrzny, Marta & Ilmari Ivaska. 2020. A Multivariate Approach to Lexical Diversity in Constrained Language. *Across Languages and Cultures* 21(2). 169–194.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1(1). 7–36. DOI: 10.1007/s40607-014-0009-9.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, vol. 5, 79–86. Phuket: AAMT. http://mt-archive.info/MTS-2005-Koehn.pdf.

Krott, Andrea, Christina L. Gagné & Elena Nicoladis. 2009. How the parts relate to the whole: Frequency effects on children's interpretations of novel compounds. *Journal of Child Language* 36(1). 85–112. DOI: 10.1017/S030500090800888X.

Kübler, Nathalie, Alexandra Mestivier & Moja Pecman. 2022. Using comparable corpora for translating and post-editing complex noun phrases in specialised texts. Insights from English-to-French specialised translation. In Sylviane Granger & Marie-Aude Lefer (eds.), *Extending the scope of corpus-based translation studies*. New York: Bloomsbury.

Lefer, Marie-Aude. 2020. Parallel corpora. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 257–282. Cham. DOI: 10.1007/978-3-030-46216-1_12.

Lefer, Marie-Aude & Martine De Clerck. 2021. L'apport des corpus intermodaux en lexicologie contrastive : étude comparative de la traduction écrite et de l'interprétation simultanée des séquences de noms. In Hanote Sylvie & Nita Raluca (eds.), *Morphophonologie, lexicologie et langue de spécialité*, 145–162. Rennes: Presses Universitaires de Rennes.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Marco, Josep. 2021. Testing the Gravitational Pull Hypothesis on modal verbs expressing obligation and necessity in Catalan through the COVALT corpus. In Mario Bisiada (ed.), *Empirical studies in translation and discourse*, 27–52. Berlin: Language Science Press. DOI: 10.5281/zenodo.4450079.

Marco, Josep & Ulrike Oster. 2018. The gravitational pull of diminutives in Catalan translated and non-translated texts. Paper presented at the Using Corpora in Contrastive and Translation Studies Conference.

Nicoladis, Elena. 2002. What's the difference between'toilet paper'and'paper toilet'? French-English bilingual children's crosslinguistic transfer in compound nouns. *Journal of child language* 29(4). 843.

Paillard, Michel. 2000. *Lexicologie contrastive anglais-français: Formation des mots et construction du sens*. Paris: Editions OPHRYS.

Plevoets, Koen & Bart Defrancq. 2018. The cognitive load of interpreters in the European Parliament. A corpus-based study of predictors for the disfluency uh(m). *Interpreting* 20(1). 1–29. DOI: 10.1075/intp.00001.ple.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/.

Role, François & Mohamed Nadif. 2011. Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011)*, 218–223. Scitepress. DOI: 10.5220/0003655102260231.

Rychlỳ, Pavel. 2007. Manatee/Bonito-A Modular Corpus Manager. In Petr Sojka & Aleš Horák (eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, 65–70. Brno: Masarykova Univerzita.

Schäfer, Martin & Melanie J. Bell. 2020. Constituent polysemy and interpretational diversity in attested English novel compounds. *The Mental Lexicon* 15(1). 42–61. DOI: 10.1075/ml.00013.sch.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: Journal des traducteurs* 43(4). 486–493. DOI: 10.7202/004136ar.

Shlesinger, Miriam & Brenda Malkiel. 2005. Comparing modalities: Cognates as a case in point. *Across Languages and Cultures* 6(2). 173–193.

Tirkkonen-Condit, Sonja. 2004. Unique items-over-or under-represented in translated language? In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist*, vol. 48 (Benjamins Translation Library), 177–186. Amsterdam: Benjamins.

Vandevoorde, Lore. 2020. *Semantic differences in translation : Exploring the field of inchoativity* (Translation and Multilingual Natural Language Processing 13). Berlin: Language Science Press. DOI: 10.5281/zenodo.2573677.

Chapter 6

Cohesion through the lens of EPTIC-SI: Sentence-initial connectors in interpreted, translated and non-mediated Slovene

Tamara Mikolič Južnič[a] & Agnes Pisanski Peterlin[a]
[a]University of Ljubljana

Due to a lack of appropriate resources, few studies are devoted to comparing linguistic characteristics across different modes of production (speech and writing). This paper focuses on contrasting the use of sentence-initial connectors in mediated spoken and written Slovene and non-mediated spoken and written Slovene, by comparing EPTIC-SI, two monolingual reference corpora of Slovene, GOS for spoken and KRES for written discourse, and a subsection of a comparable Slovene corpus of parliamentary discourse, siParl. The EPTIC corpus and its subcorpus for Slovene, EPTIC-SI, are intermodal compilations of European Parliament speeches, their verbatim reports, interpretation transcripts and verbatim reports translations. This structure allows for direct comparison of the same content in different modes of production; however, the current size and monolithic genre of the corpus would make generalizations unreliable. For this reason, reference corpora of spoken and written discourse were used to complement the EPTIC corpus. The results show notable differences between the two modes of production, and at the same time reveal other influencing factors, such as genre and mediation.

1 Introduction

Traditional linguistic research on Slovene has focused above all on the standard written variety of the language, which means that there is much less data available for other varieties. This can present a challenge because, due to historical

Tamara Mikolič Južnič & Agnes Pisanski Peterlin. 2022. Cohesion through the lens of EPTIC-SI: Sentence-initial connectors in interpreted, translated and non-mediated Slovene. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 155–182. Berlin: Language Science Press. DOI: ??

circumstances, there is a considerable gap between spoken Slovene and the standard written variety in terms of phonology, grammar and discourse. However, in recent decades, there has been increasing interest in compiling various kinds of corpora for Slovene to allow researchers an insight into different types of actual language use. At present, the majority of corpus resources available for Slovene focus on single modes of production, i.e., written texts (e.g., Gigafida,[1] KRES[2]), web discourse (e.g., Janes[3]) and spoken discourse (e.g., GOS[4]). While the compilation of new corpora has fostered a number of recent studies on non-standard and spoken Slovene (for instance, Fišer et al. (2020), Verdonik (2015)), there has been less research interest in comparing the linguistic characteristics of Slovene across different language varieties or modes of production. One of the reasons for this may be that such comparisons are often difficult to carry out because the complex differences in content, genre, length, context, participants etc. make direct contrasting of different types of materials challenging.

The present study attempts to address this gap. EPTIC-SI, the Slovene component of the EPTIC corpus, is used as a common platform for comparing the two modalities. A key advantage of EPTIC-SI is that it contains a spoken (interpreting) and a written (translation) version of the same content, which allows a direct comparison using a novel approach. Thus, data from EPTIC-SI can help us shed light on how the written and spoken modalities of Slovene follow distinct discourse patterns. At present, a downside of EPTIC-SI is that it is a small corpus, further limited by the fact that it contains a single, monolithic discourse genre. As a result, it is difficult if not impossible to generalize any findings based solely on its analysis. This means that complementing EPTIC-SI research with additional data from larger corpora helps increase the reliability and validity of the results.

This paper thus focuses on spoken and written varieties of mediated and non-mediated Slovene by comparing EPTIC-SI, two monolingual reference corpora of Slovene, as well as a Slovene corpus of parliamentary debates, siParl.[5] Specifically, we investigate variation in the use of sentence-initial[6] connectors, which

[1] http://www.gigafida.net/

[2] http://www.korpus-kres.net/

[3] http://nl.ijs.si/janes/o-projektu/korpus-janes/

[4] http://www.korpus-gos.net/

[5] https://www.clarin.si/noske/run.cgi/corp_info?corpname=siparl20&struct_attr_stats=1

[6] The term *sentence-initial* is used in this paper to refer to both written and spoken discourse, although *utterance-initial* or, in the case of dialogue, also *turn-initial* would be the appropriate terms for spoken discourse. A single term is used to simplify the comparison and also because of the transcription conventions used in EPTIC-SI Int, where utterances are transcribed as sentences.

constitute an important class of cohesive devices. We hypothesise that:

- There is a difference between the use of sentence-initial connectors in interpreting and translation in the EPTIC-SI corpus.
- The use of sentence-initial connectors in interpreted Slovene in EPTIC-SI is similar to their use in spoken Slovene.
- The use of sentence-initial connectors in translated texts in EPTIC-SI is similar to their use in written Slovene.

The article is structured as follows: in §2 the compilation and the main characteristics of the new, Slovene component of EPTIC, EPTIC-SI, are presented. §3 is dedicated to a brief description of sentence-initial connectors to shed light on the topic of investigation. §4 presents a complete overview of the corpora and methods used. The results are presented and discussed in §5, followed by a brief conclusion in §6

2 EPTIC-SI

EPTIC-SI is the Slovene component of the multilingual, parallel intermodal corpus known as EPTIC, or the European Parliament Translation and Interpreting Corpus,[7] comprising speeches delivered at the EU Parliament, their interpretations and translations (see Bernardini et al. 2016 for a more detailed description). The ongoing EPTIC project was first developed at the University of Bologna in collaboration with several other universities. As of 2020, the EPTIC corpus includes English, Italian, French, Slovene and Polish texts. With its intermodal and multilingual design, the EPTIC corpus fosters a range of different research perspectives, involving interpreting and translation and different types of comparisons of the different combinations of subcorpora. In addition, EPTIC allows the juxtaposition of interpretations and translations of the same content, facilitating a unique perspective on the differences between the two related yet divergent processes of interlingual communication.

At present, the Slovene language component of the EPTIC corpus, EPTIC-SI, is a collection of EU Parliament speeches, interpreted and translated into Slovene. Preselected speeches originally delivered in English on 17 January 2011 were used as source texts; the preselected speeches are the same speeches that are used in other parts of the EPTIC corpus. EPTIC-SI was compiled by a project team from

[7] https://corpora.dipintra.it/eptic/

the University of Ljubljana (UL) and UL MA-level students, consisting of Tamara Mikolič Južnič, Lia Lampe, Ana Podobnik, Polona Polc, Anina Stopinšek, Tamara Šiljak and Agnes Pisanski Peterlin. The preparation included the transcription of 64 selected speeches interpreted in Slovene and the preparation of metadata of both the transcripts and corresponding verbatim[8] reports translated into Slovene. In the narrow sense EPTIC-SI thus consists of two subcorpora: 64 transcriptions of interpreted speeches in Slovene, and 64 written Slovene translations of English verbatim reports. For the purposes of comparison, two subcorpora comprising the corresponding source texts are also used (i.e., 64 transcriptions of original speeches delivered in English and 64 corresponding verbatim reports in English). The total number of texts in the four subcorpora is hence 256 and the total number of words is 76,445. All the components of EPTIC-SI have been aligned at sentence level and time-aligned with the video recordings in complete accordance with the EPTIC guidelines, and the data, along with the standardized metadata, is available from the main EPTIC webpage. In January 2020, the EPTIC-SI development began its second stage, with new materials being compiled to be added to the corpus by the end of the year.

3 Review of the literature

3.1 Sentence-initial connectors

Sentence-initial connectors have a significant role in text organization as they are used to link units of text; in fact, the importance of inter-sentential linking in establishing cohesion has long been recognized (see Halliday & Hasan 1976). There is no single agreed upon characterization or framework of connectors (see Halliday & Hasan 1976, van Dijk 1977, Fraser 1999), and there is considerable variation in terminology[9] (see, for instance, Crawford Camiciottoli 2010: 651). Nonetheless, it is generally agreed upon that they constitute a functional category (see Becher 2011: 30), which can be realized through a range of different linguistic elements (see Crawford Camiciottoli 2010: 650), including conjunctions (*and*), adverbs (*however*) and even phrases (*as a result*).

[8] What is important to note is that, as Bernardini et al. (2016: 68) underline, even though they are called verbatims, "these reports are substantially edited" and may actually differ considerably from the transcripts, a fact, which must be taken into account when comparing the interpreted and the translated versions of the same speech (for a detailed account, see Bernardini et al. (2016: 62–70)).

[9] Terms such as connectives, discourse markers, pragmatic markers and similar are used for this functional category by different authors.

In recent years, a substantial body of corpus-based studies has provided novel insight into the function of connectors using authentic language. While some of these studies focused on the complexities of corpus annotation (e.g., **RehbeinEtAl2016**, **Crible2017**, **Crible2018**, **Crible2017**, **Crible2020**; **CribleCuenca2017**; **CriblePascual2020**), contributing importantly to identifying discourse-pragmatic phenomena in large collections of authentic texts, other studies explored the role of connectors in establishing cohesion in a text by investigating variation across languages, registers and discourse modes (**Lapshinova-KoltunskiKunz2014**; **KunzLapshinova-Koltunski2014**, **KunzLapshinova-Koltunski2015**, **KunzLapshinova-Koltunski2014**; **Carrió-Pastor2013**).

Much research attention has focused on both intra-sentential as well as inter-sentential connectors, but the distinct discourse functions of sentence-initial connectors have also been highlighted and investigated (cf. van Dijk 1977; **Moreno1995**; **Dupont2018**). As **Moreno1995** argues, sentence-initial connectors, functioning at the level of discourse, play a prominent role in the macrostructuring of the text. Several empirical studies on sentence-initial connectors, most notably Biber et al. (1999), have revealed a complex array of similarities and differences in terms of their use across spoken and written genres. The function of turn-initial connectors in spontaneous speech may, at first glance, appear quite distinct from the function of sentence-initial connectors in formal writing. However, as Dorgeloh (2004) shows, important parallels between the interactive discourse of spoken dialogue and writing can be established even for *and*, a connector that is far more frequent in speech than writing. Biber et al. (1999: 83–84) make a direct comparison of the frequency of use of selected "coordinators in sentence/turn-initial position" (*and, but, or* and *nor*) in conversation, fiction, news reportage and academic prose, revealing that they occur far more frequently in conversation than in any of the written registers, and that they occur least frequently in academic prose. Biber et al. (1999: 84) suggest that the somewhat more frequent use of coordinators in sentence-initial position in literature and news reportage may result from the fact that these two registers contain more dialogue. Biber et al. (1999: 83) also point out that there is "a well-known prescriptive reaction against beginning an orthographic sentence with a coordinator" (see also Dorgeloh 2004 and Bell 2007 for more information on the proscription against the sentence-initial *and* and *but* in English writing).

Finally, as EPTIC-SI comprises interpreted and translated discourse, the impact of mediation on connector use should also be considered, above all in terms of two mediation-related phenomena: transfer and explicitation. Transfer is the potential impact of source language conventions as reflected in the source texts on the target texts. Important cross-linguistic differences in the use of connectors have been identified for various pairs of languages in contrastive studies (e.g., Pit

2007; **Lapshinova-KoltunskiKunz2014**, **KunzLapshinova-Koltunski2014**, Balažic Bulc & Gorjanc 2015), including Slovene and English (Pisanski Peterlin 2015). The use of sentence-initial connectors in mediated discourse, including the texts in EPTIC-SI, may, in some cases, be influenced by transfer. Explicitation, i.e., the tendency of the target text to be more explicit than the source text, may result in an increased number of connectors in mediated texts. Musacchio & Palumbo (2010: 2) argue that "/c/onnectives are a good indicator of tendencies towards explicitation in translation as they can often be seen as optional elements." Furthermore, Gumul (2006) shows the importance and frequency of explicitation in the form of connectors in simultaneous interpreting, where they are largely a subconsciously added item in an automated mediation process. Based on the analysis of French-to-English and French-to-Dutch interpretations and translations of EU parliamentary speeches, Defrancq et al. (2015) confirm that interpreters add connective items for different reasons, including explicitation.

3.2 Hybrid speech-writing modes

ChafeTannen1987 underline that the difference between speech and writing began to receive research attention relatively late, as traditionally linguistics attempted to describe written language. The emergent focus on the differences between speech and writing has also contributed to a growing awareness that the relationship between the two modalities is not necessarily dichotomous. As **ChafeTannen1987** argue, "there is no single feature or dimension that distinguishes all of speaking from all of writing". In this context, **Wikström2017** highlights the contribution of the so-called "continuum" models, which "suggest that if particular registers such as everyday conversation and academic prose are taken as constituting poles of 'maximum' spokenness and writtenness respectively, most registers and genres of spoken and written discourse actually fall somewhere in between those poles as regards any given linguistic feature or discourse characteristic".

Outlining the fluid orality-literacy osmosis from a historical perspective, **Soffer2020** touches upon the concept of secondary orality brought about by the advent of electronic media. He argues that"/i/n the electronic media age that followed print, texts are written to be read aloud". This type of blending of the two modalities is found in a range of hybrid genres, such as written-to-be spoken discourse (e.g., pre-scripted speeches or television programmes), discourse spoken for transcription (e.g., medical dictation, intralingual live subtitling), digital Internet discourse (e.g., comments sections, tweeting) and mediated discourse (e.g., sight translation, interlingual subtitling). The blending is reflected in an array of linguistic

features, ranging from lexical choice to syntactic complexity (see **Wikström2017** for a detailed overview of the differences between the two modalities).

The intermodality of EPTIC provides a valuable insight into a genre that displays hybrid features of both spoken and written mode; as many of the speeches were pre-scripted, all were subsequently also transcribed as verbatim reports and underwent a substantial amount of editing.

4 Corpora and procedure

4.1 Corpora

In addition to the EPTIC-SI corpus, which comprises texts in both the spoken and written mode and is outlined in §2, two large reference corpora of Slovene comprised of written and spoken genres, as well as a comparable Slovene corpus of parliamentary debates were used in the study. These corpora were selected to enable a comparison between original and mediated texts in both modalities.

As described above, the EPTIC-SI corpus comprises EU Parliament speeches, and consists of transcripts of the Slovene interpretations of original English speeches and written translations of the English verbatim reports of the very same speeches.[10] While the original English speeches were not analysed in terms of connector use themselves, they were used to resolve any ambiguities about the function of a connector in the Slovene versions (only inter-sentential function was considered), as well as to shed light on whether the differences between interpretations and translations can be explained by the differences in the source texts. Table 1 summarises the statistical data for EPTIC-SI.

Table 1: Subcorpora and statistics for the EPTIC-SI corpus.

(Sub)Corpora	No. of words
EPTIC-SI English Spoken Sources (EPTIC SS)	21,561
EPTIC-SI English Verbatim Reports Sources (EPTIC VR)	20,552
EPTIC-SI Interpreting transcripts (EPTIC-SI Int)	16,143
EPTIC-SI Translated verbatim reports (EPTIC-SI Trans)	18,189
Total	76,445

[10]The speeches and the interpretations were thus produced before the verbatims and their translation.

The reference corpus of written Slovene is the KRES corpus, the 100-million word reference corpus sampled on the (much larger) Gigafida corpus. Since Gigafida, at present the biggest corpus of the Slovene language, is composed largely of newspaper and magazine articles, KRES was designed to be its balanced counterpart, in which various written genres are represented to reflect the actual ratio of different genres encountered in the everyday life by an average Slovene reader. The texts collected in the corpus were published between 1990 and 2011, and the samples of texts included in the corpus were chosen randomly (see Logar Berginc et al. 2012 for details). The taxonomy and statistics of the corpus are given in Table 2.

Table 2: Structure and statistics of the KRES corpus.

Subcorpora	**No. of words**
Printed publications	79,830,144
• Books	35,088,699
• Literature	17,030,038
• Non-fiction	18,058,661
• Periodicals	39,727,038
• Newspapers	19,919,327
• Magazines	19,807,912
Miscellaneous	5,014,206
Internet	20,001,001
• News portals	8,000,131
• Companies and institutions	12,000,870
Total	99,831,145

As can be seen from Table 2, KRES consists of 6 subcorpora: Literature, Newspapers, Magazines, Internet, Non-fiction (mainly specialized texts) and Miscellaneous (Misc.). The vast majority of texts are written in standard Slovene, though some of the subcorpora may contain texts with elements of spoken language, displaying elements of hybridity (for instance, the Literature subcorpus in some of the dialogues or the comments which are part of the Internet subcorpus). For the present study, all the subcorpora of KRES were analysed, as they represent a range of relevant genres. Further refinement of genre/source selection within each subcorpus would have been useful, but the online concordancer for KRES does not enable for it.

The reference corpus of spoken Slovene used in the study is the GOS corpus (see Verdonik et al. 2013 for more detailed descriptions). GOS includes around 120 hours of speech, transcribed in two versions (pronunciation-based and standardized), which are linked to the corresponding audio files. Samples of spoken Slovene were collected from all the regions of Slovenia between 2004 and 2010. In total, it contains around 1 million words, and it is, to date, the only reference corpus of spoken Slovene. The structure and statistics of the corpus are presented in Table 3.

Table 3: Structure and statistics of the GOS corpus.

Subcorpora	**No. of words**
Public	583,666
• Informative and educational	353,144
• Television	104,030
• Radio	95,117
• Personal contact	153,997
• Entertainment	230,522
• Television	104,955
• Radio	125,567
Non-public	451,435
• Non-private	155,893
• Telephone	33,862
• Personal contact	122,031
• Private	295,542
• Telephone	69,012
• Personal contact	226,530
Total	1,035,101

Table 3 shows that GOS comprises 4 subcorpora, but for the purposes of the present analysis, only two were used: the Public informative and educational subcorpus (henceforth Info-Ed) and the Non-Public Private subcorpus (henceforth Private). The two subcorpora were chosen because they represent two very distinct types of spoken language. The Info-Ed subcorpus comprises fairly formal spoken discourse that has often been pre-scripted or pre-prepared to some extent. Specifically, public informative discourse covers media discourse (i.e., television and radio news), while public educational discourse encompasses lectures (e.g., in

secondary schools and universities). The Private subcorpus represents the other end of the spoken continuum as it comprises spontaneous speech from private contexts, that is spontaneous conversation among family, friends and similar. While this is quite distinct from the genre of EPTIC-SI, it provides a valuable insight into the range of differences in Slovene spoken discourse.

As none of the genres in the reference corpora are directly comparable to the genre of the texts in EPTIC-SI, a set of texts from a comparable corpus of parliamentary discourse in Slovene, siParl[11] , was also analysed. SiParl is a 200-million word corpus comprising transcriptions of parliamentary debates of the Slovene National Assembly (see **PancurErjavec2020** for details). SiParl includes different types of debates, such as regular sessions, urgent sessions, sessions of individual working bodies of the assembly, etc., with texts spanning three decades (1999–2018). During this period Slovene society underwent a profound transition which may also be reflected in discourse characteristics. A small, relatively homogenous subsection of the corpus was carefully selected for a close comparison with EPTIC-SI. The 283,908-word subsection was limited to the genre of public presentation of opinions (henceforth Opinions), which is comparable to the genre of EPTIC-SI, and to the year 2011, also corresponding to the time-frame of EPTIC-SI. In making the selection, comparability was prioritised over size, with the restricted size of the subsection making manual analysis feasible.

4.2 Procedure

The criteria used to define sentence-initial connectors in this study were both formal (sentence initial position) as well as functional (discourse cohesive function). Halliday & Hasan (1976) identify four main types of conjunctive cohesion: additive, adversative, causal and temporal; in the present study, our analysis is limited to the first three categories, i.e., additive, adversative and causal.

For the purposes of corpus analysis, a list of 7 Slovene connectors was drafted for each of the three categories (see Appendix A). These lists were prepared in three steps. As relatively little data is available for Slovene on the linguistic items that can function as connectors and may appear in sentence-initial position, the first versions of the lists were compiled using several different sources. These included Toporišič's (2004: 646–652) list of intra-sentential coordinate conjunctions, Pisanski Peterlin's (2015) study of sentence-initial adversative connectors, Balažic Bulc & Gorjanc's (2015) study of the position of connectors and Hirci & Mikolič Južnič's (2014) study of causal connectors. The initial lists were further

[11] https://www.clarin.si/repository/xmlui/handle/11356/1236

expanded in the second step using the Slovene thesaurus function of Microsoft Word. The last step involved editing the list to retain only those connectors that unambiguously occur in intra-sentential function when used in the initial position. This was done because the size of the KRES corpus makes it impossible to manually examine all the results.

The searches were carried out automatically by means of the web concordancer for KRES, NoSketch Engine for GOS and siParl, and AntConc (Anthony 2020) for EPTIC-SI. The frequency counts were normalized to their rate of occurrence per 1000 words. For KRES, siParl and EPTIC-SI, where standard punctuation was used, determining the beginning of the sentence was not problematic. In GOS, double slashes marking the end of an utterance or a turn were used to identify utterance-initial or turn-initial connectors, which were considered to be the equivalents of sentence-initial connectors in spoken discourse (see Dorgeloh 2004, for arguments supporting the comparability of sentence-initial connector use in speech and writing).

Next, all the selected sentence-initial connectors identified in EPTIC-SI, siParl and GOS were examined manually to remove any false results, i.e., cases in which the items from the search list had other functions. Such cases were extremely rare for additive and adversative connectors (only one such case was found in EPTIC-SI, with a total of 7 in siParl and 8 in GOS), and fairly rare for causal connectors (only 6 such cases were found in EPTIC-SI and a total of 88 in siParl and 142 in GOS).[12] For KRES, manual cleaning was not feasible because of the corpus size (100 million) and the total number of concordances found (123,165). As a result, the figures for KRES are unrevised. However, if we assume that the percentage of false results is at least similar (and probably lower) to that in GOS, then the figures in KRES for causal connectors, the category where false results were the most common, probably contain somewhere around 3.65% false results.

The results for the different subcorpora were compared in terms of their overall frequencies, their frequencies for the different types of sentence-initial connectors and the frequencies of the individual connectors.

Finally, the results for the two subcorpora of EPTIC-SI were compared using NoSketch Engine available from the EPTIC website, where the parallel aligned versions are available, to establish the differences and similarities between the interpreted and translated versions. The corresponding transcriptions of the original English speeches and the English verbatims were also consulted when necessary as described in §4.1.

[12] The notable difference in size between the Slovene part of EPTIC-SI on the one hand, and siParl and GOS on the other must be taken into consideration when interpreting these figures.

5 Results and discussion

The normalized quantitative results of the analysis of all the subcorpora are presented in Figure 1 below. The results are first presented as a total figure for each corpus and then separately by subcorpora.

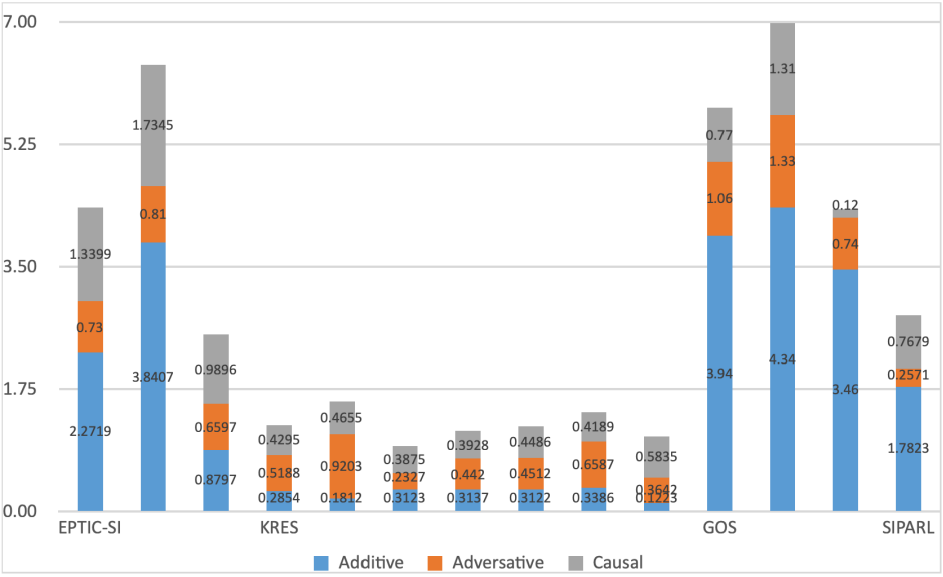


Figure 1: Occurrences of sentence-initial connectors in the analysed corpora.

The ratios of the three categories of sentence-initial connectors – additive, adversative and causal – are given for the individual subcorpora in Figure 2 below.

The results are compared and discussed in more detail in §5.1–5.3.

5.1 Sentence-initial connectors in EPTIC-SI Int and EPTIC-SI Trans

The first hypothesis examined in this paper is that there is a difference between the use of sentence-initial connectors in EPTIC-SI Int and EPTIC-SI Trans. The quantitative results of the corpus analysis are given in Table 4.

The comparison of the interpreting and translation subcorpora of EPTIC-SI reveals a substantial difference in the frequency of use of sentence-initial connectors between the two subcorpora with the ratio being approximately 2.5:1, which confirms the first hypothesis. A juxtaposition of the three categories of connectors reveals that this difference is largely due to additive connectors, which occur

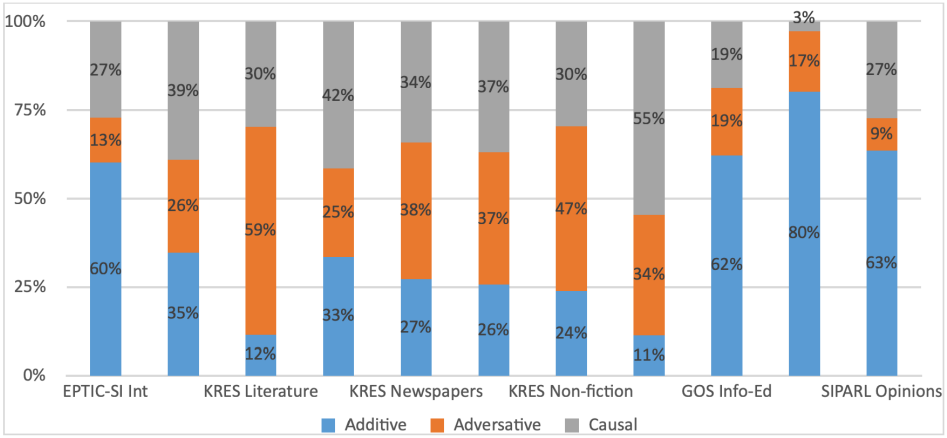


Figure 2: Ratios of the three types of sentence-initial connectors in individual subcorpora.

Table 4: Occurrences of sentence-initial connectors in the EPTIC-SI corpus

	Total EPTIC-SI		EPTIC-SI Int		EPTIC-SI Trans	
	Raw	/1k	Raw	/1k	Raw	/1k
Additive	78	2,27	62	3,84	16	0,88
Adversative	25	0,73	13	0,81	12	0,66
Causal	46	1,34	28	1,73	18	0,99
Total	149	4,34	103	6,38	46	2,53

four times as frequently in EPTIC-SI Int as in EPTIC-SI Trans. The difference in frequency is far less marked for adversative connectors that are used with almost the same frequency in both subcorpora. Finally, causal connectors are used almost twice as frequently in interpreting as in translation.

A more detailed focus on **additive connectors** shows that the marked difference is due to the use of a single connector, the sentence-initial *in* [and], which accounts for as many as 52 of the 62 additive connectors occurring in EPTIC-SI Int; only three other connectors, *poleg tega* [in addition] occurring 7 times, *prav tako* [additionally] occurring twice and *hkrati* [simlutanously] occurring once, are found in EPTIC-SI Int. In EPTIC-SI Trans, the most frequent additive connector is *poleg tega* [in addition] occurring in 8 cases, but other additive connectors are used rarely: *in* [and] in three instances, *obenem* [at the same time]

twice, *prav tako* [additionally] twice and *ob tem* [at that] once. The preference for some of these connectors is closely linked to the register, as some connectors are very formal and associated with standard written texts, while others are more often used in sentence-initial position in informal contexts. However, as Dorgeloh (2004) argues, parallels between the discourse functions in speech and writing can be observed even in the case of sentence-initial *and*, which is far more frequent speech than writing.

When the results for the two subcorpora of EPTIC-SI are compared directly using the aligned versions on the EPTIC webpage, only three cases where there are matching additive connectors in both subcorpora in corresponding passages are identified. A detailed look at the individual examples reveals that there are several other instances of matching sentence-initial additive connectors that cannot be identified automatically for various reasons, such as the use of a filler, *ehm*, immediately preceding the additive connector, but formally occurring in sentence-initial position (2 such cases), or the use of a less common sentence-initial connector not on the list used in corpus search (as in example (1)). But in the majority of cases, the manual check confirms that there are no matching sentence-initial connectors. In some of these instances, an intra-sentential additive connector is used in the corresponding passage in the other corpus, as in example (2). In other cases, no corresponding cohesive device can be identified.

(1) a. EPTIC-SI Int: *Istočasno* pa je pomembno poudariti tudi, da je Evropska unija eden največjih trgov za tropski les.[13]
[*Simultaneously*, it is important to stress that the European Union is one of the biggest markets for tropical wood.]

b. EPTIC-SI Trans: *Obenem* je tudi zelo pomembno, da poudarimo, da je EU eden izmed največjih trgov tropskega lesa.
[*At the same time*, it is very important for us to stress that the EU is one of the biggest markets for tropical wood.]

(2) a. EPTIC-SI Int: Mislim, da je to tudi eden od pomembni-, gre le za enega od kazalnikov, ampak če pogledamo celoto, zagotovo lahko

[13] Throughout the text, the following markings are used for the examples from EPTIC-SI : a. for transcriptions of the Slovene interpretations of English speeches and b. for Slovene translations of the Englsih verbatims. An English gloss, as literal as possible, is provided for all the Slovene examples in square brackets. Where necessary, c. for transcriptions of original English speeches and d. for English verbatims are added. In the examples, the relevant connectors have been highlighted in italics by the authors.

govorimo o spodbudnih dogodkih. *In* edini način, da podpremo takšen proces, je da delamo skupaj z njimi …

[I think that this is one of the importa-, it is one of the indicators, but if we look at the whole, we can certainly speak of encouraging events. *And* the only way for us to support such a process is to work together with them …]

b. EPTIC-SI Trans: Razumem, da je to samo en kazalnik, a na splošno so bile novice vzpodbudne, proces *pa* lahko izboljšamo samo, če bomo sodelovali.

[I understand that this is only one indicator, but in general there has been encouraging news, *and* we can only improve the process by collaboration.]

There seem to be two main, often interrelated reasons for these omissions. The first is the register, or more specifically, the degree of formality. As certain additive connectors, above all *in* [and], are associated with speech and informal discourse, and are rarely used in formal, edited writing, it is not surprising that there are considerable dissimilarities in this area between the two subcorpora (example (1) illustrates such a difference in formality). The second reason is linked to the English originals. It is important to bear in mind that the interpretations and the translations are obtained using related but different source texts (see §2 and Bernardini et al. 2016: 68): in spite of their name, the verbatim reports are heavily edited and diverge from the transcriptions of the speeches in terms of register and wording. As there is a strong proscription against using sentence-initial *and* in English (see Biber et al. 1999, Dorgeloh 2004, Bell 2007), it is not surprising that this is one of the features in which the source transcriptions and the verbatims in English differ greatly. Example (3) illustrates the difference between the two English versions, as well as the difference between interpreting and translation.

(3) a. EPTIC-SI Int: *In* še to za konec. Zelo hvaležna sem, da sem danes lahko predstavljala Evropsko komisijo pri tej točki. Podpredsednici Redingovi bom sporočila vse, kar ste povedali, tudi nekatera zastavljena vprašanja, vprašanje poslanca, kjer se pričakuje odgovor …

[*And* to finish. I am very grateful that I have been able to represent the European Commission on this topic today. I will convey to Vice-President Reding all of what you have said, including some of

the questions, the question raised by an MEP where an answer is expected ...]

b. EPTIC-SI Trans: Podpredsednici Reding bom tudi prenesla vse, kar je bilo povedano nocoj, vključno z vprašanjem, ki ga je postavil eden izmed poslancev in pri katerem se pričakuje odgovor.
[I will convey to Vice-President Reding all of what has been said today, including the question posed by one of the MEPs where an answer is expected.]

c. EPTIC SS: *And* my fifth and final point is that I'm very grateful that I have been here on behalf of the Commission this evening. I will convey to Vice-President Reding the points that have been made, including a question that has been raised here by one of the MEPs that that an answer is expected.

d. EPTIC VR: Finally, I will convey to Vice-President Reding the points that have been made here this evening, including the question raised by one member in relation to which an answer is expected.

At first glance, the comparison of **adversative connectors** reveals surprising similarities between examples in EPTIC-SI Int and EPTIC-SI Trans: 12 instances of the adversative connector *vendar* [however] occur in each subcorpus; in addition, there is only a single instance of another adversative connector *po drugi strani* [on the other hand] in the interpreting subcorpus. Nevertheless, a juxtaposition of the two sets of examples shows, somewhat unexpectedly, that there are only two matching expression of *vendar* [however] in the two subcorpora. An examination of the remaining instances of *vendar* [however] in both subcorpora reveals that, for most of them, markers signalling adversative relations can be found in the corresponding passages of the translations and interpreted speeches. However, these markers are not identified through corpus search for several reasons: a) they are not used in sentence-initial position, b) they are not typical adversative connectors and are therefore not on the list of sentence-initial connectors used in this study, c) they may express adversative relations, but when used in sentence-initial position, they typically do not function as adversative connectors and are therefore not on the list used in corpus search. In about one third of the cases, no corresponding adversative marker can be identified in the parallel subcorpus. As in the case of additive connections, this often occurs when there is already a discrepancy between the transcription of the original English speech and the English verbatim, as in example (4) below.

(4) a. EPTIC-SI Int: *Vendar* pa dolgoročen cilj humanitarne pomoči ni ehm to.
[*But* the long-term goal of the humanitarian aid is not ehm that.]

b. EPTIC-SI Trans: Humanitarna pomoč pa seveda ni pravi instrument, ki bi imel dolgoročen vpliv.
[Humanitarian aid of course is not the right instrument that would have a long-lasting impact].

c. EPTIC SS: Ehm *but*, for long-lasting impact, humanitarian aid of course is not the instrument.

d. EPTIC VR: Of course, for a long-lasting impact, humanitarian aid is not the right instrument.

The omission of *but* in the verbatim can very likely be attributed to the proscriptions against using sentence-initial *but* in writing in English (cf. Bell 2007: 183); as Bell (2007: 194) points out this proscription is far less strong than the proscription against sentence-initial *and*, but it nevertheless needs to be taken into account. While there are no such restrictions against using *vendar* [however] in initial position in written Slovene, the fact that the Slovene translation is based on the English verbatim necessarily means that some of the adversative connectors are not found in the translations.

As with the other two categories, there is relatively little variety in **causal connectors**. Only three such connectors occur in the interpreting subcorpus: zato [therefore] in 16 cases, *torej* [thus] in 7 cases and *zaradi tega* [because of that] in 5 cases, with 28 cases all together. In the translation subcorpus, all 18 causal connectors are instances of *zato* [therefore].

A close comparison of the results of the two subcorpora reveals that there are five matching causal connectors occurring in corresponding passages in both interpretations and translations. In several other cases, markers of causal or resultative relations can be found in the corresponding passages, often in the form of a clause, as in example (5). It seems that this reflects the complexity of the cause-effect relation which, unlike the additive meaning, tends to be overtly expressed.

(5) a. EPTIC-SI Int: *To je tudi razlog*, zakaj predvidevamo finančno pomoč za izboljšanje trgovskih zmogljivosti …
[*This is also the reason* why we expect financial aid for enhancing trade capacity …]

b. EPTIC-SI Trans: *Zato* je tu tudi finančna pomoč, ki bo okrepila trgovinsko zmogljivost.
[*Therefore,* financial aid is available to enhance trade capacity.]

Another interesting observation concerns the question of sentence boundaries and the parallels between intra-sentential and inter-sentential expressions of causality. It is noteworthy that when it comes to causal connectors, there are several instances where sentence boundaries diverge considerably between the interpreted speeches and the corresponding translations. In such cases, a sentence-initial causal connector would have a corresponding intra-sentential cause-result connector, as in example (6).

(6) a. EPTIC-SI Int: Želim odkrit razgovor z vami, sicer bom …*Torej* ehm vi ste v glavnem govorili tudi v angleščini, zato bom tudi jaz govoril v angleščini. Rekli ste, da naj si pogledamo …
[I wish to speak openly with your, otherwise I will …*So* ehm you have been mainly speaking in English, so I will speak in English as well. You have said that we should take a look …]

b. EPTIC-SI Trans: Besedilo imam v portugalščini, vendar bom improviziral v angleščini, *saj* ste v delu svojega govora, ki je bil po mojem mnenju najpomembnejši, uporabili ravno ta jezik …
[My text is in Portuguese, but I will improvise in English, *since* you have used this language in the part of your speech that I consider to be the most important part …]

Finally, a single case of a sentence-initial causal connector in EPTIC-SI Trans and a corresponding passage in EPTIC-SI Int with a combination of a sentence-initial additive connector *in* [and] immediately followed by a causal connector was found through corpus search (see example (7)). Once again, this type of difference clearly illustrates the disparity between less formal, more loosely organized spoken discourse (the metadata confirms that the speech in question is an impromptu speech), and structured, edited, written text.

(7) a. EPTIC-SI Int: *In zato* je treba pozdraviti z vsem srcem takšen sporazum in upam, da se bo tudi izvajal, kajti če se ne bo izvajal, bo škoda papirja, na katerem je napisan.
[*And therefore* this agreement should be welcomed wholeheartedly and I hope that it will be implemented, because if it isn't, it will not be worth the paper it is written on.]

b. EPTIC-SI Trans: *Zato* je ta sporazum treba pozdraviti odprtih rok in upam, da se bo tudi izvajal, kajti če se ne bo, potem ne bo vreden papirja, na katerem je napisan.
[*Therefore*, this agreement should be welcomed enthusiastically, and I hope that it will be implemented because if it is not, it will not be worth the paper it is written on.]

5.2 Sentence-initial connectors in interpreted and spoken Slovene

The second hypothesis tested was that the use of sentence-initial connectors in EPTIC-SI Int is similar to their use in spoken Slovene in GOS and siParl. The quantitative results are given in Table 5 and Figure 3.

Table 5: Occurrences of sentence-initial connectors in GOS, siParl and EPTIC-SI Int

	GOS						EPTIC-SI		SIPARL	
	Info-Ed		Private		Total		EPTIC-SI Int		Opinions	
	Raw	/1k	Raw	/1k	Raw	/1k	Raw	/1k	Raw	/1k
Additive	1533	4,34	1023	3,46	2556	3,94	62	3,84	506	1,78
Adversative	469	1,33	220	0,74	689	1,06	13	0,81	73	0,26
Causal	462	1,31	35	0,12	497	0,77	28	1,73	218	0,77
Total	2464	6,98	1278	4,32	3742	5,77	103	6,38	797	2,81

Two subcorpora of GOS were used in the present study. A comparison of the frequency of sentence-initial connectors in EPTIC-SI Int and in GOS (Total) shows considerable similarities. The frequencies of sentence-initial connectors in the comparable texts, siParl Opinions (public presentation of opinions from 2011), on the other hand, are considerably lower compared to both EPTIC-SI Int, as well as GOS and its subcorpora. However, as Figure 1 shows, the frequencies in siParl Opinions are still much higher than in all the written subcorpora of KRES, but only marginally higher than in EPTIC-SI Trans.

A closer look at the ratios of the three types of connectors for written and spoken corpora in Figure 2 reveals a clearer distinction between speech and writing in terms of sentence-initial cohesive devices.

Figure 3 reveals interesting distinctions between speech and writing. While sentence-initial **additive** connectors constitute the most frequently used category of connectors in all spoken subcorpora, this is not the case in the written texts of EPTIC-SI Trans and KRES, where causal and adversative connectors play

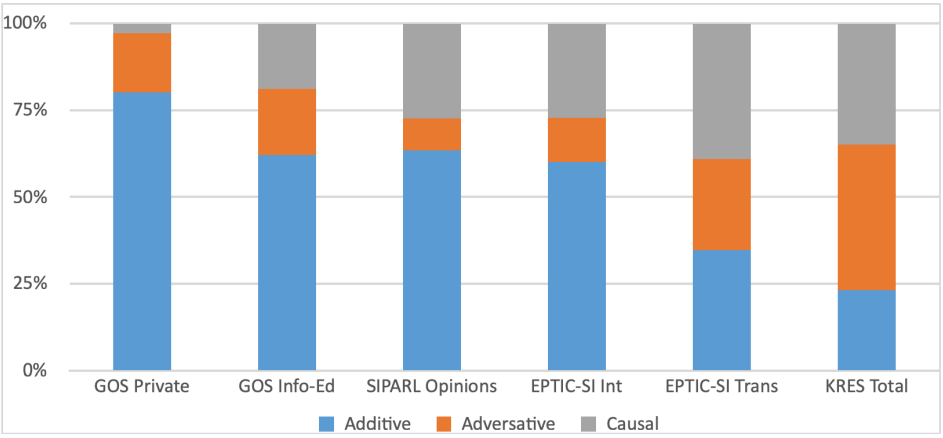


Figure 3: Ratios of the three types of sentence-initial connectors in spoken and written discourse.

a greater role in establishing inter-sentential cohesion. Moreover, the ratios in three of the spoken subcorpora, GOS Info-Ed, siParl Opinions and EPTIC-SI Int, are far more similar than in the fourth spoken subcorpus, GOS Private. This very likely reflects the fact that the Private subcorpus of GOS contains casual spontaneous conversation (see example (8)), an informal dialogical genre quite distinct from the content of EPTIC-SI Int. The discourse of the Info-Ed bears closer similarity to the genre of EPTIC-SI (see example (9)). The discourse of siParl Opinions (see example (10)) is, of course, most comparable to that of EPTIC-SI Int, as both include structured, pre-prepared formal and monological genre of parliamentary speeches. Nevertheless, the comparison with GOS Private offers important insight into commonalities across a range of varieties of spoken discourse compared to written texts.

(8) GOS Private: //*in* kaj je narobe z njimi? / sandale // tiščijo me ona ma bl mičkano nogu ku jst // kdu? / *in* kaj pa če bi mi jih meni dala ? // [name:personal] // ja pomir si // [gap] // ja točnu tud ti pomir si // sej ne vem kire si mela

[//*and* what's wrong with them?/ sandals // they are too tight her feet are smaller than mine // who? / *and* what if you gave them me to me? // [name:personal] // well try them on // [gap] // well sure you try them on, too // I don't know which ones you had][14]

[14]The annotations used in GOS include pauses, gaps, utterance beginnings/endings, etc. As noted

(9) GOS Info-Ed: // eee mislim da teh upov ni več eee vlada je na današnji seji sprejela sklepe s katerimi je dala soglasje za odprtje oziroma zaprtje štirih poglavij / *in hkrati* dala soglasje oziroma ne izdala soglasja za izd [gap] odprtje sedmih poglavij
[// erm I think that these hopes are long gone erm in today's cabinet meeting the government has passed agreements with which it gave its approval for the ope [gap] opening or closing of four chapters / *and at the same time* it gave its approval or denied its approval for the opening of seven chapters]

(10) Ustavite ga, tudi vi, gospod državni tožilec. Hkrati naj na koncu opozorim še na eno zadevo, ki se danes dogaja še vedno, mislim, da ni nobenih sprememb po prihodu novega generalnega državnega tožilca.
[Stop him, you too, Mr. Public Prosecutor. At the same time let me point out another matter that is still happening today, I believe there have been no changes after the arrival of a new general public prosecutor]

The relatively frequent use of **causal connectors** in the EPTIC-SI Int subcorpus might be explained by the fact that the genre of EU Parliament speeches is generally argumentative in nature and tends to use causal connectors as means of building arguments (cf. Didriksen & Gjesdal 2013). In siParl Opinions, the overall use of causal connectors is much lower than in EPTIC-SI Int; nevertheless, causal connectors constitute one quarter of sentence-initial connectors in both EPTIC-SI Int and siParl Opinions, underlying the argumentative character of parliamentary speeches. Interestingly, a comparatively high frequency of causal connectors occurs in the Info-Ed subcorpus of GOS, especially compared to the Private subcorpus, though the more diverse nature of the individual genres of Info-Ed (news reports, lectures), which may be more or less argumentative, probably accounts for the somewhat lower frequency of causal connectors than in EPTIC-SI Int. Moreover, connector use may be more frequent in interpreted texts due to explicitation and transfer (see §3), although a comparison with the corresponding original English texts, which is beyond the scope of the present paper, would be necessary to provide insight into translation-related phenomena.

The second hypothesis was thus partly confirmed: in terms of ratios, the results show a distinct cline with an overwhelming reliance on additive connectors in non-mediated spontaneous speech, and a more even distribution of the types

in §4, utterance beginings/endings and turn taking in dialogue are marked with double slashes, while pauses are marked using single slashes. However, it is essential to bear in mind that determining utterance boundaries is not as clear-cut as establishing sentence boundaries.

of connectors in non-mediated writing. Although the frequencies of sentence-initial connectors also showed some degree of similarity among the spoken subcorpora, the tendencies are somewhat less homogenous.

5.3 Sentence-initial connectors in translated and written Slovene

The third hypothesis, that the use of sentence-initial connectors in EPTIC-SI Trans is similar to their frequency in written Slovene in KRES, is based on the assumption that the translated verbatim reports in EPTIC-SI follow the norms of written Slovene. As Table 6 shows, the quantitative results of our analysis support the third hypothesis only partially.

Table 6: Occurrences of sentence-initial connectors in the KRES subcorpora and in EPTIC-SI Trans

			Additive	Adversative	Causal	Total
KRES	**Literature**	**Raw**	3086	15673	7927	26686
		/1k	0.18	0.92	0.47	1.57
	Internet	**Raw**	6246	4655	7751	18652
		/1k	0.31	0.23	0.39	0.93
	Newspapers	**Raw**	6249	8805	7825	22879
		/1k	0.31	0.44	0.39	1.15
	Magazines	**Raw**	6185	8938	8886	24009
		/1k	0.31	0.45	0.45	1.21
	Non-fiction	**Raw**	6114	11895	7565	25574
		/1k	0.34	0.66	0.42	1.42
	Misc.	**Raw**	613	1826	2926	5365
		/1k	0.12	0.36	0.58	1.07
	Total KRES	**Raw**	28493	51792	42880	123165
		/1k	0.29	0.52	0.43	1.23
EPTIC-SI	**EPTIC-SI Trans**	**Raw**	16	12	18	46
		/1k	0.88	0.66	0.99	2.53

As noted in the Introduction, there is a substantial divergence between spoken and written genres in Slovene. The corpus data for KRES and GOS (see Figure 1) very much reflect this divergence between the two modalities, as the sentence-connectors analysed here occur far more frequently in spoken discourse. However, the comparison of the frequency of sentence-initial connectors in EPTIC-SI Trans and their overall frequency in KRES also shows a prominent difference: sentence-initial connectors are used twice as frequently in EPTIC-SI Trans as

in KRES. A more detailed look at the categories of sentence-initial connectors shows that all are used less frequently in KRES, with the difference being particularly noticeable for additive and causal connectors.

The more frequent use of sentence-initial connectors in EPTIC-SI Trans may result from the hybrid nature of the source texts, i.e., the verbatim reports, which are based on speeches. As they are written to be delivered in the spoken mode, they share the characteristics of both written and spoken discourse.

A comparison with the individual subcorpora of KRES shows the same tendencies for the categories of **additive** and **causal** connectors and for the total number of connectors in each subcorpus. **Adversative** connectors, on the other hand, reveal a different picture: they are actually used more frequently in the Literature subcorpus of KRES (see example (10)) than in EPTIC-SI Trans, while their frequency is exactly the same in the Non-fiction subcorpus and in EPTIC-SI Trans. Of all the subcorpora of KRES, sentence-initial connectors are used most frequently in the Literature subcorpus, possibly reflecting the imitations of speech (dialogue) found in literature, as shown in example (11) (see also Biber et al. (1999: 84) for similar findings).

(11) Saj ni nič posebnega,« je priznal. »*Toda* nikamor drugam te ne morem odpeljati
[It's nothing special," he admitted. "*However*, I can't take you anywere else]

(12) »Moj palček? *In* ...kako veš, da sem mislila, da si pikapolonica?«
["My gnome? *And* ... how do you know that I thought you were a ladybird?"]

To sum up, due to its hybrid nature, EPTIC-SI Trans exhibits a frequency of sentence-initial connectors that is quite different from the spoken genres analysed as well as from the written genres in KRES, albeit the results are somewhat closer to those of the KRES corpus, compared to spoken genres. However, in addition to the influence of genres outlined above, another potential reason for the relatively high frequency of sentence-initial connectors in the EPTIC-SI Trans corpus should be considered. As it contains mediated discourse, explicitation of cohesive links as well as transfer from the source texts may well have contributed to the fairly frequent use of sentence-initial connectors in EPTIC-SI Trans.

6 Conclusion

The aim of the present study was to contrast the use of sentence-initial connectors, an important category of cohesive devices, both in spoken and written Slovene as well as in mediated and non-mediated discourse. Using EPTIC-SI, two large reference corpora for Slovene and a subsection of a comparable Slovene corpus of parliamentary discourse, we have shown that patterns of use of sentence-initial connectors reflect important differences for both dimensions, modality and mediation, thus substantiating the potential of this type of corpus research. The expected difference between mediated spoken and written discourse in the first hypothesis was confirmed, but the second and third hypotheses were only partly confirmed. For spoken non-mediated and mediated discourse, the results show a greater complexity, as the similarities depend on the type of connector. The written mediated discourse of EPTIC-SI Trans appears to display hybrid characteristics of both spoken and written discourse.

The EPTIC corpus offers a unique perspective on different modes of interlingual mediation and the complexities of language use, as it provides the same content in two different modalities and multiple languages. For Slovene as a peripheral language, the contribution of EPTIC-SI is particularly valuable because it enables us to directly observe and reflect on the differences between the same content worded in speech and writing, opening a range of additional research paradigms. We believe that the present study corroborates the multidimensional investigation potential of EPTIC and EPTIC-SI, providing insight into the intricacies of language reality.

Finally, the specific characteristics of language identified in EPTIC-SI may also shed light on other important issues in future research. The varieties of languages evolving in EU contexts, shaped by a variety of factors, including language mediation, have already been recognized as distinct forms of language production for other languages, most notably English (see, for instance, Trebits 2009, whose study focuses on the use of conjunctive cohesion in EU documents in English). However, the specific features of Slovene as used in EU contexts have not yet received systematic research attention; in fact, there seems to be little research awareness of new patterns developing in administrative and public discourse in Slovene as a result of the language contact in EU institutions. It therefore seems that as EPTIC-SI is gradually expanded, also to include original Slovene speeches delivered at the EU Parliament, it will offer an invaluable resource for studying this emerging new variety of Slovene.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0218 and No. P6-0215).

Appendix A List of sentence-initial connectors used in the corpus search

- Additive connectors:
  - In
  - Hkrati
  - Obenem
  - Ob tem
  - Poleg tega
  - Prav tako
  - Podobno
- Adversative connectors:
  - Na drugi strani
  - Nasprotno
  - Po drugi strani
  - Toda
  - Vendar
  - Vendarle
  - V nasprotju
- Cause-result connectors:
  - Kot posledica
  - Posledično
  - Torej
  - Zaradi tega
  - Zategadelj
  - Zato
  - Zatorej

References

Anthony, Laurence. 2020. *AntConc*. http://www.laurenceanthony.net/software/antconc/ (28 July, 2020).

Becher, Viktor. 2011. When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies* 23(1). 26–47. DOI: 10.1075/target.23.1.02bec.

Bell, David M. 2007. Sentence-initial And and But in academic writing. *Pragmatics* 17(2). 183–201. DOI: 10.1075/prag.17.2.01bel.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC |Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of written and spoken English*. Harlow: Longman.

Balažic Bulc, Tatjana & Vojko Gorjanc. 2015. The position of connectors in Slovene and Croatian student academic writing: A corpus-based approach. In Sonja Starc, Carys Jones & Arianna Maiorani (eds.), *Meaning making in text: Multimodal and multilingual functional perspectives*, 51–71. London: Palgrave Macmillan UK. DOI: 10.1057/9781137477309_4.

Crawford Camiciottoli, Belinda. 2010. Discourse connectives in genres of financial disclosure: Earnings presentations vs. earnings releases. *Journal of Pragmatics* 42(3). 650–663. DOI: 10.1016/j.pragma.2009.07.007.

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies*, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.

Didriksen, Anders Alvsåker & Anje Müller Gjesdal. 2013. On what is not said and who said it: Argumentative connectives in Nicolas Sarkozy's speeches to the European Parliament. In Kjersti Fløttum (ed.), *Speaking of Europe. Approaches to complexity in European political discourse*, 85–110. Amsterdam/Philadelphia: John Benjamins.

Dorgeloh, Heidrun. 2004. Conjunction in sentence and discourse: Sentence-initial and and discourse structure. *Journal of Pragmatics*. Pragmatics of Discourse 36(10). 1761–1779. DOI: 10.1016/j.pragma.2004.04.004.

Fišer, Darja, Nikola Ljubešić & Tomaž Erjavec. 2020. The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation* 54(1). 223–246. DOI: 10.1007/s10579-018-9425-z.

Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics*. Pragmatics: The Loaded Discipline? 31(7). 931–952. DOI: 10.1016/S0378-2166(98)00101-5.

Gumul, Ewa. 2006. Explicitation in Simultaneous Interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures* 7(2). 171–190. DOI: 10.1556/Acr.7.2006.2.2.

Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London & New York: Routledge.

Hirci, Nataša & Tamara Mikolič Južnič. 2014. Korpusna raziskava rabe vzročnih in pojasnjevalnih povezovalcev v prevodih iz angleščine in italijanščine. In Agnes Pisanski Peterlin & Schlamberger Brezar, Tamara (eds.), *Prevodoslovno usmerjene kontrastivne študije*, 150–70. Ljubljana: Znanstvena založba Filozofske fakultete.

Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt & Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: Gradnja, vsebina, uporaba*. Ljubljana: Trojina, Zavod za uporabno slovenistiko.

Musacchio, Maria Teresa & Giuseppe Palumbo. 2010. Following norms, taking risks: A study of the use of connectives in a corpus of translated economics articles in Italian. In Carmen Heine & Jan Engberg (eds.), *Reconceptualizing LSP: Online proceedings of the XVII European LSP symposium*, 1–11.

Pisanski Peterlin, Agnes. 2015. Sentence-initial adversative connectives in Slovene-English translation of academic discourse: A corpus study. In Mojca Schlamberger Brezar, Limon, David & Gruntar Jermol, Ada (eds.), *Contrastive analysis in discourse studies and translation.* 68–82. Ljubljana: Znanstvena založba Filozofske fakultete.

Pit, Mirna. 2007. Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast* 7(1). 53–82. DOI: 10.1075/lic.7.1.04pit.

Toporišič, Jože. 2004. *Slovenska slovnica [Slovenian grammar]*. Maribor: Založba Obzorja.

Trebits, Anna. 2009. Conjunctive cohesion in English language EU documents – A corpus-based analysis and its implications. *English for Specific Purposes* 28(3). 199–210. DOI: 10.1016/j.esp.2009.04.004.

van Dijk, Teun A. 1977. *Text and context explorations in the semantics and pragmatics of discourse*. New York: Longman.

Verdonik, Darinka. 2015. Internal variety in the use of Slovene general extenders in different spoken discourse settings. *International Journal of Corpus Linguistics* 20(4). 445–468. DOI: 10.1075/ijcl.20.4.02ver.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek & Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation* 47(4). 1031–1048. DOI: 10.1007/s10579-013-9216-5.

Chapter 7

Exploring linguistic variation in mediated discourse: translation vs. interpreting

Heike Przybyl[a], Alina Karakanta[b], Katrin Menzel[a] & Elke Teich[a]

[a]Saarland University [b]Fondazione Bruno Kessler / University of Trento

This paper focuses on the distinctive features of translated and interpreted texts in specific language combinations as forms of mediated discourse at the European Parliament. We aim to contribute to the long line of research on the specific properties of translation/interpreting. Specifically, we are interested in mediation effects (translation vs. interpreting) vs. effects of discourse mode (written vs. spoken). We propose a data-driven, exploratory approach to detecting and evaluating linguistic features as typical of translation/interpreting. Our approach utilizes simple word-based n-gram language models combined with the information-theoretic measure of relative entropy, a standard measure of similarity/difference between probability distributions, applied here as a method of corpus comparison. Comparing translation and interpreting (including the relation to their originals), we confirm the previously observed overall trend of written vs. spoken mode being strongly reflected in translation and interpreting output. In addition, we detect some new features, such as a tendency towards more general lexemes in the verbal domain in interpreting or features of nominal style in translation.

1 Introduction

We present the results of a corpus-based analysis of translations, interpreting, and comparable original written and spoken texts – four modes that are habitually produced and consumed in the domain of the European Parliament. The

Heike Przybyl, Alina Karakanta, Katrin Menzel & Elke Teich. 2022. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 183–208. Berlin: Language Science Press. DOI: ??

overarching goal of the paper is to contribute to a more nuanced understanding of the characteristics of translated and interpreted language and to the empirical foundations of theories of mediated discourse. Specifically, we are interested in the following main questions: How can we investigate linguistic differences in interpreted and translated language compared to each other and to non-mediated language? If there are differences, on which linguistic levels do they manifest themselves? Focusing on the target languages English and German, two rather closely related languages from a historical point of view but with important structural differences, we ask more specifically whether interpreting generally is more similar to spoken non-mediated (i.e. original) discourse than to written translations as suggested by Shlesinger & Ordan 2012 in their experimental and corpus-based studies for mediated texts. We may assume that simultaneous interpreting is first and foremost a form of speech with distinct features due to the cognitive complexity involved in listening, analysis, language transfer, production and articulation and not essentially the same as written translation, although both tasks involve language mediation.

We pursue a data-driven, exploratory approach using techniques from computational language modeling combined with a more hypothesis-driven micro-analysis. We employ word-based unigram language models and relative entropy (Kullback-Leibler Divergence; KLD) as a measure of the similarity/difference between modes and for highlighting the lexico-grammatical items typical of translation/interpreting that warrant deeper linguistic analysis. For inspection, we use a word-cloud visualization of the words detected as typical by KLD, where 'typical' is a gradient notion. From the highly typical words we engineer more complex features that undergo further analysis. For example, among the highly typical items for translations are definite determiners. This is an indication of a more pronounced nominal style in translations compared to written originals, so we further inspect nominal use. For interpreting, we find, for instance, that it is more varied in the use of verbs, including auxiliaries, so we inspect verbal use further (see §5).

The remainder of the paper is structured as follows. In §2 we discuss related work and show the benefits of relative entropy being used for comparative, corpus-based studies. §3 gives information on the corpora used and explains the KLD approach. This is followed by detailed descriptions of the KLD results, comparing written translations with simultaneous interpreting, but also translations to written originals and interpreted speech to spoken originals in order to observe if the features shown are typical for mediated discourse or rather distinctive for the written or spoken mode (§4). Features highly typical of interpreting/translation

as shown by relative entropy are then analysed in more detail (§5). §6 concludes the paper with a brief summary and outlook.

2 Background and related work

A long-standing question in translation studies is whether translations have specific linguistic properties in common which distinguish them from comparable original texts. These are linguistic effects of the translation process found in the translation product labelled as "translationese" (written translation) or "interpretese" (oral translation/interpreting). Effects have been categorised as simplification, explicitation, normalization, shining through etc. (Baker 1993, Laviosa 1998, Teich 2003). Some scholars have referred to the specific effects of translation as "translation universals", trying to relate them to the way in which translators process the source text (S-universals) and the way in which translators use the target language (T-universals) (Chesterman 2004: 39). The term "translationese" may seem to have become slightly outmoded to some translation scholars after divergent and sceptical views on the existence of translation universals or on the lack of sound methods to investigate this phenomenon have been expressed, e.g. by Becher (2010) and House (2008). However, the research community has been left to take up the challenge of revising this framework and gathering suitable data, methods and empirical evidence for or against its assumptions (cf. Vandevoorde (2020: 22ff) on a recent discussion on this still unresolved debate and Oakes (2021) for a discussion of various sets of statistical methods that have been used in the study of translation corpora for the identification of the characteristics of translationese). Despite a rich body of research on written translations (Hansen-Schirra et al. 2013, Lapshinova-Koltunski 2015, Evert & Neumann 2017) and some studies on the spoken mode (Sandrelli & Bendazzoli 2005, Kajzer-Wietrzny 2012, Shlesinger & Ordan 2012, Bernardini et al. 2016, He et al. 2016, Dayter 2018) a unifying explanation of the observed effects is still lacking.

Due to the availability of interpreting data in large enough quantity, the majority of corpus-based interpreting studies of recent years has been based on political discourse studied on European Parliament data (EPIC: Bendazzoli & Sandrelli 2005, Monti et al. 2005, Sandrelli & Bendazzoli 2005, Sandrelli et al. 2010, Russo et al. 2012, Bernardini et al. 2016; EPICG: Defrancq 2018, Plevoets & Defrancq 2018; TIC: Kajzer-Wietrzny 2012, 2015) or discourse within the United Nations (SIREN: Dayter 2018). Our study adds a recently compiled, relatively large dataset of transcribed material, enriched with relevant metadata for the language pair German-English to the investigation of European Parliament discourse. Most relevant to

our work are studies on EPTIC (Bernardini et al. 2016, Ferraresi et al. 2018) and TIC (Kajzer-Wietrzny 2012, 2015) as some components of the data used overlap. Bernardini et al. (2016) studied simplification via lexical density, mean sentence length, core vocabulary coverage and list head coverage in EPTIC, an intermodal, comparable and parallel European Parliament corpus for English-Italian. Comparing SI (simultaneous interpreting) with TR (translations) they find that SI is simplified regarding lexical density and larger use of frequent words. They also find SI simpler compared to spoken originals on the lexical (list head coverage and core vocabulary) and syntactic level (shorter sentences) and see this trend also for TR vs. written originals, but not as strong as for the spoken comparison: "Simplification thus appears to be both a feature of orality and a feature of mediation, such that interpreted texts, being both spoken and mediated, occupy one extreme of the simplicity cline, whose other extreme is occupied by written non translated texts." (Bernardini et al. 2016: 81). They also observe differences between the languages studied for some of the parameters.

In previous studies on EPIC (the spoken part of EPTIC, including not only English and Italian, but also Spanish) Russo et al. (2012) also report a tendency to higher lexical density in interpreted speech than in original spoken, but with some exceptions to this trend. This trend is opposite to previous findings for translations (Laviosa 1998). Kajzer-Wietrzny (2012) does not observe greater simplification in interpreting vs. spoken originals, studying English original spoken and simultaneous interpreting into English from different source languages, regarding core vocabulary and lexical density, only with respect to analysing list heads. Especially for lexical density, the languages studied, either as target or source language, seem to influence the result to a large extent. Dayter (2018) looks at the language pair English-Russian and finds simplification for SI into Russian (with lower lexical density and use of more high frequency words in SI than in originals). For English, she observes the opposite: higher lexical density and more variation in SI (with Russian as source) - also contrary to results for the English corpora in EPIC (with Italian and Spanish as source). Furthermore, Dayter (2018) also finds SI into Russian more explicit than original spoken Russian (higher proportion of nominal to pronominal reference) and again, the opposite for SI into English, which is less explicit than original English for SIREN.

Explicitation and normalisation have also been studied for TIC. Kajzer-Wietrzny (2012) confirms the universal of TR being more explicit than comparable originals for her written dataset. However, the spoken part shows mixed results. For syntactic explicitness SI behaves like TR (higher use of optional connectives following reporting verb), but no general pattern in SI was observed for linking adverbials as another factor of explicitness. The normalisation universal was also

only confirmed by one parameter studied: SI tends to normalise like TR concerning lexical bundles, but not for the use of fixed phrases.

Thus the overall picture by using traditional measures does not show a clear trend towards simplification, explicitation and normalisation in simultaneous interpreting. The languages involved (source and target languages) seem to have an influence. However, it might well be the case that the features found to describe universals for written translations are not suitable for interpreted speech. He et al. (2016) use a data driven, comparative approach. Using text classification they find segmentation (e.g. via use of coordinating conjunctions, explicitly ”and”) as a distinctive interpretese feature for the language pair Japanese-English. This and the trend of generalisation in SI they observe can be linked to the translation universal of simplification. Repetition of content words, which they find distinctive for SI, could be an indication of explicitation. In line with the traditional translationese results are also their findings that ”that” seems to be characteristic for translations, which again, can be linked to explicitation.

In this study we also pursue an exploratory approach detecting distinctive features in a data-driven fashion. Patterns in the features detected can then provide an empirical basis for further interpretation, be it as effects of some underlying translation/interpreting-specific processing or as (reinforced) effects of oral vs. written production, as discussed in Shlesinger & Ordan (2012). In her earlier work Shlesinger (1989) also found an equalising effect on oral and literate features of source speeches: orally marked source speeches seem to become more literate SI output, source speeches with more distinct written features become more oral. As our dataset includes read out speeches that were prepared by members of the European Parliament beforehand, we assume that the source speeches contain some markers of writtenness. We build on these findings and ask specifically whether the features we detect can be interpreted as effects of mediation (translation/interpreting) or rather of discourse mode (written/oral production).

Regarding the proposed method of exploratory analysis, we draw on the recent experiences in using relative entropy to capture linguistic variation across relevant variables such as time, register, style or gender in linguistic as well as humanistic research. For example, Degaetano-Ortlieb & Teich (2019) apply the asymmetric variant of relative entropy, Kullback-Leibler Divergence, as a technique to characterize the course of diachronic change and the features involved in late modern English science writing. Klingenstein et al. (2014) apply the symmetric variant of relative entropy, Jensen-Shannon Divergence, to the speaking styles in criminal trials comparing violent with nonviolent offenses and Degaetano-Ortlieb (2018) compares the speaking styles of men and women in the same corpus of historical English court proceedings. In our own work

in translation studies, we have described the basic workings of the approach in Karakanta et al. (forthcoming) and discussed the benefits of an information-theoretic perspective of translation more broadly in Teich et al. (2020). Compared to more traditional methods in corpus linguistics, our approach based on relative entropy has the advantage of being data-driven, thus helping to avoid prior selection of (potentially irrelevant) features. Second, no separate significance testing is needed – rather, significance testing is built into the procedure. This facilitates feature selection and feature evaluation and thus provides a more objective procedure and easier to interpret results.

3 Data and method

As our dataset we use European Parliament speeches: for translation, we use the Europarl-UdS corpus (Karakanta et al. 2018)[1] containing written originals (ORG WR EN and ORG WR DE) and translations for English and German (TR DE EN for translations into English with German as source language and TR EN DE for translations into German with English as source language). The source for these written originals is a spoken event in the European Parliament which was then subsequently adapted to fulfil written conversions, i.e. false starts are left out and only complete sentences are published (cf. Bernardini et al. 2016). Translations are produced from these written originals. Both written originals and translations were used as published by the European Parliament to compile the written component of our dataset. For interpreting, we use selected English material from existing European Parliament interpreting/intermodal interpreting-translation corpora (TIC: Kajzer-Wietrzny 2012, EPICG: Defrancq et al. 2015) for English spoken originals (ORG SP EN) and simultaneous interpreting from German into English (SI DE EN). For these existing English datasets we added transcriptions of the German original speeches (for existing SI DE EN) and simultaneous interpreting into German (for existing ORG SP EN). The spoken data, referred to as EPIC-UdS, were transcribed or revised according to transcription guidelines based on EPICG (Bernardini et al. 2018) ensuring comparability across the different datasets. The spoken transcripts include typical characteristics of spoken language such as false starts, hesitations and truncated words. All data were enriched with relevant metadata such as source language, original speaker as well as speech timing, mode of delivery and speech rate for the spoken part.

We build probabilistic unigram language models of the source and target languages for interpreting and translation and calculate the relative entropy be-

[1] [1] http://fedora.clarin-d.uni-saarland.de/europarl-uds/

Table 1: Corpus overview: Europarl-UdS (written) and EPIC-UdS (spoken).

Europarl-UdS			EPIC-UdS		
	sentences	words		sentences	words
TR EN DE	137,813	3,100,647	SI EN DE	4,080	58,218
TR DE EN	262,904	6,260,869	SI DE EN	3,622	59,100
ORG WR DE	427,779	7,869,289	ORG SP DE	3,408	57,049
ORG WR EN	372,5470	8,693,135	ORG SP EN	3,623	68,548

tween the distributions obtained using KLD. The KLD between distribution P and distribution Q estimates the amount of additional bits of information needed to model interpreting by translation (and vice versa) or translation/interpreting by original text. This gives us an indication not only of how different translation and interpreting outputs are overall compared to one another and compared to originals (by the KLD score between the distributions) but also of the linguistic features (here: words) that contribute most to the difference, namely the words with the highest KLD score (Fankhauser et al. 2014). On the basis of a word-cloud visualization, we explore the words that are the strongest signals of variation by relative frequency and highest distinctivity (cf. Karakanta et al. forthcoming). The word clouds serve as an intuitive visual abstraction and provide a valuable starting point for further analysis. The distributions shown in the word clouds are subject to a t-test, all the results discussed in the following having a p-value of 0.05 or lower. We show the usefulness of our KLD-based approach in detecting and analysing variation among forms of mediated discourse by confirming the observations through a more detailed corpus analysis. To this aim we compute Standardised Type-Token Ratio (STTR), lexical density (the amount of lexical words divided by the total number of words), mean token length and carry out a part-of-speech distribution as well as pattern analysis.

4 KLD analysis: Simultaneous interpreting (SI) vs. translation (TR)

In a first step, we contrast interpreting with translation for German and English as target languages (and English and German as source languages, respectively). Consider the KLD visualization in Figure 1 for German.

(a) Interpreting (b) Translation

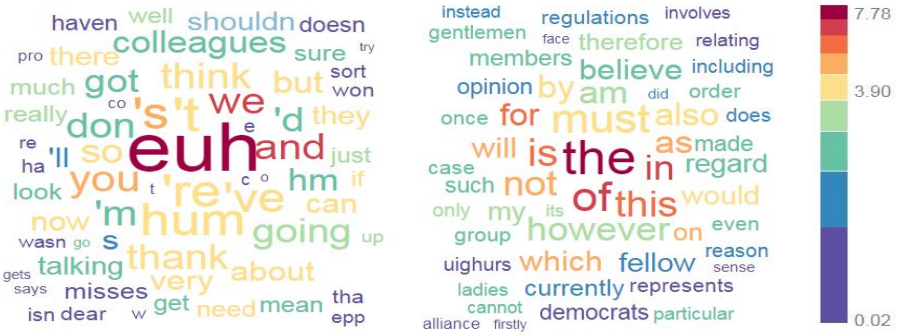
Figure 1: Variation in translation mode with German as target and English as source language. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

The KLD visualisation shows typical words for (a) interpreting (left) and (b) translation (right). The *size* of words displayed marks their distinctivity, i.e. their KLD score; *colour* represents the relative frequency of a word. Highly frequent words are visualised in red, low relative frequency is marked blue.[2] From Figure 1 we can observe that overall, interpreting exhibits more highly distinctive items than translation. The words shown for interpreting are mainly function words as well as very few but highly frequent general verbs (*haben, geben, sagen, gehen*). Closer inspection confirms that well-known features of spoken discourse appear as strikingly typical for German interpreted texts, such as hesitation markers (*euh, hum, hm*), particles, discourse markers and intensifiers (e.g. *also, ja, sehr, ganz, so*), deictics (*jetzt, hier*) and reduced forms (*hab, ne, n*). Conjunctions also seem to be more characteristic for interpreting, especially those marking parataxis (*und, aber, denn, da*). Written translations, instead, prefer the more formal *jedoch* (equivalent to *aber* in interpreting) and prepositions (*in, auf, mit, für, zu, von*). Written translations are also characterised by a more nominal style indicated by various determiners and pronouns (*der, die, dieser, diesem, ihre, seine, unser, meiner* etc.) and by more content words shown to be distinctive for translations (e.g. *Bericht, Parlament, Ansicht, Präsident, vergangenen*), however at a

[2]In our exploratory analysis we did not want to bias the results by manipulating the data severely by excluding selected parts-of-speech, e.g. by excluding content or function words. However, we also considered separate types of analyses, e.g. by masking functions words, nouns or cultural-specific items. It would go beyond the scope of this paper to also cover these different other options systematically.

much lower level and, as expected, with lower frequencies. Note that the words which are typical for translation are generally longer.

The KLD visualisation for English (Figure 2) shows a similar result: fewer and only general lexical items for interpreting. Instead, function words are most distinctive. More variation in lexical choice is observed in written translations, but their distinctivity is at a low level by KLD values.



(a) Interpreting (b) Translation

Figure 2: Variation in translation mode with English as target and German as source language. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

Like in German, spoken discourse features are the most distinctive features for interpreting: hesitations markers (*euh, hum, hm*), reduced forms, discourse markers (*well, now, so*) and intensifiers (*really, very*. In terms of logical relations, interpreting shows coordinating conjunctions (*and, but*) whereas translations are characterised by prepositions. Interestingly, written translations also use the more formal conjunction *however* (cf. the German *jedoch*) in contrast to *but* (German *aber*) used in interpreting. The contrast here cannot only be observed in style but also as a preference to coordination in spoken (also high relative frequency for *and*) vs. subordination in written. This is in line with the findings of He et al. (2016), who claim that interpreters break longer sentences into multiple smaller chunks, and therefore segmentation is a specific strategy characteristic of interpreting.

Inspection of the KLD visualisations above show that while some differences between SI and TR can be linked to effects of spoken vs. written discourse, other distinctive features do not fall into this explanation. In order to distinguish translationese/interpretese features from differences between the spoken and written mode, we next compare interpreting to spoken originals (§4.1) and translations to written original production (§4.2).

4.1 Spoken: Interpreting vs. originals

The analyses for the spoken mode show that in both languages simultaneous interpreting exhibits more spoken language features than spoken originals (see Figures 3 and 4).

(a) Interpreting (b) Spoken originals

Figure 3: Variation in spoken mode: German simultaneous interpreting vs. spoken originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

(a) Interpreting (b) Spoken originals

Figure 4: Variation in spoken mode: English simultaneous interpreting vs. spoken originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

This includes hesitations (*euh, hm, hum*), intensifiers (German: *so, ganz*; English: *really*) and a more verbal style in SI (German: *müssen, möchten, arbeiten, freuen, geben, sagen, sicherstellen*; English: *be, can, need, talk, gamble, react*). The

verbs used in German and English SI are mostly very general (more specific verbs such as *gamble* and *react* shown for English interpreting (Figure 4) have a low KLD score and are infrequent). Other features characteristic for SI when compared to TR are not distinctive between the interpreting and the spoken originals distributions, i.e. they are features that are prominent in all spoken modes (SI and originals): reduced forms (e.g. contractions, clippings) and an overrepresentation of function words.

Some language differences can also be observed: The two spoken German modes (Figure 3) are characterised by some discourse markers/particles (*ja, also*), deictics (*hier, jetzt*) as well as conjunctions (subordinating and coordinating) whereas, although also characteristic for English when comparing TR and SI, these features do not show when comparing English SI with spoken originals (Figure 4).

Overall, interpreting seems to be more spoken than originals. One explanation could be that, although all of the originals are true transcripts of originals speeches held in the European Parliament, some of the interventions had been prepared by the members of Parliament and therefore typical spoken language features might not be as strong in the spoken originals. SI on the other hand is truly spontaneous spoken production.

4.2 Written: Translation vs. originals

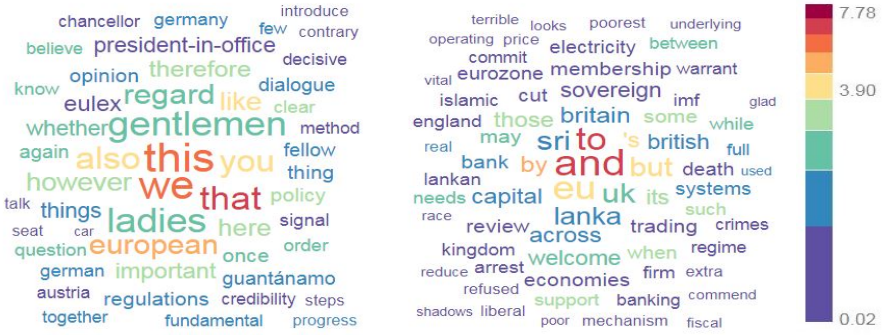
Figures 5 and 6 show the characteristic features for the written mode. Here, the results are less clear and seem to be more language-dependent: translations seem to be more nominal using various determiners (German: *der, die, des, den, seine, ihre, dieser, einer, diese, ihrer, einige, dies*; English: *this, that*). The conjunctions *jedoch* and *however* as a written feature also rank high in translations while written originals use the less formal equivalents *aber* and *but*. Translations therefore tend to be more formal/more written concerning this feature than originals. This might be due to the fact that the written originals have a spoken utterance as a basis and translators normalize to a written standard. For the other features, there does not seem to be a clear uniform trend, e.g. in German prepositions are characteristic for translations whereas for English, prepositions are typical in originals but not in translations.

4.3 Translationese vs. interpretese

In this section we attempt to tell apart purely translationese effects from purely interpretese effects. We take the perspective of TR (once against SI and once

(a) Translation (b) Written originals

Figure 5: Variation in written mode: German translations vs. written originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.



(a) Translation (b) Written originals

Figure 6: Variation in written mode: English translations vs. written originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

against written originals) for translationese (Figure 8) and the perspective of SI (against TR and spoken originals) for interpretese (Figure 7). If features are shown in both contrasts, they can be seen as distinctive of translationese and interpretese respectively.

Overall we can observe that the differences between the written and the spoken mode are greater than between SI or TR compared to the corresponding originals: TR vs. SI show more distinctive items (shown in large font) while at the same time showing more highly frequent items (shown in red and orange) than any other comparison. All models comparing the written or the spoken mode

exhibit many items with low distinctivity (shown in small font) and from lower frequency bands (shown in blue and green).

At the same time, we can observe translationese and interpretese trends: The translation model when comparing with interpreting shows similar features as the translation model when comparing to written originals, although the signal is weaker for TR vs. written originals than for TR vs. SI (same as above: more highly frequent and highly distinctive features for the written-spoken contrast). The translationese/interpretese trends seem to be more pronounced in German. The corresponding English models show a similar but weaker trend.

(a) SI vs. TR (b) SI vs. ORG

Figure 7: Variation in Interpreting: (a) Interpreting modeled on the basis of translation and (b) Interpreting modeled on the basis of spoken originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

5 Corpus analysis based on KLD findings

We have shown that KLD-based analysis brings out intuitively relevant features of mediated discourse and the sometimes subtle distinctions between different types of mediated discourse. In this section, we use the most prominent features detected by KLD-based analysis for engineering more complex features as well as for testing them further by some aggregate measure commonly employed in comparative corpus analysis.

5.1 KLD results in the context of traditional corpus measures

The KLD analysis suggests different degrees of variation in lexical choice between different production modes. Translations were shown to employ greater

(a) TR vs. SI (b) TR vs. ORG

Figure 8: Variation in Translations: (a) Translation modeled on the basis of interpreting and (b) translation modeled on the basis of written originals. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

variation in lexical items whereas fewer words were typical for interpreting. To validate this observation, we employ traditional corpus analysis measures to compute the lexical variation for the different translation modes. In order for our results to be comparable regardless of corpus size, we compute Standardised Type-Token Ratio (STTR). Table 2 shows lexical variation as STTR for the different categories. Significant differences are confirmed by a t-test (EN:t = 36.755, df = 3, p-value = 4.429e-05; DE: t = 25.299, df = 3, p-value = 0.0001354). For both languages, SI has the lowest lexical variation, followed by spoken originals. This result is in line with previous work which found SI to be less varied, more simplified compared to spoken originals. At the same time, both spoken modes are lexically less varied compared to the written modes. Surprisingly, we observe the opposite tendency for the written mode; TR show a higher STTR ratio than written originals, especially for German. This further corroborates our KLD findings suggesting that translations overemphasize features of written mode (here: vocabulary variation).

Table 2: Standardised type-token ratio.

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
STTR	0.40	0.38	0.42	0.43
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
STTR	0.47	0.43	0.49	0.52

The inspection of the KLD models (TR vs. SI) also showed a tendency for shorter words in interpreting overall. At word level, a check of mean token length reveals that SI tends towards using shorter words (see Table 3, EN: t = 99.46, df = 3, p-value = 2.241e-06; DE: t = 62.285, df = 3, p-value = 9.119e-06). Median token length is the same for all modes in both languages, except for SI DE EN (3.0 for SI vs. 4.0 for all other modes). A preference for the use of shorter words is not observed for translations. Thus we can see a tendency towards simplification in SI, but not in TR.

Table 3: Average token length.

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
mean token length	4.32	4.24	4.45	4.36
median token length	4.0	3.0	4.0	4.0
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
mean token length	5.56	5.16	5.35	5.47
median token length	4.0	4.0	4.0	4.0

A further result from the KLD analysis was that the most typical items (highly distinctive and highly frequent words) signal specific choice preferences at the level of parts of speech (POS). (Specific) function words appeared more distinctive for interpreting than for translations, which included more lexical words as distinctive (if at a low KLD and frequency level) compared to interpreting. Lexical density (amount of lexical words divided by the total number of words) is commonly used to measure this contrast. Table 4 shows lexical density for the different categories (EN: t = 104.89, df = 3, p-value = 1.91e-06; DE: t = 74.007, df = 3, p-value = 5.437e-06). For German, both SI and TR are lexically denser than comparable originals. For English, the trend is the opposite. However, as discussed in §2, lexical density often does not give a consistent trend. The method of relative entropy also picks up weaker signals in lexical choice. The contrast between the different modes does not seem to be the choice of lexical vs. content words, but rather the type of lexical item used as for example seen for SI using very general verbs.

To get a better understanding, we look at the distributions of those parts-of-speech (POS) that were highlighted by the KLD analysis: nouns, pronouns, determiners (NOUN, PRON, DET: nominal categories); main verbs, auxiliary verbs and modals (VERB, AUX, MODAL: verbal categories); adpositions, conjunctions (ADP, CONJ: relational categories). In an overall comparison of all subcorpora for each

Table 4: Lexical density.

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
lexical density	43.89	42.67	42.92	41.91
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
lexical density	47.36	47.90	45.09	47.50

target language, including SI and TR, spoken and written originals all show statistically significant differences in the POS distribution by a chi-square test (DE: X-squared = 26662, df = 21, p-value < 2.2e-16; EN: X-squared = 14266, df = 21, p-value < 2.2e-16). Figure 9 plots the part-of-speech distributions in terms of relative frequencies (y-axis shows percentages).

The largest differences in the POS distributions are observed for the nominal (NOUN, PRON, DET) vs. the verbal classes (VERB, AUX, MODAL), where nominal classes are more prominent in the distributions of written texts, while verbal classes for the spoken ones. Determiners are less frequently used in SI, and pronouns seem to be compensating a reduced use of nouns. As in previous works, we observe slightly different tendencies for the different languages. For English, the distributions show a more pronounced effect of spoken vs. written mode, since the distributions are similar for SI and spoken originals together (bars 1 and 2), and for TR and written originals together (bars 3 and 4). For German though, originals (spoken and written) seem to have a more similar distribution to each other (bars 1 and 3) than to their translated and interpreted equivalents. This might suggest stronger translationese and interpretese effects for German. This difference may be an effect of interference from the source language English, or related to linguistic prestige in mediation in the European Parliament.

To gain more information on the structures associated with these POS distributions, we further inspect selected syntactic patterns for nominal, verbal and relational categories.

5.2 Nominal use

Analysis by KLD showed various determiners as highly distinctive items for TR when compared to SI as well as compared to ORG. For German, this included words like *der, die, den, des, dem* which can also be used as relative pronouns. POS analysis is necessary to determine the grammatical function of these words. Furthermore, more nouns and adjectives were seen as typical for translations.

(a) English (b) German

Figure 9: POS distribution for selected POS for English and German.

These features together hint at a more nominal style in TR. To verify this observation, we investigate different noun patterns. When comparing the noun pattern distribution (Figure 10) it becomes clear that - although there is also a difference between the written and spoken modes - simultaneous interpreting behaves more differently than the other categories[3].

For German, spoken and written originals behave similarly (no significant difference)[4], even though within the written and spoken modes significant differences can be observed[5]. However, SI still prefers to opt for a more extensive use of pronouns whereas TR uses determiner-noun combinations instead[6].

The same comparison for English shows significant differences for all categories[7], also showing that SI prefers the use of pronouns more than the other categories.

Table 5 shows frequencies per million (fpm) for the different kinds of noun phrases and confirm that simultaneous interpreting clearly uses less complex patterns. The preference for short encodings is further corroborated by the fact that pronouns are most frequently used in the spoken mode, especially in simultaneous interpreting. Longer determiner-noun combinations are less frequent in

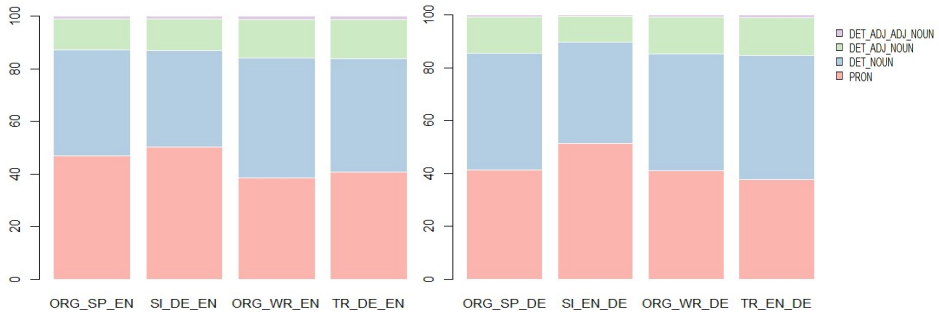
[3]DE: X-squared = 3269.4, df = 9, p-value < 2.2e-16; EN X-squared = 2022.6, df = 9, p-value < 2.2e-16

[4]ORG SP DE vs. ORG WR DE: X-squared = 1.1987, df = 3, p-value = 0.7533

[5]ORG SP DE vs. SI EN DE: X-squared = 248.73, df = 3, p-value < 2.2e-16; ORG WR DE vs. TR EN DE: X-squared = 2625.8, df = 3, p-value < 2.2e-16

[6]SI EN DE vs. TR EN DE: X-squared = 912.86, df = 3, p-value < 2.2e-16

[7]ORG SP EN vs. ORG WR EN: X-squared = 314.39, df = 3, p-value < 2.2e-16; ORG SP EN vs. SI DE EN: X-squared = 248.73, df = 3, p-value < 2.2e-16; ORG WR EN vs. TR DE EN: X-squared = 1381.2, df = 3, p-value < 2.2e-16; SI DE EN vs. TR DE EN: X-squared =303.46, df = 3, p-value < 2.2e-16



(a) English (b) German

Figure 10: Pattern distribution for PRON, DET+NOUN, DET+ADJ+NOUN and DET+ADJ+ADJ+NOUN.

SI than in all of the other modes. Both written modes prefer determiner-noun combinations rather than the use of pronouns. We also observe that the two written modes (original and translation) behave similarly whereas simultaneous interpreting stands out most. Original spoken can be placed in between (use of DET+NOUN combinations similar to the written modes, pronouns in between the frequency figures for SI and written). This might also be due to the fact that some of the original spoken utterances are prepared and read out speeches in the European Parliament. Thus, for SI we can link this preference to simplification and may well assume that the preference for shorter encodings is a mechanisms for reducing processing effort.

A more detailed observation of the most frequently used lexical types in the patterns reveal that SI seems to use more fixed, standardised phrases (e.g. *eine wichtige Rolle* (*an important role*)) does not appear in spoken originals and only rarely in the written data, but quite frequently in SI). The few occurrences in SI of the most complex patterns considered here (DET+ADJ+ADJ+NOUN) seem to be mostly filled by short words and repeating the same adjective (*the last few years/-months/weeks*). At the semantic level, we also see a tendency towards general or collective nouns in SI. However, further analysis is necessary in order to confirm this trend quantitatively.

5.3 Verbal use

A further result of the KLD-based analysis was a distinctive difference in the use of verbs across the different categories. We therefore compare the use of ver-

Table 5: Nominal patterns in fpm.

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
PRON	68,352	65,502	48,822	53,524
DET+NOUN	58,443	48,133	57,705	56,681
DET+ADJ+NOUN	17,274	15,695	18,626	19,834
DET+ADJ+ADJ+NOUN	1,434	1,256	1,512	1,581
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
PRON	80,055	95,553	77,979	74,600
DET+NOUN	85,715	71,295	83,914	93,135
DET+ADJ+NOUN	26,095	18,416	26,225	28,188
DET+ADJ+ADJ+NOUN	1,653	909	1,592	1,818

bal POS categories for the different subcorpora. The distribution of main verbs, auxiliaries and modals shows significant differences between all modes for English[8]. German originals in the written and spoken mode, again, show no significant difference in the use of verbal POS[9] whereas significant differences between other modes can be observed[10]. The normalised frequency distribution (Table 6) confirms that the spoken modes use more verbs than written overall. SI especially stands out by using verbs most frequently and therefore can be seen as being "more spoken than spoken", in line with the findings of Shlesinger & Ordan (2012).

5.4 Relational use: conjunctions

One feature shown as typical for mediated discourse in both languages is the use of *but* and *aber* in SI and *however* and *jedoch* in TR. These conjunctions were shown characteristic for the respective modes when comparing SI to TR, but also - with only one exception for English interpreting - characteristic for SI and TR when comparing to spoken and written originals.

[8] ORG SP EN vs. ORG WR EN: X-squared = 48.11, df = 2, p-value = 3.572e-11; ORG SP EN vs. SI DE EN: X-squared = 34.381, df = 2, p-value = 3.422e-08; ORG WR EN vs. TR DE EN: X-squared = 5318.5, df = 2, p-value < 2.2e-16; SI DE EN vs. TR DE EN: X-squared = 37.131, df = 2, p-value = 8.65e-09

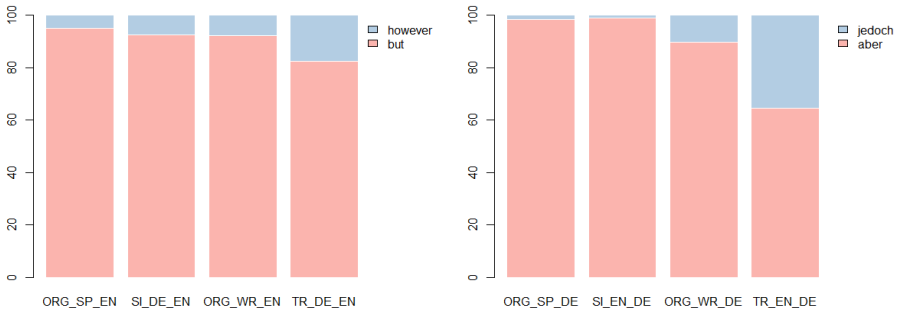
[9] ORG SP DE vs. ORG WR DE: X-squared = 0.29177, df = 2, p-value = 0.8643

[10] ORG SP DE vs. SI EN DE: X-squared = 9.9219, df = 2, p-value = 0.007006; ORG WR DE vs. TR EN DE: X-squared = 1030.4, df = 2, p-value < 2.2e-16; SI EN DE vs. TR EN DE: X-squared = 38.291, df = 2, p-value = 4.843e-09

Table 6: Verbs in fpm.

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
AUX	20,437	22,601	15,563	17,435
MODAL	18,314	22,569	16,758	21,160
VERB	144,393	142,321	133,571	129,944
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
AUX	48,868	53,317	45,680	42,250
MODAL	15,711	19,842	14,390	14,877
VERB	91,158	99,771	84,814	87,639

The distribution for the use of these conjunctions (Figure 11) and Fisher's exact test (due to scarce data points in the spoken data) partly confirm the observation made in the KLD analysis: In the spoken modes, there is no significant difference in the use of these conjunctions[11]. However, translations clearly prefer to use the more formal conjunction *however/jedoch*.[12] This can be seen as normalisation into written mode for translation whereas we might see some spoken influence in the written originals.



(a) English (b) German

Figure 11: Distribution of *but/however* and *aber/jedoch*.

See some examples below from translation and interpreting with their respective originals. The simultaneous interpretation into English keeps *but* as equivalent to the German *aber* in the source, while the English translation opts for the more formal *however* from original *aber*.

[11]ORG SP EN vs. SI DE EN: p-value = 0.1627; ORG SP DE vs. SI EN DE: p-value = 0.7176

[12]ORG WR EN vs. TR DE EN: p-value < 2.2e-16, ORG WR DE vs. TR EN DE: p-value < 2.2e-16

Table 7: aber/jedoch and but/however in fpm

English	ORG SP EN	SI DE EN	ORG WR EN	TR DE EN
but	5299	4701	2614	2806
however	281	386	221	584
German	ORG SP DE	SI EN DE	ORG WR DE	TR EN DE
aber	4057	6214	2796	1773
jedoch	67	65	320	979

Example.
*(1) ORG SP DE: "... es ist gut dass wir ihn haben / **aber** er soll eben Maßnahmen regeln die..."*
*(2) SI DE EN: "...it is very good that we have it / **but** its rule should apply to..."*
*(3) ORG WR DE: "...es ist gut, dass wir ihn haben. **Aber** er soll eben Maßnahmen regeln, die..."*
*TR DE EN: "...it is good that we have it. **However**, its rules should apply to ..."*

6 Conclusion and outlook

We have presented a data-driven, exploratory method for analysis of the typical linguistic features of the modes of communication in a mediated, multilingual setting such as the European Parliament (written vs. spoken originals, translation vs. interpreting). Focusing on the language pairs English and German, we have revisited the question of the distinctive properties of mediated discourse, i.e. translation and interpreting. Using computational language models combined with the information-theoretic measure of relative entropy (here: Kullback-Leibler Divergence), we have shown how to detect and assess features indicating major differences between the different modes in a data-driven way (§4). In a second step, the words found to be distinctive by KLD modeling have been related to known measures of corpus comparison such as type-token ratio as an indicator of vocabulary variation and used as a basis for engineering more abstract and more complex features for further analysis (parts-of-speech, grammatical patterns (§5).

Comparing translation and interpreting (including the relation to their originals), we confirm the previously observed trend of written vs. spoken mode being strongly reflected in translated and interpreted texts. Several aspects of our analyses for the language pair German and English confirm Shlesinger & Ordan's (2012) earlier observation that interpreting is strongly characterised by general

spoken language features and that it is not merely a different mode of translation. We also detected more subtle features typical of interpreting, e.g. a preference for syntactic coordination or the tendency to use general verbs, as well as differences between English and German interpreted texts, e.g. a pronounced use of deictic expressions in German. Some of the observed features and the subsequently performed linguistic analysis may be linked to traditional translationese features (e.g. simplification on the lexical level for SI) but often with different trends for interpreting and translation. Our analyses show that translation overemphasizes features associated with written mode, while interpreting tends to be "more spoken" and conceptually oral than comparable originals.

In our future work, we plan to investigate other linguistic levels, notably the morphological, semantic and the phonetic level. Word-internal structures and other aspects of morphology should shed light on the degree of term variation and consistency in mediated vs. non-mediated discourse. Variants, for instance, are probably found more typically in original texts, whereas we expect to see a higher degree of formulaicity in translations. Original texts, translations and interpreted language might make use of particular patterns indispensable for language economy in different ways. They might differ, for instance, in usage preferences for acronyms of complex terminological units with the aim to reduce articulatory or memory efforts. To better understand the mechanisms underlying lexico-semantic choice in translation and interpreting, we apply word embedding models (Bizzoni & Teich 2019); and to better understand the phonetic side of interpreting output we would also like to examine the different types of hesitations and pauses produced by interpreters and find correlations with indicators of processing effort such as entropy and surprisal.

Acknowledgements

This paper is based on research conducted in a project funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1102 / Project-ID 232722074. We are grateful to Stefan Fischer for providing the models and visualizations. We also thank the anonymous reviewers for their insightful suggestions and comments.

References

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.

Becher, Victor. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1). 1–28. DOI: 10.1556/Acr.11.2010.1.1.

Bendazzoli, Claudio & Annalisa Sandrelli. 2005. An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In Heidrun Gerzymisch-Arbogast & Sandra Nauert (eds.), *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*, 149. https://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html (5 December, 2012).

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC |Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 21–42. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-6199-8_2.

Bizzoni, Yuri & Elke Teich. 2019. Analyzing variation in translation through neural semantic spaces. In Pierre Zweigenbaum Serge Sharoff & Reinhard Rapp (eds.), *Proceedings of the 12th Workshop on Building and Using Comparable Corpora at Recent Advances in Natural Language Processing (RANLP) - Special topic: Neural Networks for Building and Using Comparable Corpora, Varna, Bulgaria.*

Chesterman, Andrew. 2004. Beyond the particular. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist*, 33–49. Amsterdam & Philadelphia: John Benjamins.

Dayter, Daria. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM: International Journal of Interpretation and Translation* 16(2). 241–264. DOI: 10.1075/forum.17004.day.

Defrancq, Bart. 2018. The European Parliament as a discourse community : Its role in comparable analyses of data drawn from parallel interpreting corpora. *The Interpreters' newsletter* 23. 115–132. DOI: 10.13137/2421-714X/22401.

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to*

discourse and translation studies, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.

Degaetano-Ortlieb, Stefania. 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the 2nd Workshop on Stylistic Variation collocated with NAACL HLT 2018*, 1–10. New Orleans, USA. https://stefaniadegaetano.files.wordpress.com/2018/06/naaclhlt2018-degaetano.pdf.

Degaetano-Ortlieb, Stefania & Elke Teich. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*. 1–33. DOI: 10.1515/cllt-2018-0088.

Evert, Stefan & Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies. New theoretical and methodological traditions*, vol. 300 (Trends in Linguistics. Studies and Monographs (TiLSM)), 47–80. Berlin: Mouton de Gruyter.

Fankhauser, Peter, Jörg Knappen & Elke Teich. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the Language Resources and Evaluation Conference (LREC), May 2014, Reykjavik, Iceland*, 4125–4128.

Ferraresi, Adriano, Silvia Bernardini, Maja Petrović & Marie-Aude Lefer. 2018. Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta* 63(3). 717–738. DOI: https://doi.org/10.7202/1060170ar.

Hansen-Schirra, Silvia, Stella Neumann, Erich Steiner, Oliver Culo, Sandra Hansen, Marlene Kast, Yvonne Klein, Kerstin Kunz, Karin Maksymski & Mihaela Vela. 2013. *Cross-linguistic corpora for the study of translations*. Berlin, Boston: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110260328.

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.

House, Juliane. 2008. Beyond intervention: Universals in translation? *Trans-kom. Zeitschriftfür Translationswissenschaft und Fachkommunikation* 1(1). 6–19. http://www.trans-kom.eu/bd01nr01/trans-kom_01_01_02_House_Beyond_Intervention.20080707.pdf (5 December, 2012).

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu. (Doctoral dissertation). https://repozytorium.amu.edu.pl/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf.

Kajzer-Wietrzny, Marta. 2015. Simplification in interpreting and translation. *Across Languages and Cultures* 16(2). 233–255. DOI: 10.1556/084.2015.16.2.5.

Karakanta, Alina, Heike Przybyl & Elke Teich. Forthcoming. Exploring variation in translation with probabilistic language models.

Karakanta, Alina, Mihaela Vela & Elke Teich. 2018. Europarl-UdS: Preserving metadata from parliamentary debates. In Darja Fišer, Maria Eskevich & Franciska de Jong (eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018) Miyazaki, Japan, May 2018*. Paris, France: European Language Resources Association (ELRA).

Klingenstein, Sara, Tim Hitchcock & Simon DeDeo. 2014. The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences* 111(26). 9419–9424.

Lapshinova-Koltunski, Ekaterina. 2015. Variation in translation: Evidence from corpora. In Claudio Fantinuoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies* (Translation and Multilingual Natural Language Processing (TMNLP)), 93–114. Berlin: Language Science Press.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4). 557–570.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli & Mariachiara Russo. 2005. Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta* 50(4).

Oakes, Michael P. 2021. Corpus statistics for empirical translation studies. In Ji Meng & Sara Laviosa (eds.), *The Oxford handbook of translation and social practices*, 543–560. Oxford: Oxford University Press.

Plevoets, Koen & Bart Defrancq. 2018. Lexis or parsing? A corpus-based study of syntactic complexity and its effect on disfluencies in interpreting. In *UCCTS 5 - Using Corpora in Contrastive and Translation Studies*. Louvain-la-Neuve. http://hdl.handle.net/1854/LU-8562920.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli & Nicoletta Spinolo. 2012. The European Parliament Interpreting Corpus (EPIC): Implementation and developments. *Breaking ground in corpus-based Interpreting Studies* 147. 53–90.

Sandrelli, Analisa, Claudio Bendazzoli & Mariachiara Russo. 2010. European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary

results on lexical patterns in simultaneous interpreting. *IJT-International Journal of Translation* 22. 165–203.

Sandrelli, Annalisa & Claudio Bendazzoli. 2005. Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). In *Proceedings from the Corpus Linguistics Conference Series*, vol. 1.

Shlesinger, Miriam. 1989. *Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum*. Tel Aviv University, Faculty of the Humanities, Department of Poetics & Comparative Literature. (MA thesis). DOI: 10.13140/RG.2.2.31471.69285.

Shlesinger, Miriam & Noam Ordan. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24(1). 43–60. DOI: 10.1075/target.24.1.04shl.

Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, vol. 5. Berlin: Mouton de Gruyter.

Teich, Elke, José Martínez Martínez & Alina Karakanta. 2020. Translation, information theory and cognition. In Fabio Alves & Arnt Lykke Jakobsen (eds.), *The Routledge handbook of translation and cognition*, chap. 20, 360–375. London: Routledge.

Vandevoorde, Lore. 2020. *Semantic differences in translation : Exploring the field of inchoativity* (Translation and Multilingual Natural Language Processing 13). Berlin: Language Science Press. DOI: 10.5281/zenodo.2573677.

Chapter 8

NLP-enhanced shift analysis of named entities in an English<>Spanish intermodal corpus of European petitions

Gloria Corpas Pastor[a] & Fernando Sánchez Rodas[a]
[a]University of Malaga

This chapter aims at presenting an NLP-enhanced corpus-based analysis of the translation and interpreting shifts observed in the named entities (NEs) of PETI-MOD, an English<>Spanish intermodal corpus of written and oral mediated texts from the Committee on Petitions of the European Parliament. Our main assumption is that shifts in institutional genres mostly occur in the transfer of NEs, and that NLP techniques such as automatic Named Entity Recognition (NER) can be applied to systematically extract and compare examples of these shifts, leading to the (possible) verification of translational and/or interpretational constraints. Results show that traits like normalisation, transformation and simplification depend not only on the language direction or the mediation mode, but also on the semantic category (person, organisation, etc.) of the NE involved. Further studies are needed in order to correlate observed shifts with different NE taxonomies.

1 Introduction

To the present day, a considerable amount of corpus-based research in translation and interpreting has relied on the European Parliament (EP) as a main or only source. Among the European Union (EU) institutions, the Parliament provides an open access repository of both official documents and speeches in a wide range of languages and topics. Before the appearance of intermodal corpora such as EPTIC (Bernardini et al. 2016), the EP had already been used as a source for building translation corpora, i.e., Europarl (Koehn 2005), the European Parliamentary Comparable and Parallel Corpora, or ECPC (Martínez & Serrat 2012),

Gloria Corpas Pastor & Fernando Sánchez Rodas. 2022. NLP-enhanced shift analysis of named entities in an English<>Spanish intermodal corpus of European petitions. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Empirical investigations into the forms of mediated discourse at the European Parliament*, 209–240. Berlin: Language Science Press.
DOI: ??

and the EU resources at Sketch Engine (Baisa et al. 2016). In the field of corpus-based interpreting studies, it was early pointed out that EP linguistic material could provide researchers with numerous advantages (Bendazzoli 2010). The European Parliament Interpreting Corpus (EPIC) is an example of this (Russo et al. 2012). However, researchers have not yet attended the call. In spite of their unquestionable relevance and high-level complexity, legislative chambers have not received that much attention from linguistics until very recently[1] (**Calzada2018**). Bibliometric analyses of Europarl (one of the largest multilingual corpora available) show that it has hardly been used in translation studies[2] (Ustaszewski 2019). Reasons for this little academic interest may include corpora distribution in a format that largely disregards the needs of translation research and practice (ibid.) and the need for unexplored, more down-to-earth studies which empirically look at the compared properties of source texts, translations and interpretations and offer a modern, technology-based twist on the methodologies involved.

Against this background, we hypothesise that texts and speeches originated in the Committee on Petitions of the European Parliament provide an excellent source for the observation of shifts in institutional translation and interpreting, and that shifts in these genres are mostly given in the transfer of Named Entities (NEs). We also assume that recent techniques based on Natural Language Processing (NLP) can be applied to the recognition, extraction and comparison of segments with NEs in two languages and/or modes, as a systematic way of observing shifts between them and proving (or not) the existence of translation and interpreting universals in the analysed texts. To this end, our main research objectives are as follows:

- compile an intermodal, bidirectional corpus (English<>Spanish) of translations and interpretations (plus their different, corresponding source texts) of suitable genres from the EP Committee on Petitions;
- apply NLP-based techniques (Named Entity Recognition) on the said corpus in order to extract relevant units for the study of shifts in both languages and modes;
- compare qualitatively and quantitatively the observed shifts in the English-Spanish translations and Spanish-English interpretations of the Committee;

[1]See Veroz González (2014a,b, 2017) and Prieto Ramos (2019) for examples of corpus-based discursive and/or linguistic analysis in this field.

[2]In order to make the wealth of linguistic data easily and readily available to the translation studies community, a toolkit named EuroparlExtract has been recently developed (Ustaszewski 2019).

- draw conclusions on the relation of three different parameters (language, mode, and semantic category of the NEs) with the presence of translation and interpreting universal features in the analysed documents, especially of simplification traits.

In connection with the objectives above, the chapter presents the following structure. After this introduction (§1), §2 covers basic notions related to communications in the Committee on Petitions. §3 describes the PETIMOD corpus, with a special focus on data collection and design criteria. The NLP-based methodology deployed in this study is spelled out in §4; the main findings are presented in §5 and then discussed in detail (§6). After considering some limitations of our study, §7 offers some concluding remarks on the implications of intermodal corpora for research in translation and interpreting, with special reference to shifts, mediation types and functions, among other relevant issues.

2 A brief overview of EU Petitions

The right to petition is set out in the European legislation. Article 44 of the Charter of Fundamental Rights of the European Union ensures the right to petition to the European Parliament. And Article 227 of the Treaty on the Functioning of the European Union states that "any citizen of the Union, and any natural or legal person residing or having its registered office in a Member State" shall have the right to address a petition to the European Parliament (European Union 2012). A petition may "take the form of a complaint, a request or an observation concerning problems related to the application of EU law or an appeal to the European Parliament to adopt a position on a specific matter" (European Parliament 2020b). After submission, original petitions are registered and given a number. Then, they are summarised (normally in English) and submitted to the members of the Committee on Petitions of the European Parliament for a decision on admissibility and follow-up (ibid.). This committee serves a core function within the governance of the Union, as it acts "as a bridge between Europeans and the EU institutions" (European Parliament 2020a).

As the Committee on Petitions plays an important, mediating role in the context of a multilingual institution and society such as the EU, translation and interpreting are especially relevant in assuring the transparency of its communications. Petition summaries are translated and published in all official EU languages on the Petitions Portal of the European Parliament right after a decision

on admissibility has been taken[3] (European Parliament 2020b). The speeches of the committee meetings are also interpreted into each official language and published in the Webstreaming section of the European Parliament Committees website.[4]

Being these institutional texts, translators and interpreters have to deal with an important amount of terminology. As Goffin (1994: 637–638) states, the language used in the EU texts, or eurolect, is no different in origin, semantic organization or morpho-syntagmatic characteristics from any other specialized dialect. Depending on the concept they represent, EU terms are classified as *euronymes,* i.e. terms coined for new institutional realities, or *hétérolexies*, i.e. terms which convey notions and designations rooted in each of the main languages[5] (Goffin 1994: 641).

This classification indicates a prominence of entities in this knowledge field. Entities are abstractions from external experience which are perceived as self-defined, that is, independent in time and space. Born out of our worldly experience, some entities are highly culture-bound, which poses a real challenge for translators and interpreters (Mayoral 1999). This is especially true for institutional references, which are usually related to the political life of a society (Martin 1997; Ortega 2002). In the Committee on Petitions, where citizens and platforms strive to expose national problems and petitions are chosen by Members of the European Parliament (MEPs) on the basis of their political relevance, it is highly important to give these relevant entities a (see §4).

3 The PETIMOD Corpus

The purpose of our compilation was to create an intermodal corpus of EU petitions suitable for the study of shifts in translated and interpreted NEs. The size of the corpus was initially limited to one month of institutional activity, and its medium written (see expanded size data in §3.2). The authorship of the documents was exclusively institutional and the topics were mostly agricultural and environmental, which was not determined by our sampling schema but given by the inherent frequency of the petitions. The publication date was a relevant criterion for the context of this research. As the elaboration of the paper ran parallel to the coronavirus crisis, a cancellation of the Committee activity and/or

[3] In fact, petitions are one of the most frequent briefings for the translation trainees of the EP Schuman Traineeships (https://ep-stages.gestmax.eu/website/homepage).

[4] https://www.europarl.europa.eu/committees/es/peti/meetings/webstreaming.

[5] Examples of the two categories extracted from our named-entity recognition would be "Eurobarometer" (*euronyme)* and "Boletín Oficial" (*hétérolexie*).

a change in the content of petitions was predicted. Therefore, the last Committee meeting before the health crisis (19th and 20th February 2020) was chosen as the main source of material. Finally, the languages of the corpus were Spanish and English in their institutional or EU varieties (for a fully-fledged study on eurolects, see Mori (2018)).

3.1 Data collection

The retrieval, storing, and conversion of materials started with the oral transcriptions. First, the audiovisual material for the meeting was accessed via the Web-streaming section of the EP Committees site. Three sessions were available for this debate: two on 19th February 2020 (morning[6] and afternoon[7] sessions) and one on 20th February (morning[8] session). We downloaded the complete recordings for both Spanish and English, obtaining six video files in high quality (HQ) .mp4 format.[9] These were moved into a folder structure and coded with the date and time of each session plus the corresponding language abbreviation (e. g. "19feb1000_EN.mp4"). The duration of each recording is indicated in Table 1.

For cost and ease-of-use reasons, Youtube was the selected application for further ASR (Automatic Speech Recognition) and ATT (Automatic Text Transcription).[10] The upload of the files was performed with a personal account in private visualisation mode to avoid copyright issues. The automatic transcription (without time marking) was generated, then copied and pasted in different TXT files, one for each intervention of the speakers. The naming pattern explained before was used, but three additional references were included for better localisation and connection with the petitions: intervention number, key word/expression related to the topic, and surname of the MEP/speaker (e.g. "19feb1430_17_-

[6] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200219-0900-COMMITTEE-PETI_vd.

[7] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200219-1430-COMMITTEE-PETI_vd.

[8] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200220-0930-COMMITTEE-PETI_vd.

[9] Audio tracks are available for the original speeches and the interpretations into any official EU language, although only one version can be downloaded at once. Download is performed through a request system which allows for choosing between the complete session and a selected part, and also between different video qualities. After this, a download link is sent to the desired email account. Downloading high-quality videos was the less time-consuming option in the long term, since low and medium quality videos had to be re-downloaded because of visualization problems. This is a relevant point, as videos are quite helpful for identifying the speakers in each petition.

[10] See Gaber et al.'s (2020) assessment of ASR systems for corpus compilation in interpreting.

Table 1: Properties of the audiovisual files used for automatic transcription.

File(s) name(s)	Length (hour, minutes and seconds)
19feb1000_EN.mp4 19feb1000_ES.mp4	02:10:19
19feb1430_EN.mp4 19feb1430_ES.mp4	03:14:43
20feb900_EN.mp4 20feb900_ES.mp4	02:32:24

ES_oranges_Rego.txt"). In the case of interpretations, the speech's original language was indicated between brackets with the mark "or-", as in this example: "19feb1430_78_EN(or-ES)_radioactivewaste_Montserrat.txt.".

Finally, the transcriptions were double-checked manually. In a first round, the EPTIC conventions for transcribing interpretations (Bernardini et al. 2018: 26–27) were applied. In a second revision, the Spanish and English versions of the EU Interinstitutional Style Guide, or ISG (**EuropeanUnion2011**), were used for spelling and capitalisation, together with other resources, such as the English Style Guide from the European Commission's Directorate-General of Translation[11] and the *Fowlers' Dictionary of Modern English Usage*.[12] Although the complete six videos in Table 1 were uploaded to Youtube and their transcriptions extracted in different TXT files, the only material revised manually and included in the transcribed component of the corpus was the one from the second session (19th February 2020 14:30–17:30). This was decided because the manual revision of all data was considered too time-consuming for the scope of this chapter. Additional reasons were that it was the longest session, and it contained the largest number of original Spanish speeches, which was in line with our goal of building a bidirectional corpus. As a result of this revision, we obtained 80 transcripts (40 transcriptions of original Spanish interventions and their corresponding 40 interpretations into English, with 18,152 and 10,530 words respectively).

A similar procedure was followed in the case of written documents. The No-

[11] https://ec.europa.eu/info/sites/info/files/styleguide_english_dgt_en.pdf.

[12] https://www.oxfordreference.com/view/10.1093/acref/9780199661350.001.0001/acref-9780199661350.

tices to Members were accessed through the eMeeting portal[13] of the European Parliament. We did not only look for the petitions mentioned in the revised session (19th February 14:30), but for the ones debated in the other two sessions as well, as this was a much quicker way of building our corpus. We browsed and downloaded the petitions in English and Spanish in PDF format. When possible, we included all the other accessible PDF documents which were not petitions but were also handled in the debates, such as reports and opinions. This was done for the sake of coherence and terminological relevance. Similarly to the transcriptions, these files were organised in a folder structure and renamed using a coding system with date and time of the meeting, language abbreviation and key word/expression related to the topic (e.g. "19feb1430_EN_oranges.pdf", "20feb900_EN_insects.pdf"). In the case of translations, the document's original language was indicated between brackets with the mark "or-", as in this example: "19feb1000_ES(or-EN)_amendment.pdf". Finally, the documents were saved as plain text (TXT) files with UTF-8 encoding for correct character recognition by any corpus software.

3.2 Design criteria

PETIMOD is a parallel intermodal corpus which contains citizens' petitions and other documents related to the Committee on Petitions of the European Parliament, as well as transcribed speeches related to these documents. It comprises two subcorpora, allowing for various types of comparison to be carried out: PETIMOD_ORIG (original texts and speeches in English and Spanish) and PETIMOD_MEDIATED (their corresponding translations and interpretations from English into Spanish, and vice versa). At the same time, PETIMOD is a *bidirectional* corpus (Olohan 2004) because the mediating activity is not only represented in B-A direction (Spanish speeches interpreted into English), but also A-B (English documents translated into Spanish). Finally, it is important to recall that, in contrast to other intermodal corpora in the field (cf. the works on EPTIC), PETIMOD comprises translations and interpretations (texts and speeches) that belong to different genres, the first being mostly Notices to Members and the second being interventions of said MEPs and speakers invited to the Committee on Petitions' sessions held in Brussels monthly.

Specifically, the corpus consists of all the petitions discussed during the three sessions of February 2020, whereas the original Spanish speeches and their English interpretations were extracted from a single session (19th February 2020

[13] https://emeeting.europarl.europa.eu/emeeting/committee/agenda/202002/PETI?meeting=PETI-2020-0219_1P&session=02-19-10-00.

14:30–17:30), as explained in §3.1 In order to diversify our corpus and investigate further correspondences, some non-petitional public documents discussed in the sessions, such as reports or opinions, were also included in the corpus.

According to classical typological parameters (Corpas Pastor 2001; Olohan 2004; Shlesinger 2008), the PETIMOD corpus can be classified as follows:

- it is *parallel,* as is composed of original texts (and speeches) plus their translations (and interpretations).
- It is *intermodal,* as it encompasses original, translated, and interpreted components which can be compared to each other in a three-way fashion.
- It is *written*, as it contains official documents (PDF and TXT) as well as transcriptions of parliamentary speeches (TXT).
- It is *bidirectional*, as it comprises English documents translated into Spanish (A-B), and also of Spanish speeches interpreted into English (B-A).

The size of the PETIMOD corpus is provided in Table 2 and Table 3 (in total, per component and per language). The total number of documents, running words (tokens) and word types (types) have been calculated by using Sketch Engine.

Table 2: PETIMOD size per component.

Counts	**Petimod_orig**	**Petimod_mediated**	**Total**
Tokens	67,575	72,775	140,350
Types	58,294	63,524	121,818
Documents	59	59	118

Table 3: PETIMOD size per language and component.

Counts	**Petimod _orig_en**	**Petimod _orig_es**	**Petimod _ mediated_es**	**Petimod _ mediated_en**
Tokens	54,220	13,355	61,181	11,594
Types	45,842	12,452	52,994	10,530
Documents	19	40	19	40

Figure 1 provides a visual representation of the composition of our intermodal corpus, in which the double arrows represent the (ordered) envisaged comparisons for analysis (A). In this study, the selected comparisons are A_5 and A_6. As it can be seen, cross-comparison of A_5 and A_6 presents differences not only in directions (EN<>ES), but also different languages in terms of origins (Anglosaxon and Romance), different modes (written and oral) and different types of linguistic mediation (translation and interpreting). This is a conscious choice which aims at raising awareness of the multifactorial nature of translation and interpreting phenomena (cf. De Sutter & Lefer 2019), but also at trying to establish generalisations between the two communicative situations by looking at a possible core set of shared factors given by the function of the institution for which they are produced, that is, the Committee on Petitions.[14]

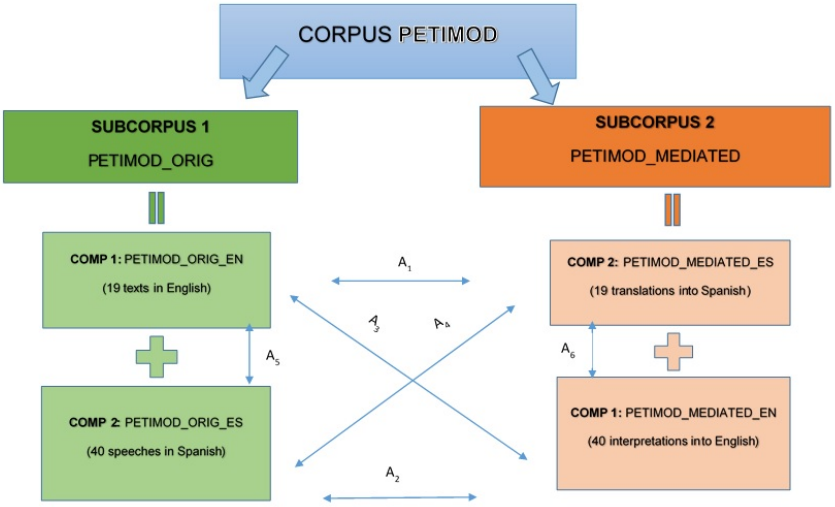


Figure 1: PETIMOD subcorpora and envisaged comparisons.

4 Methodology

In order to study shifts in translated speeches and interpretations, we have focussed on NEs and extraction techniques. Named entity recognition (NER) is the

[14] Cf. **Saldanha2009** for discussion on the bridging role of "function" and "context" in linguistic approaches to translation and interpreting.

task of identifying and categorising key information or real-world objects (entities) in text. In NLP, a NE is a real-world "object" that is assigned a name (e.g., *Donald Trump*, *United States*, *The Foreign Office*, *World Health Organisation*, etc.).

For this study both automatic and manual extraction of NEs have been performed. Both precision and recall have been calculated in order to assess the system's performance. Then, a corpus-based study of NEs in the translated and interpreted components have been carried out.

4.1 Automatic named entity recognition

Similarly to other models trained on a Wikipedia-based corpus (**NothmanEtAl2013**), for this paper we have used the VIP[15] NER annotation scheme, that distinguishes four entity types: per (named person or family), LOC (name of politically or geographically defined locations, e.g., cities, countries, regions, rivers, lakes, seas, mountains), ORG (named corporate, governmental or other organisational entities) and MISC (miscellaneous entities, e.g., laws, events, languages, products, work of art, etc.). In order to extract and identify NEs automatically, a script[16] has been programmed based on the VIP module for NE chunking, extraction, and identification. See Figure 2 for a screenshot of the Excel file generated by the script.

	A	B	C	D
1	PER	ORG	LOC	MISC
2	96/29//Euratom	//	Aachen	(Electoral Act
3	Aguilar	// the Commission	Aarhus	-now -Article 11
4	Aguilera	//i'm	Acoset	//
5	Aigües de Barcelona	20.12.2013	Annex II	// fifty cents
6	Alejandro Blasco Sánchez	2003/30/EC	Arenales of San Pedro del Pinatar	11/18
7	Alexander Edberg Thorén	2004/18/EC	Arenales of San Pedro del Pinatar	7/9
8	Ana Martínez	28.03.1987	Austria	28
9	Angel Dzhambazki	4.1.1 The Assembly	Aves	04.3.2006
10	Arianna Colonello	72/194/EEC	Barcelona	0956/2016
11	Arias Cañete	90/364/EEC	Belgium	1
12	Article 32(1	90/365/EEC	Belliardstraat	1 January 2012
13	Article 7(1)(c	A4	Berne	1 January 2013
14	Article 8b(2	AGRI	Bonn	1 January 2014
15	Article XXIII(6	AM\1197535EN.docx	Brussels	1 January 2019
16	Auken	Admissibility	Bruxelles	1 July 2019
17	B. K.	Aigües de Barcelona	Bulgaria	1 See
18	B.E.	Artificial Intelligence (AI	Cartagena	1(3
19	Bond Beter	Assembly	Craiova	1.1
20	Catherine the Great	BORM	Croatia	1.2
21	Christel Schlebusch	Barcelona Metropolitan Area'	Cyprus	1.3
22	Cristina Maestre Martín De Almagro	Bioenergy	Danube	1.4
23	Curtis	CAP	Decreto	1.5
24	Curtis-Teixeiro	CERMI	Denmark	10

Figure 2: English NEs file automatically retrieved by the VIP script.

[15] VIP (Voice-text integrated system for InterPreters) is a hub of online resources and computer-assisted tools for interpreters created by the research group Lexytrad of the University of Malaga. VIP includes a suit of interpreting-related tools with a NER module and its own annotation scheme. The platform can be accessed here: http://www.lexytrad.es/VIP/index_en.php.

[16] Authors would like to express their gratitude to Mr Francisco Javier Lima for writing the script used in this paper, which has been integrated in the VIP NER functionality.

VIP integrates spaCy[17] (a free open-source library in Python). VIP provides a user-friendly interface and allows importing NEs into an Excel file. Pre-trained spaCy models rather than custom-made NER models were used. The two pre-trained spaCy models used - es_core_news_lg (Spanish) and en_core_web_lg (English) - differ in the degree of granularity of the NER annotation scheme. The Spanish model recognises four categories (PER, LOC, ORG and MISC), whereas the English model recognises twelve additional types of entities: ORDINAL (e.g., st, *second*), DATE (*13 October, 2019*), GPE (countries, cities and states, e.g., *Madrid*), CARDINAL (*102, 67.5*), NORP (nationalities, religious or political groups, e.g. *Democrats*), FAC (buildings, airports, highways, bridges, etc., e.g. *Golden Gate*), PERCENT (percentage, including %), PRODUCT (objects, vehicle, foods, etc., e.g. *Toyota*), LAW (laws, directives, regulations, etc.), QUANTITY (measurements of weight, distance, etc., e.g., *hectare*), MONEY (e.g., *cents, dollars*), TIME (times smaller than a day), and LANGUAGE (e.g., *Spanish*). For this reason, English categories have been simplified. Thus, akin to the Spanish model, FAC and GPE have been subsumed under the category LOC and the rest have been grouped under MISC.

Precision has been calculated to measure how well our NER system performs. Precision is defined as the fraction of relevant instances among all retrieved instances, i.e. the total number of relevant NEs retrieved divided by the number of all NEs retrieved (correctly and incorrectly identified by the model).

$$Relevant\ NEs = Total\ number\ of\ correctly\ retrieved\ NEs - Errors$$

$$Precision = \frac{Relevant\ NEs}{Total\ number\ of\ extracted\ NEs}$$

For calculating the above formula, it was necessary to manually assign the retrieved NEs to three categories: (a) segments which were correctly identified as NEs ("Correct ID"), (b) segments wrongly identified as NEs ("Wrong ID"), (c) and segments correctly identified as NEs but wrongly labelled ("Wrong Class").

NER performance has been calculated in terms of precision for both languages. Two levels of analysis have been established. The first level takes all NEs correctly identified as relevant, irrespective of their classification. For instance, the non-entity sequence *[Articles 20(2)(b]*, retrieved as NE by the system, would be classified as an error, whereas the retrieved sequence *[2004/18/EC]* would be considered as relevant (correctly identified) whether it has been tagged correctly (MISC) or not (ORG). The mathematical formula for Level 1 is as follows:

[17] https://spacy.io/.

$$Precision = \frac{(Correct\ ID + Wrong\ Class)}{(Correct\ ID + Wrong\ ID + Wrong\ Class)}$$

Table 4 presents results for this wider category of relevant NEs.

Table 4: NER performance in terms of precision (Correct ID + Wrong Class).

	English	Spanish
NEs retrieved	1,726	1,183
Correct identification	1042	456
Wrong identification	522	576
Wrong class	162	151
Errors	684	727
Relevant NEs retrieved	1204	607
Precision	0.697	0.513

A further level of analysis is achieved by discriminating between NEs correctly identified and correctly tagged (for instance, *[2004/18/EC]* correctly identified as NE and classified as MISC) and NEs correctly identified but wrongly tagged (for instance NE *[2004/18/EC]* classified as ORG). The formula below allows refining results by considering wrong-labelled NEs as errors (see Table 5).

$$Precision = \frac{(Correct\ ID)}{(Correct\ ID + Wrong\ ID + Wrong\ Class)}$$

Table 5: NER performance in terms of precision (Correct ID).

	English	Spanish
NEs retrieved	1,726	1,183
Correct identification	1042	456
Wrong identification	522	576
Wrong class	162	151
Errors	522	576
Relevant NEs retrieved	1,042	456
Precision	0.603	0.385

4.2 Manual named entity extraction

In order to assess the performance of the system in terms of recall, it was necessary to identify and extract NEs manually for both languages. Recall is the fraction of retrieved instances among all relevant instances, i.e. it refers to the total number of relevant NEs retrieved versus the total number of relevant NEs found manually in our corpora. The idea was to delve into word lists generated by a corpus management tool, so we could identify NEs in the documents that had not been automatically recognised by our system. The sum of both types of NEs (automatically recognised and manually extracted) would bring the total number of relevant NEs in the corpus. The formula used to calculate recall is presented below:

$$Recall = \frac{Relevant\ NEs\ extracted}{Total\ number\ of\ relevant\ NEs\ in\ the\ corpus}$$

The selected corpus management platform was Sketch Engine, the same tool used for corpus statistics in §3.1. Sketch Engine was chosen for two reasons: it features European Parliament corpora (Ustaszewski 2019) and its interface allows for swift change when working with several subcorpora simultaneously. We uploaded the plain-text files for each of the four components of our intermodal corpus as four different monolingual comparable corpora, using the "New Corpus" functionality in the menu "Select Corpus → My Corpora".

Then, a starting point for manual NER was the wordlist generator of Sketch Engine, which was used in each component. We chose to compose a list of nouns filtered by two stopword lists (one for each language).[18] In Sketch Engine, this can be done in the "Advanced" tab of the wordlist menu, under the heading "Exclude these words"; the list has to be pasted manually, with one word per line. The PETIMOD_MEDIATED_EN subcorpus, for example, yielded a list of 643 nouns (e.g., *Commission*, *situation*, *petitioner*, *problem*, etc.).

Once the wordlist was generated (Wordlist 1), we had a basic frequency list which contained some nouns that could be used to refine the automatic NER, such as committee (22 occurrences), directive (16), agreement (13), group (11), plan (10), fund (9), etc. Then, a second wordlist (Wordlist 2) was created by sorting the nouns alphabetically and filtering out those which were neither semantically nor frequency-wise relevant (e.g. *angle*, 1) or which had been correctly recognised by the automatic NER (e.g. *Aguilar*, 1). Although Sketch Engine did not allow for

[18] Stopword lists were directly copied and pasted from http://members.unine.ch/jacques.savoy/clef/index.html. The interjection "ehm", used in the transcription conventions for representing hesitation in speech, was also added to the stopword list.

alphabetical sorting of the wordlist, nor for complete visualisation of the results in one column (the maximum is 500), it was possible to download the data in a CSV file and order the words by using the corresponding Excel function.

The next step was to search for the nouns in the wordlist manually. To this end, we opened a new window of concordances in Sketch Engine to directly search for the occurrences of each noun in the corpus. At this point, some basic functions of concordance search, such as alphabetical sort by context (left and right), file view, and wildcard search, were also used for easier and faster identification of new entities. Wildcard search proved especially useful in combination with the wordlist, as in some cases looking for lexical roots made it possible to inspect several instances of the list at once. For instance, a search for *[*omission*]* retrieved up to three instances of the wordlist simultaneously (*Commission*, *commission*, and *commissioner*).

Apart from wordlist frequency, institutionalisation was the second criterion for identifying relevant NEs. In this case, coverage in Eur-lex,[19] IATE[20] and/or TermCoord's Glossary Links[21] was taken as a reference (see Figure 3).

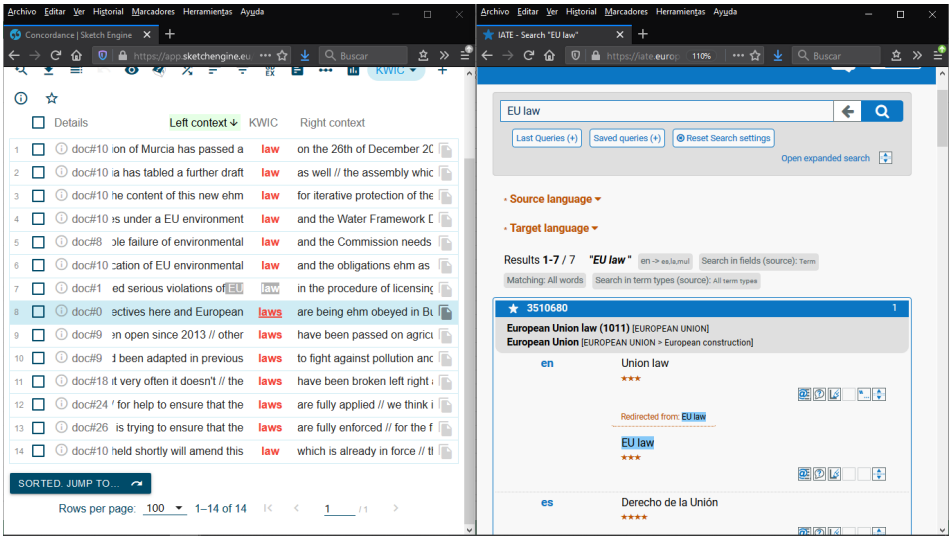


Figure 3: Example of manual NER using institutional criteria. The consulted NE ("EU law") had not been automatically recognised.

[19] https://eur-lex.europa.eu/.

[20] https://iate.europa.eu/home.

[21] A database of more than 8,000 glossaries managed by the Terminology Coordination Unit of the EP Directorate-General for Translation (https://termcoord.eu/glossarylinks/).

Following these criteria, new NEs were extracted from the concordances in each component and saved in an Excel file. Some examples of further relevant NEs manually extracted were *Directorate-General for the Mar Menor* (ORG, PETIMOD_MEDIATED_EN), *Acuerdo de Asociación Económica* (MISC, PETIMOD_MEDIATED_ES), *municipality of Real* (LOC, PETIMOD_ORIG_EN) and *Directiva de inundaciones* (MISC, PETIMOD_ORIG_ES), among others.

Finally, NER performance has been calculated in terms of recall for both languages. As in the case of precision, two granular levels of analysis have been used. The first level takes all NEs correctly identified by the automatic script as relevant, irrespective of their classification (see §4.1). For these calculations, it was necessary to sum the manually retrieved NEs for each component, combining and sorting them by language.

Table 6: NER performance in terms of recall (Correct ID + Wrong Class).

	English	**Spanish**
Total no. of relevant NEs	1,557	896
Relevant NEs retrieved automatically	1,204	607
Relevant NEs retrieved manually	353	289
Recall	0.773281	0.677455

A further level of recall analysis is achieved by discriminating between NEs correctly identified and correctly tagged by the automatic script (relevant) and NEs correctly identified but wrongly tagged (not relevant). This allows refining recall results by excluding wrong-labelled NEs from calculation.

Table 7: NER performance in terms of recall (Correct ID).

	English	**Spanish**
Total no. of relevant NEs	1,075	745
Relevant NEs retrieved automatically	1,042	456
Relevant NEs retrieved manually	353	289
Recall	0.969302	0.612080

4.3 Corpus-based analysis

For the corpus-based analysis described in this section, all relevant NEs in the Excel files (correctly identified, correctly identified but mislabelled, and manually extracted) were prepared by listing them together in a new file, manually sorting them by category and language. Figure 4 below shows the two columns for the PER category (English and Spanish), the first one attending to the VIP annotation scheme order described above.

Once all NEs were prepared, the next step was analysing the observable shifts in their translation and/or interpretation. We decided to perform the shift analysis both in the EN>ES translation (components PETIMOD_ORIG_EN vs. PETIMOD_MEDIATED_ES, or direction A1 in Figure 1) and in the ES>EN interpretation (components PETIMOD_ORIG_ES vs. PETIMOD_MEDIATED_EN, or direction A2 in Figure 1). The reasons for this decision were two: it comprised all the different components in our corpora, it and the cross-comparison of translation and interpreting analysis would expectably show interesting findings.

Providing that the raw material for analysis (i.e., the NEs) were already extracted, labelled, and sorted by language, the next three steps to be taken were: (1) contrasting them across languages to observe (possible) changes; (2) searching for them in the corpora in order to extract contextual exemplification of the shifts and identify their direction; (3) categorising the shifts. Step 1 could be done directly in the Excel file, underlining those units already analysed and/or not shifted. For Step 2, we prepared a mosaic-style panel of four windows, one for each uploaded component in Sketch Engine, in order to identify the directions and the exact alignment of the document in which the shift occurred (see Figure 5). A total of 142 shifts (69 for EN>ES translation, 73 for ES>EN interpretation) were identified and analysed. Regarding Step 3, the bottom-up transfer operations typology of Bernardini (2016) was chosen to categorise the shifts (see §5 for further description). The category was annotated next to the extracted concordances, in a table-like fashion. The Excel file for shift analysis included retrieved NEs and categorised shifts sorted by direction. As it can be inferred, the previous work with the entities in the automatic and manual NER phases was very helpful for this analysis, and allowed for quick identification and remembrance of the nature and direction of several shifts. Again, the institutional resources cited in §4.2 (Eur-lex, IATE, Glossary Links) were occasionally used in combination with generic searches in Google and/or Wikipedia in order to gain insight into the possible motivations behind some of the shifts encountered.

	A	B
1	PER-EN	PER-ES
2	Ádám Kósa	Ádám Kósa
3	Aguilar	Aguilar
4	Aguilera	Aguilera
5	Alejandro Blasco Sánchez	Alejandro Blasco Sánchez
6	Alexander Edberg Thorén	Alexander Edberg Thorén
7	Álvarez	Alexandrov
8	Ana Martínez	Ana Martínez Vidal
9	Angel Dzhambazki	Angel Dzhambazki
10	Arianna Colonello	Antoaneta Rizova-Kalapish
11	Arias Cañete	Arianna Colonello
12	Auken	Arias Cañete
13	B. K.	Auken
14	B.E.	B. E.
15	Catherine the Great	B. K.
16	Christel Schlebusch	Catalina la Grande
17	Cristina Maestre Martín De Almagro	Christel Schlebusch
18	Domènec Ruiz Devesa	Clara Aguilera
19	Dzhambazki	Cristina Maestre Martín De Almagro
20	Fernando López Miras	Domènec Ruiz Devesa
21	Fernando Novella Asensio	Eduardo Salazar Ortuño
22	Fragkos	Fernando López Miras
23	Giovanni Cortese	Fernando Novella Asensio
24	Isabel Rubio Perez	Giovanni Cortese
25	Ismael Antonio López Pérez	Gloria
26	J.B.	H. E.
27	Joaquín Pérez Gómez	Isabel Rubio Pérez
28	Jordi Cañas	Ismael Antonio López Pérez
29	José Luis Álvarez	J. B.
30	José Luis Álvarez Rubio	Joaquín Pérez Gómez

Figure 4: Screenshot of the Excel file with extracted PER NEs (English/Spanish).

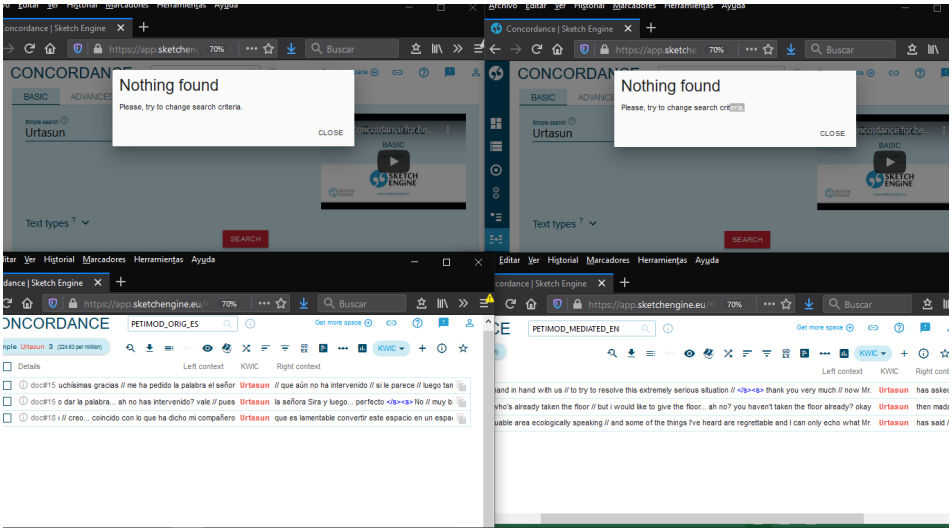


Figure 5: Four-window panel in Sketch Engine to track shifts in the corpus.

5 Shift analysis and results

The term "shifts" commonly refers to "changes which occur or may occur in the process of translating [and interpreting]" (Bakker et al. 2009: 269). Shifts of translation (and interpreting) can be distinguished from the systemic differences which exist between source and target languages and cultures. Systemic differences, which pertain to the level of competence, are part of the opening conditions for translation (and interpreting). Shifts, on the other hand, result from attempts to deal with systemic differences (ibid). In this study, only NEs that experienced shifts during translation/interpreting were analysed, whereas translations/interpretations where no shifts in NEs occurred were ignored. As stated in the previous section, the bottom-up transfer operations typology from Bernardini (2016: 140), used to categorise shifts in the intermodal corpus EPTIC, was chosen for this analysis. It includes register shifts (either upwards or downwards), quantitative meaning shifts (contraction, expansion, clarification, broadening), and transformational meaning shifts (partial and total), as well as cases akin to normalisation.[22] In the next paragraphs, each of these categories will be

[22]For the sake of clarity, the original name of this category was rephrased for this chapter (from "more collocational" to "normalisation").

described and illustrated with examples from our corpus.[23] However, as Bernardini (ibid.) puts it:

> As often happens with language in use, some instances were impossible to assign indisputably to one category only. In these cases a decision was made based on a close reading of the co-text and, inevitably, intuition as to the main reason for making a certain choice. (Bernardini 2016: 140)

The first type of shifts, categorised under "register" (up and down), could indeed be sometimes confused with contraction and expansion changes. Illustrating them with the use of acronyms helps establish a clear-cut separation between register shifts (formal) and meaning shifts. In example (1), the acronym is avoided in the EN-ES direction, which increases the level of formality. It is shifted to the modifier *de la Unión*, which in the Spanish eurolect can be considered even more formal than the alternative *de la Unión Europea* because of its specificity. Exactly the same change can be further found in the same sentence (from *EU Member States* to *Estados Miembros de la Unión*).

(1) Register up shift (EN-ES translation)
   a. The EU Delegation in Japan and the authorities of EU Member States [PETIMOD_ORIG_EN]
   b. La <Delegación de la Unión> en Japón y las autoridades de los Estados miembros de la Unión [PETIMOD_MEDIATED_ES]

On the contrary, in the ES-EN interpretation in example (2), the acronym *ENVI* is preferred instead of the denomination *Comisión ENVI* (already shortened in the original). As the Spanish ISG always recommend the use of the word *Comisión* when referring to these bodies (cf. **EuropeanUnion2011**: 172), this can be considered a shift which downgrades register. In fact, some shifts of the same nature can be observed in the surrounding verbs (*pedimos* → *pass*, *realice* → *carry out*).

(2) Register down shift (ES-EN interpretation)
   a. le pedimos a la Comisión ENVI que realice una visita [PETIMOD_ORIG_ES]
   b. we should pass it on to <ENVI> and ask them to carry out a study... visit [PETIMOD_MEDIATED_EN]

[23] We followed the same conventions of Bernardini (2016: 140). The underlined NE in the source roughly corresponds to the NE or segment in the target (in angle brackets).

Moving to quantitative meaning shifts, contraction implies changing from an informative detailed NE or NE sequence to a shorter and more under-defined equivalent (Bernardini 2016: 141). Although the author does not put it explicitly, it can be deduced from the given examples that contraction and expansion are related, as the reduction (or addition) of meaning also conveys a reduction or addition in the number of words (ibid.). In example (3), the English word referring to the region (*Galicia*) is omitted in the Spanish translation, as it is (supposedly) not necessary for a standard Spanish reader.

(3) Contraction shift (EN-ES translation)
   a. in an existing business park, on a green field plot, in Curtis-Teixeiro, La Coruña , Galicia, Spain. [PETIMOD_ORIG_EN]
   b. en un parque de actividades económicas ubicado en un terreno no urbanizado de Curtis-Teixeiro <(La Coruña„ España)> [PETIMOD_MEDIATED_ES]

A similar example, but this time of expansion, could be extracted from the ES-EN direction. Here we also have a LOC NE referred to a quite specific Spanish area (*Campo de Cartagena*), but the interpreter's decision is the opposite one: adding the modifier *region* to specify the nature of the named entity, thus increasing the number of words.

(4) Expansion shift (ES-EN interpretation)
   a. él estaba contentísimo con el modelo agrícola del Campo de Cartagena [PETIMOD_ORIG_ES]
   b. they were very happy with the agricultural model in the <Campo de Cartagena region> [PETIMOD_MEDIATED_EN]

Like expansion shifts, clarifications are instances of addition, in which meanings that are implicit in the sources are made explicit in the targets. As a rule of thumb, Bernardini (2016: 140) states that "in the case of clarification words used are more explicit, whereas in the case of expansion there is also an increase in the number of words (though admittedly the difference is not always clear-cut)." For improved distinction, it could be added that clarification seemingly implies adding *less* words that any expansion would. Again, the LOC label provides a suitable example in the EN-ES translation. In example (5) the unit *municipality of Real*, which initially refers to a geopolitical entity and could imply demanding information from any office contained in these borders, is shifted to a more explicit reference (*Ayuntamiento de Real* or town hall). Interestingly, by performing this operation, the nature of the NE is also shifted (from LOC to ORG).

(5) Clarification shift (EN-ES translation)

a. the Environmental Inspection Service requested the municipality of Real to inform [PETIMOD_ORIG_EN]

b. En 2012, el Servicio de Inspección Medioambiental pidió al <Ayuntamiento de Real> información [PETIMOD_MEDIATED_ES]

The third possible case of quantitative meaning shift is broadening, or generalisation through vaguer or emptier terms. In example (6), two PER NEs are generalised through the common, more neutral noun *petitioners*. This is a quite prototypical example, as additionally the first PER (*Eduardo Salazar Ortuño*) is not one of the petitioners, but a lawyer who is present on behalf of them (this is contextual information which can be found in the corpus some interventions before). Other aspects worth mentioning are the double nature of the shift and the extended broadening phenomena in the two MISC NEs *dos minutos*, which are suppressed in favour of the more general idea conveyed by *conclude*.

(6) Broadening shift (ES-EN interpretation)

a. para concluir esta petición le daríamos la palabra por dos minutos al señor Eduardo Salazar Ortuño // y luego le daríamos dos minutos más al señor José Luis Álvarez-Castellanos Rubio [PETIMOD_ORIG_ES]

b. let's close the debate on that and we will conclude this point by giving the floor back to <our two petitioners> [PETIMOD_MEDIATED_EN]

Transformational shifts include two different grades (partial and total). Partial transformation involves a reformulation with approximately the same co-textual meaning, but using an unrelated expression with a different out-of-context meaning (Bernardini 2016: 142). Again, the ES-EN interpretation provides a prototypical example of partial transformation. The collocation *flourishing ecosystem* in example (7) does not convey the same specialised meaning as *Zona de Especial Conservación*, but serves as equivalent in the context of the inversion operated in the target sentence. As already observed in example (6), the shift affects more than one particular NE and can be analysed even at the sentence level.

(7) Partial transformation shift (ES-EN interpretation)

a. en la cuenca del Mar Menor la Red Natura 2000 es una etiqueta formal que no responde a una gestión eficiente de lo que sería una Zona de Especial Conservación [PETIMOD_ORIG_ES]

b. Natura 2000 is an official label that should lead to efficient management of what should be a <flourishing ecosystem> [PETIMOD_MEDIATED_EN]

Total transformation, on the other hand, may sometimes override the limits of equivalence and fall closer to the notion of translation error (see for example Hurtado 2017). In example (8), the translator seems to have looked for the real (and very different) equivalent of the generic NE underlined in (o), but has made a mistake in the process (*Consejería de Turismo y Cultura* instead of *Consejería de Turismo, Cultura y Medio Ambiente*).[24] This is a very similar case to the one illustrated by Bernardini (2016: 142), in which an error is produced in the search of a salient collocation (here, a NE) in the target language.

(8) Total transformation shift (EN-ES translation)

a. the creation of a specific Directorate-General for the Mar Menor, within the regional Department for the Environment [PETIMOD_ORIG_EN]

b. la creación de una Dirección General del Mar Menor, dentro de la <Consejería de Turismo y Cultura> [PETIMOD_MEDIATED_ES]

The last shift category presented in this typology is normalisation. In the words of Bernardini (Bernardini 2016: 142), here "the difference from source to target seems to be one of *collocationality*: i.e., the inherent motivation for using a certain turn of phrase seems to be its salience as a phrase, or status as a collocation, in the target language." In this study, however, the analysed normalisation shifts are not performed on collocations, but on multi-word terms or NEs, such as the ones in example (9). In this translation, a subtle shift in a preposition (*National Assembly in France* → *Asamblea Nacional de Francia*) reveals a more frequent[25] multi-word term in the target language than *Asamblea Nacional en Francia*.

(9) Normalisation shift (EN-ES translation)

[24] See https://www.borm.es/services/anuncio/ano/2017/numero/3482/pdf?id=757271. In fact, the name of the supervising office is now "Consejería de Agua, Agricultura, Ganadería, Pesca y Medio Ambiente" (https://administracion.gob.es/pagFront/espanaAdmon/directorioOrganigramas/fichaUnidadOrganica.htm?idUnidOrganica=123379&origenUO=comunidadesAutonomas&comunidadesAutonomas=true&volver=comunidadesAutonomas&idCCAA=14#.X-DOye1Ce00).

[25] For example, a search in the Spanish reference corpus CORPES (https://webfrl.rae.es/CORPES/view/inicioExterno.view) yields five results against zero.

a. on 16 February 2019, the National Assembly in France has adopted the law of programming 2019–2022 and the justice reform [PETIMOD_ORIG_EN]

b. la <Asamblea Nacional de Francia> adoptó el 16 de febrero de 2019 la ley de programación 2019–2022 y la reforma judicial [PETIMOD_MEDIATED_ES]

6 Discussion

In this section, the overall results of our analysis are discussed, focusing on three different quantifications for both translation and interpreting: 1) distribution of the type of shifts retrieved; 2) distribution of the labels of the shifted entities; and 3) the detailed shift entity relationship with all the subcategories of shifts as described above. Then, we will relate our findings to results reported in related literature on intermodal corpora.

Figure 6 quantifies certain tendencies within English-Spanish translations and Spanish-English interpretations in the Committee on Petitions. The most prominent shifts are quantitative shifts (75 instances in both language pairs) and register shifts (33), followed by transformational shifts (18) and normalisations (16). There is a predominance of register shifts in EN-ES translations (20 against 13) and a fairly more balanced number in the case of quantitative meaning shifts (36 against 39). Transformational meaning shifts are more numerous on ES-EN interpretations (2 against 16); inversely, normalisations are more present in the translations into Spanish (11 against 5).

Figure 7 shows the distribution of shifted NEs per label, as illustrated in §3 of this chapter. MISC entities are the most frequent (58), closely followed by ORG (51); LOC (22) and PER (11) are considerably less represented in the shifts. The miscellaneous entities are more subject to shifts in the interpretations into English (26 against 32); conversely, organisational entities are prone to shifts in the translations into Spanish (30 against 21). The number of locations remains fairly equal in both directions (12 against 10). Finally, shifts in named persons are almost inexistent in EN-ES translations (only 1 result) in comparison to ES-EN interpretations (10).

In Table 8, 9 and 10, the two types of data commented above (type of shifts and type of entities) are cross-related and broken down into the nine shift subcategories used for this study.

Table 10 contrasts the subcategories of shifts encountered in both directions (EN-ES translations vs. ES-EN interpretations).

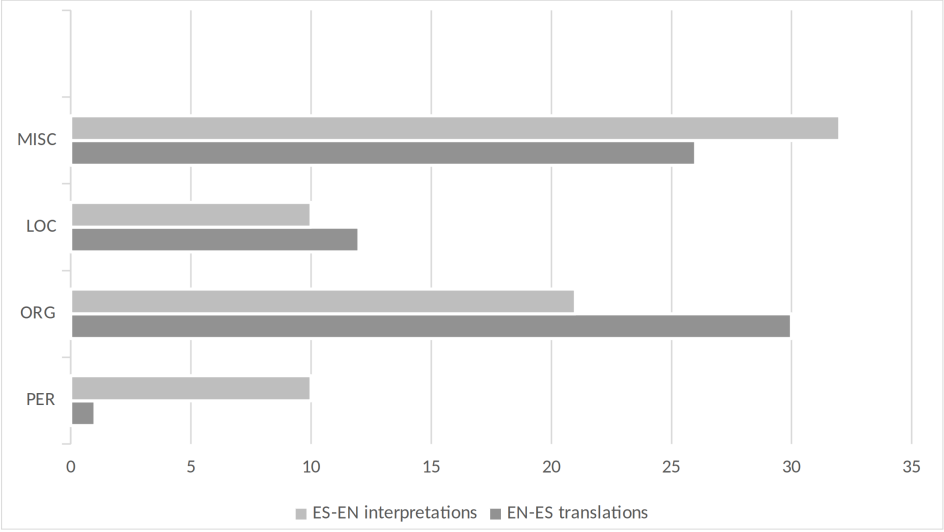


Figure 6: Type of shifts distribution.

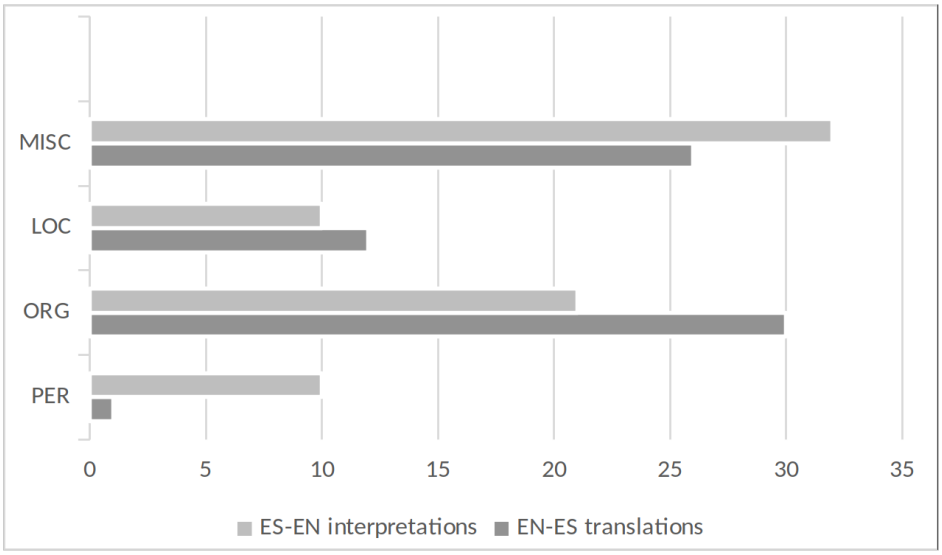


Figure 7: Shifted entities distribution.

Table 8: Detailed shift-entity relationship (EN-ES translations)

Type of shift	PER	ORG	LOC	MISC	Total
Register up	0	6	3	10	19
Register down	0	0	1	0	1
Contraction	0	4	2	9	15
Clarification	1	0	2	1	4
Expansion	0	9	1	4	14
Broadening	0	1	1	1	3
Partial transformation	0	0	0	1	1
Total transformation	0	1	0	0	1
Normalisation	0	9	2	0	11

Table 9: Detailed shift-entity relationship (ES-EN interpretations)

Type of shift	PER	ORG	LOC	MISC	Total
Register up	0	0	0	1	1
Register down	0	5	2	5	12
Contraction	4	2	1	3	10
Clarification	0	1	0	2	3
Expansion	1	0	1	0	2
Broadening	3	6	6	9	24
Partial transformation	1	1	0	6	8
Total transformation	1	1	0	6	8
Normalisation	0	5	0	0	5

In this comparison, major differences can be found which help characterising the shifting profile of each type of transfer separately. It appears that, when operating with named entities:

- English-Spanish translations tend to upgrade register (19), change meaning by contracting (15) and expanding (14), and to normalise multi-word terms (11).
- Spanish-English interpretations, contrarily, tend to downgrade register (12), change meaning by contracting (10) and broadening (24), and to present more transformations, be them partial (8) or total (8).

Table 10: Comparison of shift subcategories in both directions

Type of shift	EN-ES translation	ES-EN interpretation
Register up	19	1
Register down	1	12
Contraction	15	10
Clarification	4	3
Expansion	14	2
Broadening	3	24
Partial transformation	1	8
Total transformation	1	8
Normalisation	11	5

In general, the results show clear differences in the nature of shifts between EN-ES translations and ES-EN interpretations in the Petitions Committee. Translations from English into Spanish present more frequently register (e.g. *RAMSAR* → Convención de Ramsar)[26] and normalisation shifts (e.g. *Government of Valencia`s Ministry of Agriculture* → *Consejería de Agricultura de la Generalitat Valenciana*). In the case of register, practically all changes are upwards (*Bulgarian Ministry of Environment and Water* → *Ministerio Búlgaro de Medio Ambiente y Recursos Hídricos*), as opposed to the downward tendency of the shifts in the interpretations into English (*Comisión de Medio Ambiente del Parlamento Europeo* → ENVI Committee). The fact that Spanish translators tend to be more formal than English interpreters was a previous intuition confirmed by the data, similarly to the results obtained by Bernardini (2016: 143–144) in her comparative analysis of Italian-English translations. Moreover, results in normalisation bring a new perspective to previous studies, as this is a newly introduced shift category which focuses on changes in specialised multi-word terms instead of general-language collocations. In the case of Bernardini (ibid.), results showed an increased tendency by Italian-English translators to insert general language collocations. Our data show that normalisation of specialised phraseology is preferred when translating into the Romance language (Spanish) instead.

Moving on to quantitative meaning shifts, the interpretations present a slightly higher amount of them, although it must be specified that they are not of the same type in both directions and modes. While contraction and clarification are

[26]These examples were extracted from the most common NE categories in each case according to the correlation Table 8 and Table 9.

more or less equal, expansion prevails overwhelmingly in the EN-ES translations, as in the example: *Association for the Renaissance of Craiova (ARC)* → «Association for the Renaissance of Craiova » (ARC) (Asociación para el Renacimiento de Craiova). Inversely, broadening is much more numerous in the ES-EN interpretations (*nueve_mil_seiscientas hectáreas ilegales* → *considerable illegal construction*). Considering that broadening shifts could be regarded as a simplification feature, our results for the English-Spanish/Spanish-English pair are in line with the bidirectional English<>Italian study of Bernardini et al. (2016), in which interpreters were found to simplify the input more than translators.

Finally, transformations are substantially more present in the ES-EN interpretations, where they are equally distributed among partial (*Ley de Protección Integral del Mar Menor law for iterative protection of the Mar Menor*) and total (*Planes de Ordenación de los Recursos Naturales* → natural protection ehm plans). This is an interesting finding because it presents both similarities and divergences with previous intermodal studies. In Bernardini (2016), for example, transformations were also absent from English-Italian translations, but far more present in the other subcorpora, and the "partial" category outnumbered the "total" one. Although this could be the result of different conceptualisations by the researchers on what "transformation" means, it can also be argued that dissimilarities in transformational behaviour are connected to Ferraresi & Miličević's (2017: 1) "cognitive and task-related constraints" characterising the translation and interpreting processes. In other words, the number and nature of the transformations operated by the translator and/or interpreter could be strongly dependent on factors beyond language direction or mode, such as the communicative situation in which he/she is working (e.g., whether the context is a plenary session of the Parliament or a Committee meeting) or even the topic of the source text.[27]

Precisely with the goal of shedding some light on the connections between topic (or specialisation field, etc.) and the shifts involved in translation and interpreting, discussion should also centre on the shifted NEs label distribution shown in Figure 7. A clear majority of miscellaneous and organisational entities over locations and proper nouns of persons can be observed both in EN-ES translation and ES-EN interpreting. These results picture a cognitive domain of a rather political nature, in which parties, public platforms and similar organisations discussing policies and agreements are more important than the places where the problem occurred or the persons who complained in the first place. This perspective suits the function of the Committee on Petitions and points indeed towards

[27]These factors could also affect the degree of relation between total transformation shifts and translation/interpreting errors suggested in §3

a supranational way of making politics which permeates through the shifts encountered in translation and interpreting. What is more, a closer examination of Table 8 and 9 reveals that there is a high degree of relationship between the frequency of shifts and entities in both analysed directions and modes. For example, the frequent upward register shifts in EN-ES translation often occur in MISC NEs (*EU law* → *Derecho de la Unión*), and the numerous broadening shifts in the ES-EN interpretations are usually operated on ORG NEs (*departamentos de Ecología de la Universidad de Murcia* → the University of Murcia and its researchers). Introducing this new parameter in the analysis of shifts could add a new variant to the conclusions of Ferraresi et al. (2018) and lead us to hypothesise that simplification is a contingent feature which depends not only on the mediation mode and the source languages involved, but also on the topic of the source text. This is in line with calls for multifactorial research designs in empirical translation/interpreting studies (**CorpasPastor2008**, **DeSutterLefer2018**), since studies that take into account only one or two explanatory factors fall short of explaining the complexity of real-world translation/interpreting phenomena. Under this view, the analysed ES-EN interpretations of the Committee on Petitions would be more simplified than the EN-ES translations not just because they are an oral mediation performed into English, but also because the interpreters (consciously or not) would apply certain strategies aimed at approaching the content of their message to a broader audience than translation. This would imply neutralising or simplifying institutional-specific MISC and ORG NEs (EU legislation, international agreements, public bodies, etc.), paradoxically the most unfamiliar in the ears of the European citizens who could also exercise their right of petition.

7 Conclusion

The study presented in this chapter can be regarded as innovative for various reasons. To the best of our knowledge, it is one of the first corpus-based studies which relies on translated and interpreted documents from the European Parliament Committee on Petitions. Secondly, it does not only build and employ a type of resource which is still in its infancy (intermodal corpora), but also introduces a new methodological layer through manual and state-of-the-art automatic named-entity recognition (the latter performed by spaCy). This approach added new aspects to the analysis of translation and interpreting shifts (a new shift category called "normalisation" and the possibility of correlating shifts to the semantic labels of the NEs involved), which in turn helped establish interesting findings in relation with previous studies (normalisation as a language-dependent feature

of translation, transformation and simplification as contextual, topic-dependent features of interpreting).

Our study presents some limitations, though. Concerning methodology, the suitability of the selected transcription conventions must be revised. Even though we introduced certain modifications to the system, some of the proposed features seem more adequate for multimodal corpora and are counterproductive when recognising NEs (consider for example the hesitation particle *ehm* in *the Socialist for ehm Party ehm from the ehm Murcia region*). NE recognition and corpus-based shift analysis could also be extremely facilitated with the addition of an intermediate alignment phase to cope with terminology variation. In fact, monolingual terminological variation within NEs (e.g., *The Court, Court of Justice of the European Union*, *CJEU*, etc.) turned manual pairing into an exhausting job. As to NE labelling, a more fine-grained taxonomy is also needed for both languages, especially in the MISC category, where additional subtypes not available in the VIP scheme could be traced during the analysis (e.g. agreements like *Ramsar* or quasi-legal documents such as *Estrategia de Gestión Integrada*, among others). Undoubtedly, a tailor-made labelling system like this would considerably increase the quality of the correlating shift-entity results. In addition, the spaCy script integrated in the VIP NER module has been trained on two different language models, which could also account for the differences in precision and recall (685,000 word vectors in English as opposed to 500,000 word vectors in Spanish).

Finally, the NLP-enhanced orientation to the analysis of intermodal corpora presented in this chapter helped envisage a new line of research which does not hold translation and interpreting universals as an unconditional reality, but as a theoretical basis which is given in different degrees in the texts, depending on variants such as the languages and directions involved, the mode of mediation, and even the semantic content of the named entities conveyed. Therefore, multifactorial research designs are needed to capture the multitude of factors that have an influence on the observed phenomena. Although more studies are needed to determine the exact relevance of these semantic categories in translation and interpreting shifts, it would seem that the final goal is finding a transversal set of norms which could break the theoretical differences between translation and interpreting, focussing the discussion on the coordinates or *function* of the mediation instead of the mediating mode itself.

Acknowledgements

The research reported in this article has been funded by the Spanish Ministry of Education and Professional Training (ref. FPU18/05803). The article has also

been carried out in the framework of the projects VIP (FFI2016-75831-P), TRIAGE (UMA18-FEDERJA-067) and MI4ALL (CEI-RIS3).

References

Baisa, Vít, Jan Michelfeit, Marek Medved & Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2799–2803.

Bakker, Matthijs, Cees Koster & Kitty Van-Leuven Zwart. 2009. Shifts. In *Routledge encyclopedia of translation studies*, 269–274. London: Routledge.

Bendazzoli, Claudio. 2010. The European Parliament as a source of material for research into simultaneous interpreting: Advantages and limitations. In Lew N. Zybatow (ed.), *Translationswissenschaft: Stand und Perspektiven*, 51–68. Frankfurt am Main: Peter Lang.

Bernardini, Silvia. 2016. Intermodal corpora: A novel resource for descriptive and applied translation studies. In Gloria Corpas & Miriam Seghiri (eds.), *Corpus-based approaches to translation and interpreting: From theory to applications*, 129–148. Frankfurt: Peter Lang. DOI: 10.3726/b10354.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC |Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, vol. 1 (New Frontiers in Translation Studies), 21–42. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-6199-8_2.

Corpas Pastor, Gloria. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS. Revista de Traductología* 5. 155–184. DOI: 10.24310/trans.2001.v0i5.2916.

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* 0(0). 1–23. DOI: 10.1080/0907676X.2019.1611891.

European Parliament. 2020a. *European Parliament Committees: About PETI.* https://www.europarl.europa.eu/committees/en/peti/about.

European Parliament. 2020b. *Petitions FAQ*. https://petiport.secure.europarl.europa.eu/petitions/en/faq/det?questionor=1&sectionor=1.

European Union. 2012. *Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union: charter of fundamental rights of the European Union*. Publications Office. DOI: 10.2860/58644.

Ferraresi, Adriano, Silvia Bernardini, Maja Milicevic Petrovic & Marie-Aude Lefer. 2018. Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta : Journal des traducteurs / Translators' journal* 63. 717–738. DOI: 10.7202/1060170ar.

Ferraresi, Adriano & Maja Miličević. 2017. Phraseological patterns in interpreting and translation. Similar or different? In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies. New methodological and theoretical traditions*, 157–182. Berlin: Mouton/De Gruyter. DOI: 10.1515/9783110459586-006.

Gaber, Mahmoud, Gloria Corpas Pastor & Ahmed Omer. 2020. Speech-to-Text technology as a documentation tool for interpreters: A new approach to compiling an ad hoc corpus and extracting terminology from video-recorded speeches. *TRANS. Revista de Traductología* 5. 263–281. DOI: 10.24310/TRANS.2020.v0i24.7876.

Goffin, Roger. 1994. L'eurolecte: Oui, jargon communautaire: Non. *Meta: Journal des traducteurs* 39(4). 636–642. DOI: 10.7202/002930ar.

Hurtado, Amparo. 2017. *Traducción y traductología: Introducción a la traductología*. 9th. Madrid: Cátedra.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, vol. 5, 79–86. Phuket: AAMT. http://mt-archive.info/MTS-2005-Koehn.pdf.

Martin, Anne. 1997. *Tratamiento de las referencias de carácter institucional del mundo de habla inglesa en la prensa española*. Universidad de Granada. (Doctoral dissertation).

Martínez, José & Iris Serrat. 2012. ECPC: El discurso parlamentario europeo desde la perspectiva de los estudios traductológicos de corpus. *Linguamática* 4(2). 65–73.

Mayoral, Roberto. 1999. La traducción de referencias culturales. *Sendebar: Revista de la Facultad de Traducción e Interpretación* 10-11. 67–88.

Mori, Laura (ed.). 2018. *Observing Eurolects: Corpus analysis of linguistic variation in EU law* (Studies in Corpus Linguistics 86). Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/scl.86.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London; New York: Routledge.

Ortega, Juan Miguel. 2002. La traducción de referencias culturales de carácter institucional y político a través de un caso práctico. *Puentes* 1. 21–32.

Prieto Ramos, Fernando (ed.). 2019. *Institutional translation for international governance: Enhancing quality in multilingual legal translation* (Bloomsbury Advances in Translation Studies). New York: Bloomsbury Publishing.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli & Nicoletta Spinolo. 2012. The European Parliament Interpreting Corpus (EPIC): implementation and developments. In Francisco Straniero Sergio & Caterina Falbo (eds.), *Breaking ground in corpus-based interpreting studies*, 53–90. Bern: Peter Lang. DOI: 10.3726/978-3-0351-0377-9.

Shlesinger, Miriam. 2008. Towards a definition of interpretese: An intermodal, corpus-based study. In Gyde Hansen, Andrew Chesterman & Heidrun Gerzymisch-Arbogast (eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile*, 237–253. Amsterdam: John Benjamins. DOI: 10.1075/btl.80.18shl.

Ustaszewski, Michael. 2019. Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives: Studies in Translation Theory and Practice* 27(1). 107–123. DOI: 10.1080/0907676X.2018.1485716.

Veroz González, María Azahara. 2014a. El Registro Público de Documentos del PE como recurso documental en la traducción especializada: Elaboración de bases de datos terminológicas con corpus en Multiterm. *Hikma* 13. 125. DOI: 10.21071/hikma.v13i.5229.

Veroz González, María Azahara. 2017. Translation in the European Parliament: The study of the ideational function in technical texts (EN/FR/ES). *Meta* 62(1). 19–44. DOI: 10.7202/1040465ar.

Veroz González, María Azahara. 2014b. *La traducción en el Parlamento Europeo. Estudio de los textos técnicos y de comunicación administrativa*. Universidad de Córdoba. (Doctoral dissertation).

Empirical investigations into the forms of mediated discourse at the European Parliament

The purpose of this book is to showcase a diverse set of directions in empirical research on mediated discourse, reflecting on the state-of-the-art and the increasing intersection between Corpus-based Interpreting Studies (CBIS) and Corpus-based Translation Studies (CBTS). Undeniably, data from the European Parliament (EP) offer a great opportunity for such research. Not only does the institution provide a sizeable sample of oral debates held at the EP together with their simultaneous interpretations into all languages of the European Union. It also makes available written verbatim reports of the original speeches, which used to be translated. From a methodological perspective, EP materials thus guarantee a great degree of homogeneity, which is particularly valuable in corpus studies, where data comparability is frequently a challenge.

In this volume, progress is visible in both CBIS and CBTS. In interpreting, it manifests itself notably in the availability of comprehensive transcription, annotation and alignment systems. In translation, datasets are becoming substantially richer in metadata, which allow for increasingly refined multi-factorial analysis. At the crossroads between the two fields, intermodal investigations bring to the fore what these mediation modes have in common and how they differ. The volume is thus aimed in particular at Interpreting and Translation scholars looking for new descriptive insights and methodological approaches in the investigation of mediated discourse, but it may be also of interest for (corpus) linguists analysing parliamentary discourse in general.