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

Marta Kajzer-Wietrzny[a], Adriano Ferraresi[b], Ilmari Ivaska[c] & Silvia Bernardini[b]

[a]Adam Mickiewicz University [b]University of Bologna [c]University of Turku

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 2020). 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 to investigate 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 transcribe data and account for spoken language-specific traits. At the beginning,

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.), *Mediated discourse at the European Parliament: Empirical investigations*, iii–xi. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977036

interpreting corpora were mostly comparable. Today, most of them are (also) parallel and aligned at sentence level, with a few also including alignment with corresponding translated texts and original videos (Ferraresi & Bernardini 2019), or even sound-to-text alignment at word level (cf. Chmiel et al. 2022 [this volume]). Scholars compiling their corpora make use of such technological advancements as speech recognition to speed up the transcription process like in the European Parliament Translation and Interpreting Corpus (EPTIC), the Polish Interpreting Corpus (PINC) or the PETIMOD corpus (Ferraresi & Bernardini 2019; Koržinek & Chmiel 2021; Pastor & Rodas 2022 [this volume]) or speaker identification to disambiguate interpreter voices, e.g., in PINC (Koržinek & Chmiel 2021; Chmiel et al. 2022 [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 (2020: 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 and task expertise. This approach aims at shedding light on which linguistic features typically associated with translation may result from bilingual activation in general (as opposed to monolingual language production), or from the process of reworking a text (as opposed to producing it anew). From a more sociolinguistic and discourse-related perspective, translation and interpreting scholars have also made use of corpus methods to explore 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 their multilingual nature, European Parliament (EP) 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: these are spoken, written and re-written for various recipients in diverse forms depending on the communicative goal. The activities connected with the EP plenaries involve Members of the Parliament (MEPs) delivering a speech either impromptu or upon earlier preparation, which is usually based on existing documentation at various stages of completion. 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, EP data constitute a valuable source of texts that in Kotze's (2020: 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 EP 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 EP 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 delivered and interpreted, and are only later 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, Defrancq and Plevoets examine speeches delivered by MEPs and their interpretations. After theoretical considerations on whether MEPs have more expertise in the genre of EP plenary speeches than interpreters or the other way round, the empirical part concentrates 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 EP 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 3, by Chmiel, Korzinek, Kajzer-Wietrzny, Janikowski, Jakubowski and Polakowska, introduces PINC, a corpus of European Parliament Polish 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 4 in the volume adopts a more qualitative approach to address sensitive, and hence challenging issues 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 5, 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 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 6, 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 7, 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 8 by Corpas Pastor and Sánchez Rodas presents an NLP-enhanced analysis of shifts in the rendition of named entities in an English<>Spanish subcorpus of PETIMOD, the translation and interpreting corpus of the Committee on Petitions of the EP. The outcomes suggest that tendencies such as normalisation, transformation and simplification depend on language direction, mediation mode, and 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 dominant research avenues also 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 (Reynaert et al. 2021). 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 intermodal analyses, collecting and pre-processing the required amount of data lies beyond the capacity of single scholars. 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 EPIC-UdS corpus (Przybyl et al. 2022 [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*, 149–161. https : / / www . euroconferences . info / proceedings / 2005 _ Proceedings/2005_proceedings.html.

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.

Castagnoli, Sara. 2020. Translation choices compared: Investigating variation in a learner translation corpus. In Sylviane Granger & Marie-Aude Lefer (eds.), *Translating and comparing languages: Corpus-based insights*, 25–44. Louvain-la-Neuve: Presses universitaires de Louvain.

Chmiel, Agnieszka, Danijel Koržinek, 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.), *Mediated discourse at the European Parliament: Empirical investigations*, 63–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977042.

Collard, Camille & Bart Defrancq. 2016. How to use a nanocorpus. Enriching corpora of interpreting. Paper presented at Corpus Linguistics in the South 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. 2020. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28(1). 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*, 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*, 123–139. Amsterdam/Philadelphia: John Benjamins.

Kajzer-Wietrzny, Marta. 2012. *Interpreting universals and interpreting style*. Adam Mickiewicz University, Poznań. (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.

Koržinek, Danijel & Agnieszka Chmiel. 2021. Interpreter identification in the Polish Interpreting Corpus. *Revista Tradumàtica* 19. 276–288.

Kotze, Haidee. 2020. 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: Routledge.

Kruger, Haidee & Bertus van Rooy. 2010. The features of non-literary translated language: A pilot study. In Richard Xiao (ed.), *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS)*.

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* 50(4). 141–158. DOI: 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.

Pastor, Gloria Corpas & 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.), *Mediated discourse at the European Parliament: Empirical investigations*, 219–251. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977052.

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

Przybyl, Heike, 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.), *Mediated discourse at the European Parliament: Empirical investigations*, 191–218. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977050.

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* 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 Carol Taylor Torsello, Katherine Ackerley & Erik Castello (eds.), *Corpora for university language teachers*. Bern: Peter Lang.