From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective

Published Version:
Bernardini, S., Ferraresi, A., Miličević, M. (2016). From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. TARGET, 28(1), 61-86 [10.1075/target.28.1.03ber].

Published:
DOI: http://doi.org/10.1075/target.28.1.03ber

Bernardini, S., Ferraresi, A. and Miličević, M. 2016. "From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective". *Target* 28, 1: 61-86.

The final published version is available online at: **http://dx.doi.org/10.1075/target.28.1.03ber**

From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective[1]

Abstract

This article introduces EPTIC (the European Parliament Translation and Interpreting Corpus), a new bidirectional (English<>Italian) corpus of interpreted and translated EU Parliament proceedings. Built as an extension of the English<>Italian subsection of EPIC (the European Parliament Interpreting Corpus), EPTIC is an *intermodal* corpus featuring the pseudo-parallel outputs of interpreting and translation processes, aligned to each other and to the corresponding source texts (speeches by MEPs and their written up versions). As a first attempt at unearthing the potential of EPTIC, we investigate lexical simplification replicating the methodology proposed by Laviosa (1998a, 1998b), but extending it to encompass both a monolingual comparable and an intermodal perspective. Our results indicate that the mediation process reduces complexity in both modes of language production and both language directions, with interpreters simplifying the input more than translators, and evidence of simplification being more lexical in English and more lexico-syntactic in Italian.

Keywords: corpus-based approach, intermodal corpora, interpreting, translation, lexical simplification, English, Italian

1. Introduction

Translation and interpreting studies have largely developed independently of each other, the former being concerned mainly with linguistic, literary and sociocultural perspectives and the latter focusing more on psychological (and more recently social/relational) aspects (see e.g. the various articles in Schäffner 2004). However, it has also been pointed out, for instance by Gile (2004, 10) that they "share epistemological, methodological, institutional and wider sociological concerns" and that "[i]t, therefore, makes much sense for both disciplines to work together in spite of the differences".

In both translation and interpreting studies, the corpus-based approach has attracted substantial interest for over two decades, ever since the publication of Mona Baker’s (1993) seminal paper concerning translation and the corresponding proposal made by Miriam Shlesinger (1998) with regard to interpreting. Within the corpus-based approach, translated and interpreted outputs are typically compared either to their source texts (within *parallel* corpora), or to comparable original texts constituted by written or oral non-mediated production (*comparable* corpora); the former are needed if the goal is to discern local differences between specific source and target texts, while the latter have mostly been employed to study patterns typical of translated/interpreted texts in general, e.g. when looking at translation/interpreting universals.

In line with the interest in investigating the common ground between translation and interpreting as two different *modes* or *modalities* of translation broadly conceived, and thanks to the shared methodology, some researchers have recently started to investigate the potential of *intermodal* corpora, i.e. corpora containing parallel or comparable outputs of translation and interpreting. However, while extremely rewarding in terms of the insights they can offer, intermodal corpora have proved challenging to construct due to the shortage of texts that are both translated and interpreted in authentic settings.

Our first aim in this article is to describe our attempt at building an intermodal corpus out of what is probably the largest and most accessible source of interpreted and translated texts, namely the European Parliament plenary sessions. Specifically, we detail the steps in the construction of EPTIC, the European Parliament Translation and Interpreting Corpus, which consists of independently produced translational and interpretational outputs (into English and Italian) based on input from the European Parliament sessions, as well as the input source texts themselves (in Italian and English respectively). Our second aim is to offer an example of how EPTIC can be used to investigate lexical simplification, one of the purported translation/interpreting universals, bidirectionally and adopting a combined intermodal and comparable perspective. Our results show that interpreted language in EPTIC is simpler than translated language, that mediated language is simpler than non-mediated language, and that parameters of simplification apply differently to different languages – English simplifying at the lexical level, and Italian at the lexico-syntactic one.

The paper is structured as follows. We first outline the background on intermodal corpora and lexical simplification. The process of building EPTIC is explained next, after which we move on to describe our case study, providing a detailed account of the method used. Following the presentation and discussion of our main findings, we end by making suggestions for future research based on EPTIC and sketching plans for its further development.

2. Background: intermodal corpora and lexical simplification

2.1 Intermodal corpora for comparing interpreting and translation

In contrast to earlier work, the last decade has witnessed a growing interest in the comparison of translation and interpreting, on the assumption that both (sub)disciplines can profit from the synergy. In the words of Shlesinger and Ordan (2012, 44):

> [...] translation scholars can learn about the process and product of (written) translation by finding out more about interpreting – and interpreting scholars can infer about this high-pressure form of translation by observing the slower, more readily observable process and product of (written) translation.

As part of the shared ground between interpreting and translation studies, Gile (2004, 27) mentions the goal of target text description and the comparison of source and target text, while Pöchhacker (2004, 115) refers specifically to corpus linguistics as a paradigm "initially explored by translation scholars and subsequently applied to conference interpreting corpora", found to be enriching to both (sub)disciplines.

The move toward an integration of translation and interpreting corpora can be traced back to Miriam Shlesinger. In her early work, the idea is put forward that comparable corpora in interpreting studies should be extended to cover not only interpreted texts and original oral discourses delivered in similar settings, but also written translations of such texts (Shlesinger 1998, 488). However, as this ideal is difficult to attain, given the inherent difficulties of finding authentic settings in which texts are both translated and interpreted, in more recent work Shlesinger (2009) uses the term *comparable intermodal corpora* to describe corpora formed of target texts resulting from different translation modalities, not necessarily sharing the same source text. She further notes that intermodal corpora can be composed either of authentic translations/interpretations, or else of translations/interpretations produced under experimental conditions. Resorting to experimental data, while justified in light of the difficulties of finding real-life settings in which texts are both translated and interpreted, and coherent with the interpreting studies tradition reviewed by Gile (2004), is clearly at odds with the corpus orthodoxy, which requires that corpora contain authentic texts (see e.g. the definition in McEnery et al. 2006, 5). We thus argue that the ideal comparable intermodal corpus is still the one imagined by Shlesinger in 1998, i.e. one that features collections of *authentic* interpreted texts, *authentic* written translations of the same texts, and *authentic* original discourses (on similar topics and) delivered in similar settings.

To the best of our knowledge, three attempts have been made so far at constructing intermodal corpora, prior to our own. Shlesinger (2009) constructed a small-scale monolingual intermodal corpus comprising experimental data, i.e. the interpretational and translational outputs (English>Hebrew) of the same written input by six professional translators/interpreters, who first rendered it orally and then, three years later, in writing. The corpus size is 8,327 tokens for simultaneous interpreting and 8,968 tokens for written translation. As a follow-up to this study, Shlesinger and Ordan (2012) constructed a larger monolingual comparable and intermodal corpus in the academic domain (about 24,000 tokens per subcorpus), comprising authentic translational and interpretational English>Hebrew output as well as spontaneous Hebrew speeches in the same domain. This corpus allowed the authors "to explore (again) differences between the oral and written modalities of translation, [and] to observe the effects of the ontology variable (original vs. translated) as well" (Shlesinger and Ordan 2012, 47). Both corpora support intermodal comparisons, since they make available the outputs of translation and interpreting processes, but they differ in the nature of the relationship between the two mediated subcorpora. While the former includes interpreted and translated versions of the same source text, forming a sort of intermodal-*parallel* corpus, the discourses and the texts in the latter are linked though an intermodal-*comparable* bond, i.e. they simply belong to (approximately) the same domains, but were produced independently of each other.

Thirdly, the Translation and Interpreting Corpus (TIC) described in Kajzer-Wietrzny (2012) is a monolingual comparable and intermodal-parallel corpus also based on the European Parliament plenary sessions. It contains texts in English interpreted and translated from French, Spanish, German and Dutch, as well as texts originally produced in English, for a total of 10 subcorpora (5 written and 5 spoken, 8 mediated and 2 non-mediated) and an overall size of over 500,000 words. TIC does not include the source speeches, nor any alignment linking up translated and interpreted outputs.

2.2 Lexical simplification: translation, interpreting and intermodal perspectives

Lexical simplification is one of the core universals proposed by Baker (1993) and subsequently investigated in both translation and interpreting settings. Within the monolingual comparable corpus approach, lexical simplification can be defined as the hypothesized tendency for translators and interpreters to produce texts that are less informationally dense and less lexically varied than those produced by people writing or uttering their own texts or speeches in the same language, in similar circumstances. For practical purposes, lexical simplification has usually been operationalized through the following parameters:

- lexical density: the percentage of lexical versus grammatical words in a text;
- type-token ratio: the ratio between the number of different words and the total number of words in a text;
- list head coverage: the percentage of corpus covered by the *n* top words of its frequency list;
- core vocabulary coverage: the percentage of text covered by *n* most frequent words of the given language, established on the basis of a reference corpus;
- mean sentence length: the average number of words per sentence in a text.

While these parameters are clearly an approximation that cannot hope to do justice to the complexity of the notion of simplicity – for sure “the concept of ‘plain and simple’ is itself very far from being plain and simple”, Halliday and Mathiessen (2004, 654) –, they do provide a methodological point of reference. First proposed in two seminal articles by Laviosa (1998a, 1998b), they have since been replicated in several works.

Laviosa first applied them to subsets of the English Comparable Corpus (ECC), a two-million-word monolingual corpus comprised of texts translated from different source languages (fiction and newspaper articles) and similar texts originally produced in English. The results of her studies suggest that newspaper articles translated into English are lexically simpler than comparable non-translated texts, irrespective of the source language, in terms of lexical density, list heads, core vocabulary, mean sentence length, and variance in these measures; some of these regularities hold true across different genres, being confirmed by an analysis of narrative texts (lexical density, list heads and core vocabulary), while others do not (sentence length and variance). More detail on these parameters is provided in Section 4.1.2.

Laviosa's parameters of simplification were developed with (written) translation in mind. Applying them to interpreting, and even more so to the comparison of the interpreted and translated modalities, raises issues about their relevance to spoken discourse. However, several empirical studies of lexical simplification adopting Laviosa's methodology have been conducted on interpreted language. On the basis of the EPIC corpus (a description of EPIC can be found in Section 3, as well as the works cited in this section), Sandrelli and Bendazzoli (2005) investigate lexical density and lexical variety in two of its monolingual comparable subcorpora, i.e. Italian original speeches vs. interpretations from English/Spanish and English original speeches vs. interpretations from Italian/Spanish. Their results indicate slight (significance testing is not performed) and inconsistent differences in lexical density in the various interpreting directions, such that a clear pattern does not emerge in this case, contrary to Laviosa. As for lexical variety, list heads of speeches interpreted into English display less variety than those of original English speeches (in line with Laviosa), but the opposite is true of Italian, possibly indicating a directionality effect. Focusing on the Spanish monolingual comparable subcorpus of EPIC, Russo et al. (2006, in Sandrelli et al. 2010) find that the speeches interpreted into Spanish from Italian and English have a higher lexical density than those originally delivered in Spanish, contrary to expectations (and Laviosa's studies). Results of the list head comparisons on the other hand confirm Laviosa's, with Spanish interpreted speeches displaying a higher percentage of high-frequency words than original speeches, i.e. less lexical variety (and consequently more simplification). In these studies, as well as in Sandrelli et al. (2010), the authors refer to Laviosa's results and conclude that, disregarding some inconsistencies that may be due to differences in the size of the various subcorpora, the pattern emerging might be in keeping with the one proposed by Shlesinger (1989, in Shlesinger 2009, 241) whereby "interpreting exerts a levelling effect: oral texts become more literate, literate texts become more oral."

A similar replication of Laviosa's methodology is found in Kajzer-Wietrzny (2012), who looks at three indicators: lexical density, core vocabulary and list heads. Whilst TIC as a whole is intermodal, for this study only the oral portion of the corpus is examined. Lexical density and the proportion of core vocabulary point to interpreted discourse not being simplified at all compared to original speeches. The only parameter that does, to some extent, reflect simplification in interpreted texts is that of list heads. Kajzer-Wietrzny finds that this result may in fact be specific to certain language combinations: interpretations from Spanish, German and Dutch (but not from French) in fact display higher list head coverage than originals. Also, it may be affected by the mode of delivery of the source texts (unscripted or semi-scripted versus read out): this parameter, examined on the Spanish>English subcorpus alone, shows that interpretations of un/semi-scripted texts once again have higher list head coverage than the originals as well as the read-out texts. On the basis of these results concerning list heads, Kajzer-Wietrzny concludes that the "tendency to repetitiveness, or the lack of it, is heavily contingent on the source language and subject to interference" (2012, 122). On the other hand, she argues that the increased lexical density of interpreted texts may be a result of condensation techniques used to save time, with shifts from referential to lexical cohesion possibly at play too, as simultaneous interpreters may tend to substitute pro-forms or ellipsis used in the original text with lexical items, which are either repetitions or synonyms.

Even though we are not aware of any intermodal studies prior to ours that attempt to replicate Laviosa’s approach, studies by Shlesinger (2009) and Shlesinger and Ordan (2012) investigated parameters closely related to lexical simplification in intermodal corpora. Specifically, Shlesinger (2009) compares lexico-grammatical features of interpreting and translation in her experimental corpus containing translational and interpretational outputs (English>Hebrew) of the same source text by six professional translators/interpreters. Even though her focus is not specifically on simplification, some of her findings are relevant in this respect. For instance, she finds that the type-token ratio is higher in written translations both on average, and for each of the six subjects taken individually. While this result would point to interpreting being more lexically simple than translation, Shlesinger questions whether this is indeed a distinguishing feature of interpreting (vs. translation), rather than orality (vs. writing). This issue is subsequently addressed by Shlesinger and Ordan (2012), who look at a monolingual comparable and intermodal corpus and find a statistically significant difference in type-token ratio between interpreted texts on the one hand and both translated and spontaneous oral data on the other. For lexical density such a difference only exists between the interpreted and spontaneously produced oral texts. Based on these and other findings (not related to lexical simplification), the authors argue that "one may see interpreting as, in a sense, an extreme case of translation, one in which those features that have been found to distinguish between translated and original texts [...] are found to be all the more salient" (Shlesinger and Ordan 2012, 54).

Summing up, the results of studies focusing on interpreting from a monolingual comparable perspective paint a more blurred picture than the one emerging from similar studies based on translated texts. Lexical simplification in interpreted vs. non-interpreted discourse is found to be dependent on the language combination/direction, and possibly also on the mode of delivery of the source text, thus hardly universal. From an intermodal perspective, some evidence exists to the effect that interpreted texts are more simplified than comparable translated texts; however, findings pointing to this conclusion mostly come from small-to-medium sized corpora and are prevalently based on a single lexical measure, the type-token ratio. The seemingly contradictory results – interpretations are less simplified in comparison to oral originals than translations in comparison to written originals, but more simplified when directly compared to translations – once again show that, despite some important efforts, what still appears to be missing is a large enough resource combining intermodality, monolingual comparability, and multi-directionality.

3. Introducing EPTIC: Construction and use

3.1 From EPIC to EPTIC

EPTIC is an intermodal corpus adding a translational component to the well-known EPIC, the European Parliament Interpreting Corpus. Given that EPIC itself, and details about its design and construction, are described in detail in several works (see in particular Sandrelli and Bendazzoli 2005, Bendazzoli 2010), in this section we focus on the expansion process, singling out the features of EPIC that are particularly relevant for the creation of EPTIC.

3.2 Corpus design and data collection

Despite belonging to a specialized domain and not being fully representative of the general language, the European Parliament materials have been widely used in corpus/computational linguistics, due to their authenticity, homogeneity, availability, and the number of languages represented. Importantly for translation and interpreting studies, the institutional setting also guarantees that translators and interpreters involved are accredited professionals (Vuorikoski 2004).

EPIC is one of a series of corpora derived from the European Parliament documents (cf. Europarl, Koehn 2005; TIC, Kajzer-Wietrzny 2012). It is a trilingual (English<>Italian<>Spanish) corpus of European Parliament speeches and their corresponding interpretations delivered at the Parliament's part-session held in February 2004. The speeches and interpretations were recorded off the news channel Europe by Satellite, and subsequently digitized, transcribed, and turned into a machine-readable corpus.[2] Both source and target texts are included, making for a total of nine subcorpora, three of source speeches, and six of interpreted speeches.

The European Parliament materials also offer a unique opportunity for constructing intermodal corpora. For each plenary session the Parliament publishes ‘verbatim’ reports of proceedings, i.e. written versions of the speeches, as well as their translations into all EU official languages. The reports are first published in a provisional version, soon followed by a final one that carries the indication ‘revised version’. Despite being called ‘verbatim’, these reports are at times substantially edited. The changes include the addition of punctuation, the correction of mistakes such as false starts, unfinished sentences or mispronunciations, and the removal of context-related comments on the part of the speakers (see Table 1). The amount of intervention differs from text to text, depending mainly on the mode of delivery of the speech, with read-out texts undergoing smaller changes than the unscripted ones. As concerns translations of the proceedings, several EU officials consulted on this issue confirmed that they result from an independently performed translation process based on the revised verbatim reports, without any reference to the interpreters’ outputs.

Table 1. A fragment from a transcribed speech and the corresponding verbatim report

Transcript of the original speech	**Verbatim report**
`ehm is the microphone wor-. well thank you ehm President and ehm welcome to our Commissioner David Byrne. I'm deeply concerned about some of the issues that have been raised today in this meeting.`	`Mr President, I welcome our Commissioner, Mr Byrne. I am deeply concerned about some of the issues that have been raised in this meeting today.`

In order to create the oral/interpreted components of EPTIC, we obtained the transcriptions of interpreted talks and of their source texts from EPIC, including at this stage only speeches in English and Italian; for the written/translated components, we downloaded from the European Parliament website the final (‘revised’) versions of the verbatim reports of proceedings and their translations.[3] EPTIC is therefore a bilingual and bidirectional corpus (English<>Italian). Considering all its subcorpora, comprising simultaneous interpretations paired with their source texts, plus corresponding translations and source texts (a total of eight components), it can be classified as an intermodal, comparable and parallel corpus. The structure of EPTIC is shown in Figure 1.

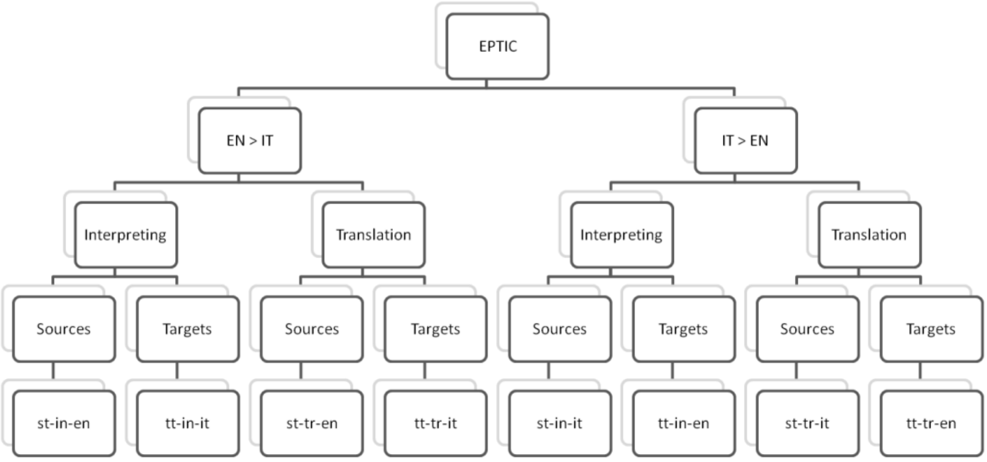

Figure 1. EPTIC corpus structure: the st- and tt- prefixes indicate source and target texts, the -in- and -tr- affixes interpretations and translations.

The total size of EPTIC is just above 175,000 words (disregarding truncated words in interpreted texts), and the sizes of individual subcorpora are as shown in Table 2. The difference in size between the English>Italian and the Italian>English parts is substantial, but at this stage there was no attempt to correct this imbalance, to preserve comparability with EPIC.

Table 2. Sizes of EPTIC subcorpora

Subcorpus	**N. of texts**	**Total word count**	**% of EPTIC**
st-in-en	81	41,869	23.91
st-tr-en	81	36,685	20.95
tt-in-it	81	33,675	19.23
tt-tr-it	81	36,876	21.06
Subtotal	**324**	**149,105**	**85.14**
st-in-it	17	6,387	3.65
st-tr-it	17	6,234	3.56
tt-in-en	17	6,577	3.76

tt-tr-en	17	6,819	3.89
Subtotal	**68**	**26,017**	**14.86**
TOTAL	**392**	**175,122**	**100.00**

3.3 Text pre-processing

All files were saved in plain text format. Since the main aim of EPTIC is to allow comparisons between interpreted and translated texts, rather than focus on the specific features of either modality, we further discarded information on mispronounced words (in oral/interpreted texts), and emphasis as signalled by italics (in written/translated texts).

As for metadata, we reproduced those in EPIC, modifying the structure of the text headers to make it applicable both to the oral/interpreted texts and to the written/translated ones. Headers in EPTIC thus have the same structure across all its components, but depending on the subcorpus different metadata may or may not be available. The metadata available for the different subcorpora are shown in the Appendix.

3.4 Corpus processing

To enable more advanced automatic analyses, EPTIC was tagged for parts of speech (POS). Even though EPIC is POS-tagged as well, for reasons of consistency across all corpus components, linguistic mark-up in EPTIC was performed independently. All subcorpora were tagged and lemmatized using the TreeTagger,[4] and indexed with the Corpus WorkBench (CWB).[5] Thanks to corpus indexing, metadata can be used to perform complex queries based on specific characteristics of the texts and the speakers who delivered them. In the case of target texts, searches can also be restricted on the basis of the respective source texts (e.g. only interpretations of unscripted speeches), and, for interpreted texts, of interpreters' characteristics (e.g. only interpretations by male/female interpreters; unfortunately no corresponding information was available for translators).

In addition, bidirectional sentence-level alignment was performed using CWB's built-in aligner, with no manual correction, both for parallel (source-target) and intermodal (translation-interpretation) pairs. Alignment is a new feature compared to EPIC, whose aligned version is planned for the next stage of corpus development (see Bendazzoli 2010, 134). The output of a command line query to the aligned corpus (in this case for the lemma 'activity') returns concordances in the format shown in Table 3.

Table 3. Parallel concordances from EPTIC (command-line interface)

`269:`	`And then there's a second aspect which concerns`

	`the Commission's <activities> we know ehm Commissioner Lamy told us`
`-->tt-tr-en:`	`The second aspect concerns the Commission's work. As Commissioner Lamy announced,`
`-->st-in-it:`	`Poi c'è un secondo aspetto che riguarda l'attività della Commissione // Noi sappiamo ce l'ha annunciato il Commissario Lamy`
`-->st-tr-it:`	`Il secondo aspetto riguarda l'attività della Commissione. Come annunciato dal Commissario Lamy`

4. Exploring simplification in mediated discourse through EPTIC

4.1 Method

4.1.1 *Directions of comparison: monolingual comparable and intermodal perspectives*

To illustrate the potential of EPTIC, two complementary perspectives are adopted. Adopting an intermodal perspective, we contrast the interpreted and translated English subcorpora on the one hand, and the interpreted and translated Italian subcorpora on the other. When significant results are obtained for this dimension, a set of monolingual comparable comparisons are carried out, contrasting the translated and non-translated subcorpora on the one hand, and the interpreted and non-interpreted subcorpora on the other. The ultimate aim is to visualize how instances of oral and written, mediated and non-mediated discourse position themselves with respect to each other in terms of simplification parameters.

Concerning the more novel, intermodal analyses, it should be noted that the comparisons carried out are monolingual *near*-parallel, since they apply to interpretations and translations of spoken and written sources that are very closely related (i.e., recorded speeches and their written up versions). These sources, making up the source text subcorpora, are checked to control for unrelated source text variables. If significant differences are found both between the target texts and between the corresponding source texts, then major editorial changes must have been made to the original speeches when turning them into written texts. In such cases, the differences found between target texts may be the result of differences in their sources (and thus irrelevant to the interpreted/translated opposition), and are not considered here.[6]

This approach means that the parameters of simplification described in 4.1.2 below are tested in a total of 6 sets of comparisons, namely:

1a. Interpreted vs. translated English (monolingual parallel)
1b. Benchmark: Italian sources of interpreted English vs. Italian sources of translated English
2a. Interpreted vs. translated Italian (monolingual parallel)
2b. Benchmark: English sources of interpreted Italian vs. English sources of translated Italian
3. Interpreted vs. non-interpreted English (monolingual comparable)
4. Interpreted vs. non-interpreted Italian (monolingual comparable)
5. Translated vs. non-translated English (monolingual comparable)
6. Translated vs. non-translated Italian (monolingual comparable)

4.1.2 *Simplification parameters and statistical testing*

Our study aims to replicate as closely as possible the method adopted by Laviosa in her 1998 studies of simplification in newspaper and fiction texts, distancing itself only when the new research setup or methodological advances in the discipline impose it. We examine four simplification parameters: lexical density, mean sentence length, core vocabulary coverage and list head coverage. We ignore the relative percentage of present and past auxiliary forms, a marginal feature that was found by Laviosa to be specific to newspaper language, and is thus of limited relevance here. All measures, except for list head coverage, are computed on a single text basis, to account for variability among texts within the different subcorpora. Given that the measures we look at are influenced by text size, by-text analyses also ensure the validity of comparisons between subcorpora of different sizes.

Following Laviosa (1998a, 1998b), who in turn refers back to Stubbs (1996, 172), lexical density is defined as the proportion of lexical to function words, and calculated "by subtracting the number of function words […] from the number of running words (which gives the number of lexical words) and then dividing the result by the number of running words" (Laviosa 1998a, note 3). Running words are counted through a regular expression matching sequences of numbers and/or letters, which may include apostrophes and hyphens.[7] As in Sandrelli and Bendazzoli (2005), function words are identified relying on automatic POS tagging: a list of all POS tags present in the four subcorpora for each language is obtained, and tags are manually checked and classified as lexical (adjective, noun, digit, verb and open-class adverb tags)[8,9] or functional (all other tags excluding those for punctuation signs, disfluencies and foreign words). A total of 19 function word tags are identified in this way for English and 25 for Italian. Compared to the method used by Laviosa (1998a), as well as Kajzer-Wietrzny (2012), relying on plain text corpora and an external list of function words, this procedure allows for closer fitting to the actual words in the corpus, and provides a principled way of extracting comparable data from corpora in different languages.

Mean sentence length is calculated by dividing the number of running words in each text in each EPTIC subcorpus by the number of sentences in that text. In the case of the written components of EPTIC, we rely on sentence boundaries assigned by the tagger; for the spoken components we perform the counts based on sentence-like units as defined by the EPIC transcribers, who segmented the text "on the basis of the speaker's intonation and syntactic information available in the sentence" (Russo et al. 2012, 59).

Core vocabulary coverage refers to the proportion of high frequency words to low(er) frequency words, where high frequency words are defined as the 200 most frequent words in large reference corpora of English and Italian. To obtain the number of occurrences of low frequency words, we sum up the number of occurrences of each high frequency word in each EPTIC text and subtract this figure from the total number of running words for that text. We express the proportion of high to low frequency words as a percentage. Laviosa's high frequency word list (also used by Kajzer-Wietrzny 2012) was obtained from the Collins Cobuild Bank of English. For this study we opted for lists extracted from the ukWaC and itWaC corpora respectively (Baroni et al. 2009). With all the provisos applying to web data, this solution has the advantage of providing more recent, comparable data about the most frequent words in the two languages in our study.

List head coverage is defined as the proportion of each subcorpus accounted for by the top hundred words of their frequency lists. Unlike core vocabulary coverage, which is determined with respect to an external point of reference, list head coverage is a (sub)corpus-internal measure, whereby list head status is defined based on cumulative frequencies at subcorpus level. This might be the reason that led Laviosa (1998a) to perform a *by-corpus*, rather than a *by-text*, analysis with respect to this measure; to ensure comparability of results, we adopt the same method. To obtain the counts of non-list head words, we sum up the number of occurrences of each list head word in its subcorpus and subtract this figure from that for the total number of running words in that subcorpus.

In the case of by-text analyses (lexical density, mean sentence length and core vocabulary coverage) we perform statistical testing using Kruskal-Wallis tests to compare all subcorpora (translation and interpreting sources and targets) in each of the two languages, followed where appropriate by post-hoc pairwise intermodal and comparable comparisons using Wilcoxon rank sum tests (with Bonferroni correction). We use non-parametric tests based on the results of preliminary checks on the normality of the distributions (Shapiro-Wilks tests), which revealed that the majority of data is not normally distributed. The Chi-square (χ^2) test is used for by-corpus analyses (list head coverage). All statistical analyses are performed using the R software.[10]

Finally, as an indicator of (dis)homogeneity within the different subcorpora, we compute variances pertaining to those measures that are calculated on a single text basis and evaluate the significance of differences across the different subcorpora using the Ansari-Bradley test.

4.2 Data analysis

Table 4 shows the median values for the four studied parameters in the eight subcorpora. Tables 5 and 6 present the complete set of results of significance tests obtained for the by-text comparisons (core vocabulary, lexical density and sentence length), while results concerning the comparisons of list head coverage, which are carried out on a by-corpus basis, are reported in Table 7.

Table 4. Median values for the eight sub-corpora and the four parameters

	English	Italian	
tt-in	56.95%	46.17%	Core vocabulary coverage
tt-tr	52.46%	44.82%	
st-in	52.77%	45.67%	
st-tr	53.25%	46.40%	
tt-in	57.03%	44.24%	List head coverage
tt-tr	54.58%	42.76%	
st-in	52.97%	45.94%	
st-tr	51.14%	45.43%	
tt-in	44.94%	46.71%	Lexical density
tt-tr	48.80%	48.90%	
st-in	48.34%	46.67%	
st-tr	48.14%	47.85%	
tt-in	20.30	18.59	Sentence length (words per sentence)
tt-tr	31.05	24.20	
st-in	22.43	22.90	
st-tr	23.61	31.35	

Our main focus being intermodal, we first of all look for significant differences between interpreted and translated data (the ‘intermodal’ comparisons in Tables 6 and 7). Where any such differences are observed (following a significant Kruskal-Wallis test result in Table 5), we check that they do not also apply to the corresponding source text comparisons (‘intermodal, control’ comparison in the other language). If they do, the results are discarded. Such significant but irrelevant results – e.g., sentence length in interpreted and translated English texts and in their Italian sources – are surrounded by parentheses in the tables and not commented on further.

Where a significant difference is observed in an intermodal comparison that does *not* also apply to its corresponding benchmark comparison, this is considered a valid result, i.e. one that is not due to source text dissimilarities. These significant and relevant results, shown in bold and underlined in Tables 6 and 7, apply to lexical density and sentence length in Italian and core vocabulary and list heads in English. Limited to these language/parameter combinations, we then look at the relevant monolingual comparable comparisons (‘comparable, interpreting’ and ‘comparable, translation’) to see if significant differences also exist between mediated and non-mediated subcorpora (we show these values in bold).

Table 5. Results of the Kruskal-Wallis H test (df=3)

English	Italian	

H= 10.1402
p= 0.01741
	H= 5.2287
	p= 0.1558 (n.s.)
		Core vocabulary

H = 7.833
p= 0.04959
	H= 13.5538
	p= 0.00358
		Lexical density

H= 21.6034
p= 7.888e-05
	H= 55.4084
	p = 5.618e-12
		Sentence length

Table 6. Results of the pairwise comparisons (Wilcoxon rank sum test with Bonferroni correction)

	English	Italian	
Intermodal (tt-in vs. tt-tr)	**p=0.0046**	▨	Core vocabulary

Intermodal, control (st-in vs. st-tr)	n.s.		
Comparable, interpreting (tt-in vs. st-in)	**p=0.0016**		
Comparable, translation (tt-tr vs. st-tr)	n.s.		
Intermodal (tt-in vs. tt-tr)	n.s.	**p=0.0031**	Lexical density
Intermodal, control (st-in vs. st-tr)	n.s.	n.s.	
Comparable, interpreting (tt-in vs. st-in)		n.s.	
Comparable, translation (tt-tr vs. st-tr)		n.s.	
Intermodal (tt-in vs. tt-tr)	(p=0.0018)	**p=1.1e-08**	Sentence length
Intermodal, control (st-in vs. st-tr)	n.s.	(p=0.0065)	
Comparable, interpreting (tt-in vs. st-in)		**p=0.018**	
Comparable, translation (tt-tr vs. st-tr)		**p=0.0053**	

Table 7. Results of the χ^2 test for the by-corpus comparisons of list head coverage (df=1)

	English	Italian	
Intermodal (tt-in vs. tt-tr)	**χ^2=8.0437,** **p=0.004566**	(χ^2=15.7206, p=7.342e-05)	List head
Intermodal, control (st-in vs. st-tr)	(χ^2=26.2893, p=2.939e-07)	n.s.	
Comparable, interpreting (tt-in vs. st-in)	**χ^2=37.4966,** **p=9.157e-10**	▨	
Comparable, translation (tt-tr vs. st-tr)	**χ^2=27.1857,** **p=1.848e-07**	▨	

We plot the significant intermodal comparisons alongside the corresponding monolingual comparable and non-mediated spoken and written comparisons by means of box plots (Figures 2-4) and bar plots (Figure 5). The intermodal comparisons are shown on the left-hand side, the control non-mediated comparisons on the right-hand side, while comparable comparisons are non-adjacent. The line connecting the median values obtained for the four subcorpora shows how they position themselves with respect to each other in terms of greater/lesser simplification.

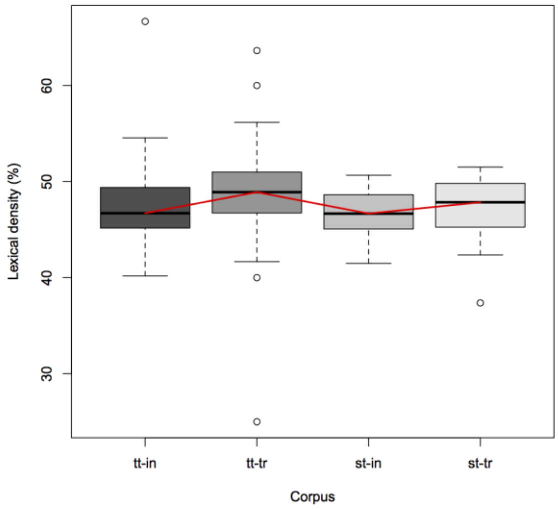

Figure 2. Lexical density in Italian

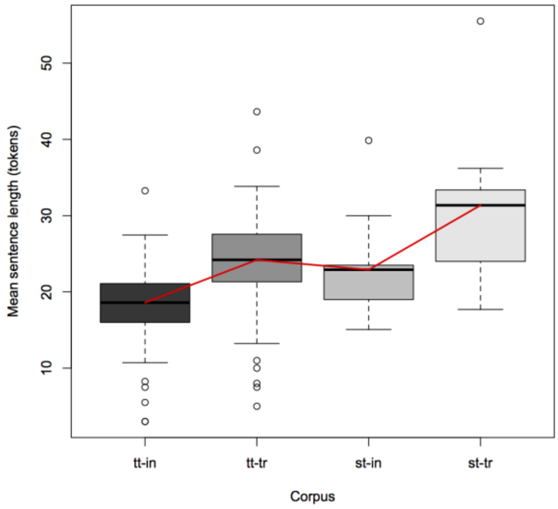

Figure 3. Mean sentence length in Italian

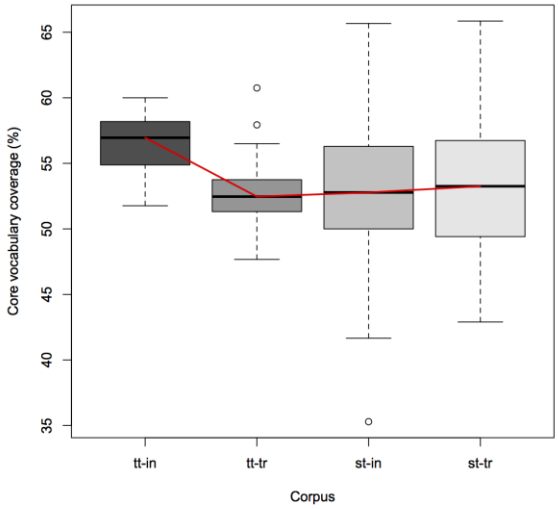

Figure 4. Core vocabulary coverage in English

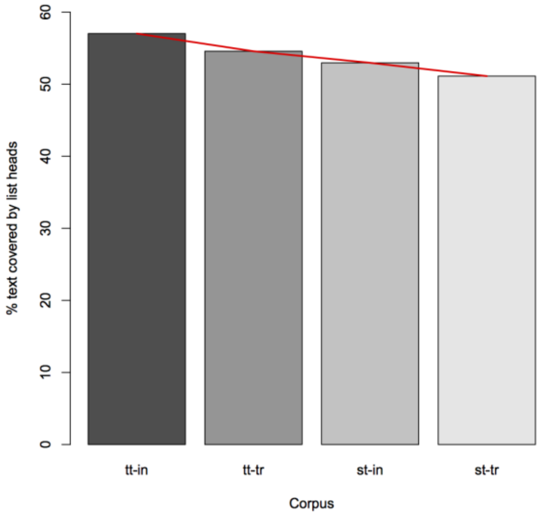

Figure 5. List head coverage in English

A final word on the results of variance testing. There was no intermodal comparison for which the Ansari-Bradley test returned significant differences, therefore monolingual comparable comparisons were not performed, coherently with the general approach adopted in this study, and we concluded that individual variation could be disregarded as a factor affecting the intermodal analysis of simplification in EPTIC.

4.3 Discussion of results

Focusing first on the intermodal (interpreted vs. translated) comparisons, the picture that emerges is one in which interpreted texts in EPTIC are consistently simpler than their translated counterparts. More specifically, interpreted Italian texts have significantly lower lexical density and mean sentence length than their translated versions, while interpreted English texts make larger use of frequent words, both when measured corpus-internally (list heads) and externally (core vocabulary). Simplification parameters thus seem to apply differently to different languages. Interpreters into English rely more on purely lexical resources (greater use of common words and text-internal repetitions), and interpreters into Italian on lexico-syntactic ones (shorter sentences and more function words, the latter possibly indicating a greater incidence of verbal rather than nominal structures, sometimes due to a closer adherence to the source input), when producing texts that are simpler than their translated counterparts.[11]

Moving on to the monolingual comparable perspective, all comparisons returning significant results follow a trend whereby the mediated corpus component is simpler than the corresponding non-mediated one. Significant differences are observed in terms of higher list head coverage in interpreted and translated English, higher core vocabulary coverage in interpreted English, and lower values for sentence length in interpreted and translated Italian, with respect to the corresponding non-mediated modes. The trend thus applies to translation and interpreting alike, though it seems slightly stronger in interpreting, since the translated and non-translated English subcorpora do not differ in terms of core vocabulary coverage, while the interpreted and non-interpreted ones do.

Our method and the structure of EPTIC also allow us to put both the intermodal and the monolingual comparable results in perspective by setting them against the background of a third kind of comparison, namely that between spoken and written (non-mediated) subcorpora. If we now focus on the results for the non-interpreted vs. non-translated subcorpus comparisons, the pattern observed at the intermodal and monolingual comparable levels can be further refined: the transcriptions of oral speeches are simpler than the written versions of the same texts in three out of four comparisons (even though not significantly for lexical density in Italian). The only exception is core vocabulary in English, whose coverage in the written subcorpus is however only 0.48% higher than in the spoken one, a negligible, non-significant difference. The non-mediated subcorpora thus display the same trend as the corresponding mediated ones: in both cases, spoken language is simpler than written language. But while the non-mediated subcorpora differ significantly in terms of one parameter each (sentence length in Italian, list head coverage in English), the mediated subcorpora differ in terms of two (the former two plus lexical density in Italian and core vocabulary in English). 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.

Since no intermodal study of simplification exists, the relationship between these results and those obtained in previous studies can only be discussed with reference to the monolingual comparable perspective. Laviosa (1998a, 1998b) found ample evidence of simplification in newspaper articles and narrative texts translated into English from a variety of languages. Translated and non-translated EU parliamentary proceedings differ significantly along a more restricted set of dimensions. Yet the general pattern of greater simplicity in translated language is confirmed for this text type as well. With respect to interpreting, our study offers a clearer picture than that emerging from Kajzer-Wietrzny's (2012) study. Her analyses of core vocabulary and lexical density run counter to the hypothesis of greater simplification in interpreted vs. non-interpreted language, which was instead confirmed by her analysis of list heads. In our study all parameters coherently point to greater simplification in interpreted (vs. non-interpreted) texts. Finally, in line with previous studies, we also found evidence of language-specific patterns: repetitiveness as measured through list head coverage constitutes the most prominent simplification parameter in interpreted English (cf. Sandrelli and Bendazzoli 2005, Kajzer-Wietrzny 2012), but the same is not true of Italian (see also Sandrelli and

Bendazzoli 2005), where the central property distinguishing between different subcorpora is sentence length. Core vocabulary coverage and lexical density appear to be less stable as simplification indicators and more difficult to compare with previous studies, possibly due to differences in the methods used to calculate them.

5. Conclusions

In this article we have presented EPTIC, a new bidirectional corpus of interpreted and translated texts in English and Italian containing texts from the EU Parliament plenary sessions. As a first attempt at uncovering its potential, we have replicated a methodology proposed by Laviosa (1998a, 1998b) and previously applied to monolingual comparable corpora of both translation and interpreting. The combined availability of mediated and non-mediated, written and spoken, comparable and parallel, English and Italian texts made it possible for us to observe features of translation and interpreting against a multifaceted background. We concluded that interpreted texts are simpler than translated ones and that mediated texts are simpler than non-mediated ones in both English and Italian, even though different parameters of simplification apply differently to the two languages. In particular, while the mediation process reduces complexity in both modes of language production, the fact that interpreted and translated texts differ in terms of more simplification parameters than the corresponding non-mediated texts hints at the fact that interpreters simplify their input more than translators do. At least as far as simplification is concerned, we thus concur with Shlesinger and Ordan’s (2012, 54) view of interpreting as "an extreme case of translation".

A note of caution is in order at this point. Though fascinating, the multifaceted perspective afforded by an intermodal corpus has a number of limits, particularly when the corpus is small and does not provide access to multimedia source contents. Taking into consideration situational factors (e.g. impromptu vs. prepared speeches) and modality-specific parameters (e.g. prosodic structures), for instance, would certainly allow us to paint a more accurate picture than we have been able to do here. [12]

There are several other ways in which the analytical work presented here could be extended. First, the growing body of hypotheses (see e.g. the survey in Zanettin 2013) about features of translation (and interpreting) could be scrutinized from an intermodal perspective. Second, in this study we made no attempt to factor in the bilingual parallel (source vs. target) perspective, yet a three-way aligned corpus like EPTIC allows the straightforward comparison of the product of independent decisions taken by interpreters and translators when faced with the same problems, under genuine working conditions.

However, we believe that the most pressing issues concern the availability and further expansion of the corpus. As pointed out by Shlesinger and Ordan (2012, 44), the literature in the field of interpreting studies abounds with examples of small-scale corpora used almost exclusively by the researchers who constructed them. Such a situation clearly impedes replication studies, as the corpora are often not available to other researchers. EPTIC could easily be made available through a web interface or distributed to those who request it: the nature of EU Parliament proceedings is such that there should be no restriction to their distribution, particularly for non-profit research purposes.

In terms of development, the corpus could continue to grow thanks to the efforts of a small group of researchers supported by Master’s and Doctoral students, but we believe that there are very special conditions that make EPTIC a perfect corpus for a shared task. The absence of copyright issues, the availability of data in a very large number of languages and directions, and the very object of study, that requires scholars in two sister (yet not fully integrated) disciplines to work together and share their methodological assumptions, mean that a fully-fledged corpus cannot and should not be constructed single-handedly.

Our aim in the near future is therefore to set up the infrastructure allowing interested researchers to contribute to EPTIC and to provide space for discussion (about metadata, transcription conventions, corpus encoding, etc.), as well as for sharing methods and results. In so doing we hope that a consensus emerges on "methods and procedures […] for corpus description and presentation, to allow sharing and comparability, and perhaps more elusively, on a theoretical framework and procedures for analysis" (Setton 2011, 36), which would be highly beneficial to both corpus-based interpreting and translation studies, and could strengthen the common ground between the two.

Notes

1. We would like to thank the creators of EPIC for assisting us in the initial stages of its transformation into EPTIC (we are especially indebted to Claudio Bendazzoli and Maria Chiara Russo). We are also grateful to students of interpreting and translation at the University of Bologna, who are enthusiastically contributing to the further enlargement of the corpus, in particular Rita Micchi, Niccolò Morselli and Manuela Santandrea. The research reported here was first presented at the Workshop on New Ways of Analyzing Translational Behaviour in Corpus-Based Translation Studies (SLE 2013, Split). It benefited greatly from insightful comments from the other workshop participants, as well as from the careful reading of two anonymous reviewers. Finally, this work is dedicated to the memory of Miriam Shlesinger, to whom it owes so much.

2. The European Parliament has since made available for download high-quality recordings of speeches as originally delivered and of their interpretations (starting from September 2008), accessible at http://www.europarl.europa.eu/plenary/en/debates-video.html and http://www.europarl.europa.eu/ep-live/en/plenary/.

3. Verbatim reports and translations can be downloaded from: http://www.europarl.europa.eu/RegistreWeb/search/typedoc.htm?language=EN, under "Documents relating to parliamentary activity > Plenary documents".

4. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

5. http://cwb.sourceforge.net/

6. An anonymous reviewer suggests that a more rewarding procedure to deal with such cases would be to carry out a regression analysis on the source texts, so that single texts could be discarded instead of the whole set of results. We leave it to future work to investigate this possibility, as the use of a regression would impede comparability of our results with those of previous studies on lexical simplification.

7. The extraction of running words is performed via the *cwb-scan-corpus* tool, using the following regular expression: `[a-zA-Z0-9àèéìòùÈ\'\-]+[^\-]` (where `[^\-]` excludes truncated words).

8. Since the Italian tagset does not assign a separate tag to modals, whereas both the English and Italian tagsets do tag auxiliaries separately from other verbs, to ensure comparability modals are considered lexical and auxiliaries are considered functional in both languages.

9. Open-class adverbs in Italian are assigned a specific tag by the TreeTagger (*ADV:mente*), and this is excluded from the list of function word tags; for English, since open-class adverbs are assigned the same tag as closed-class ones, we exclude from the count of function words adverbs ending in *-ly* (except *only*).

10. http://www.r-project.org

11. Specialized translators from English into Italian often render verbal forms in the source language with nominal forms in the target language as a means of raising the register (Scarpa 2001, 135). Though unsystematic, spot checks of parallel concordances suggest that interpreters resort to this procedure to a more limited extent than translators do. This may be due to the simple fact that they reformulate less, or to a preference for a more informal register.

12. It is worth pointing out that, all through this article, no value judgment is implied by the use of the terms “simple” and “simplification”. We agree with an anonymous reviewer that, “simpler may in fact be better by being more communicative”.

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications." In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis and Elena Tognini-Bonelli, 233–250. Amsterdam: John Benjamins.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora." *Language Resources and Evaluation* 43 (3): 209–226.

Bendazzoli, Claudio. 2010. *Corpora e Interpretazione Simultanea*. Bologna: Asterisco.

Gile, Daniel. 2004. "Translation Research versus Interpreting Research: Kinship, Differences and Prospects for Partnership." In Schäffner 2004, 10–34.

Halliday, Michael A. K. and Christian Mathiessen. 2004. *An Introduction to Functional Grammar*. London: Arnold.

Kajzer-Wietrzny, Marta. 2012. *Interpreting Universals and Interpreting Style*. Doctoral dissertation. Adam Mickiewicz University, Poznan.

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *Machine Translation Summit* X, 79–86.

Laviosa, Sara. 1998a. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta* 43 (4): 557–570.

Laviosa, Sara. 1998b. "The English Comparable Corpus. A Resource and a Methodology." In *Unity in Diversity? Current Trends in Translation Studies*. Ed. By Lynne Bowker, Michael Cronin, Dorothy Kenny and Jennifer Pearson, 101-112. Manchester: St. Jerome.

McEnery, Anthony, Richard Xiao and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book.* London: Routledge.

Pöchhacker, Franz. 2004. "I in TS: On Partnership in Translation Studies." In Schäffner 2004, 104–115.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli. 2006. “Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC.” *Forum* 4 (1): 221-254.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli and Nicoletta Spinolo. 2012. “The European Parliament Interpreting Corpus (EPIC): Implementation and Developments.” In *Breaking Ground in Corpus-based Interpreting Studies*, ed. by Francesco Straniero Sergio and Caterina Falbo, 35-90. Bern: Peter Lang.

Sandrelli, Annalisa, and 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 1*. University of Birmingham, Birmingham.

Sandrelli, Analisa, Claudio Bendazzoli, and Mariachiara Russo. 2010. "European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in Simultaneous Interpreting." *International Journal of Translation* 22 (1-2): 165–203.

Scarpa, Federica. 2001. *La Traduzione Specializzata*. Milano: Hoepli.

Schäffner, Christina (ed.). 2004. *Translation Research and Interpreting Research. Traditions, Gaps and Synergies*. Clevedon: Multilingual Matters.

Shlesinger, Miriam. 1989. *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. Unpublished M.A. thesis. Tel Aviv: Tel Aviv University.

Shlesinger, Miriam. 1998. "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." *Meta* 43 (4): 486–493.

Shlesinger, Miriam. 2009. "Towards a Definition of Interpretese: An Intermodal, Corpus-Based Study." In *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, ed. by Gyde Hansen, Andrew Chesterman and Heidrun Gerzymisch-Arbogast, 237–253. Amsterdam: John Benjamins.

Shlesinger, Miriam, and Noam Ordan 2012. "More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting." *Target* 24 (1): 43–60.

Setton, Robin. 2011. "Corpus-Based Interpretation Studies: Reflections and Prospects." In *Corpus-based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach and Jeremy Munday, 33-75. London: Continuum.

Stubbs, Michael. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Vuorikoski, Anna-Riitta. 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*. Doctoral dissertation, University of Tampere.

Zanettin, Federico. 2013. "Corpus Methods for Descriptive Translation Studies." *Procedia. Social and Behavioural Sciences* 95: 20-32.

Appendix

Metadata available for the different EPTIC subcorpora

Metadata on	**Metadata element**	**Possible values and explanation**	**Available in sub-corpus**
Text	**id**	*[a univocal progressive number]*	all
	date	*[date of delivery of speech]*	all
	length_words	*[length of text in word tokens]*	all
	length	short *[<300 words];* medium *[301-1000 words];* long *[>1000 words]*	all
	duration_seconds	*[duration of text in seconds]*	st-in, tt-in
	duration	short *[<2 minutes];* medium *[2-6 minutes];* long *[>6 minutes]*	st-in, tt-in
	speed	slow *[<130 w/m];* medium *[131-160 w/m];* high *[>160 w/m]*	st-in, tt-in
	delivery	read; impromptu; mixed; interpreted *[mode of delivery of the speech; interpreted texts always have value "interpreted"]*	st-in, tt-in
	topic	*[macro-topic of the text, e.g. Health]*	all
	specific_topic	*[specific topic, e.g. Asian bird flu]*	all
	type	st-in; tt-in; st-tr; tt-tr *[type of sub-corpus]*	all
	comments	*[(optional) comments by transcriber]*	all
Speaker	**name**	*[name of speaker, e.g. Francesco Fiori]*	all

	gender	M *[male]*; F *[female]*	all
	country	*[country of origin of speaker]*	all
	native	yes; no *[status of speaker with reference to the language in which the speech is delivered]*	all
	political_function	*[political function of speaker, e.g. MEP]*	all
	political_group	*[political group of speaker, e.g. PPE-DE]*	all
Source text	**length_words**	*[value duplicated from the corresponding ST]*	tt-in, tt-tr
	length		
	duration_seconds		
	duration		
	speed		
	delivery		
Interpreter	**gender**	M *[male]*; F *[female]*	all
	native	yes; no *[status of interpreter with reference to the language into which the speech is interpreted]*	tt-in