Wikström, Peter. 2017. *I tweet like I talk : Aspects of speech and writing on Twitter.* Karlstads: Karlstads universitet. (Doctoral dissertation). http : / / urn . kb . se / resolve?urn=urn:nbn:se:kau:diva-64752 (15 February, 2021).

Chapter 7

Exploring linguistic variation in mediated discourse: Translation vs. interpreting

Heike Przybyl[a], Alina Karakanta[b,c], Katrin Menzel[a] & Elke Teich[a]

[a]Saarland University [b]Fondazione Bruno Kessler [c]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 the 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.), *Mediated discourse at the European Parliament: Empirical investigations*, 191–218. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977050

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 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 level (list head coverage and core vocabulary) as well as the 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 a target or a 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 the 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 work in translation studies, we have described the basic workings of the approach in Karakanta et al. (2021) 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 the 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.

[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	372,547	8,693,135	ORG SP EN	3,623	68,548

We build probabilistic unigram language models of the source and target languages for interpreting and translation and calculate the relative entropy between 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). Based on 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. 2021). 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 more detailed corpus analysis. To this aim, we compute Standardised Type-Token Ratio (STTR), lexical density (the number 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

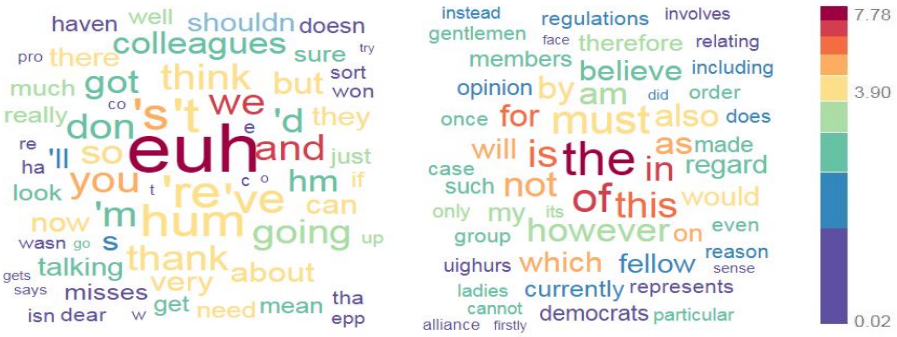
Figure 1: Variation in translation mode with German as target and English as source language. Relative frequency (RelF) is indicated by colour (high RelF red, low RelF blue), distinctivity is visualized by size.

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 (e.g. *der, die, dieser, diesem, ihre, seine, unser, meiner*) and by more content words shown to be distinctive for

[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.

translations (e.g. *Bericht, Parlament, Ansicht, Präsident, vergangenen*), however at a 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 the 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 shows 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. 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 modes of German (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).

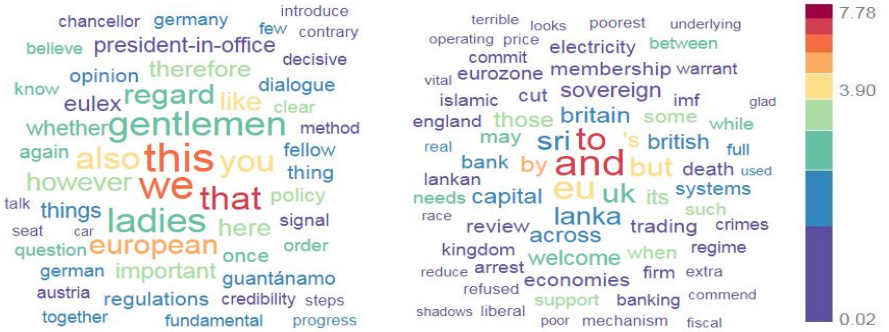
Overall, interpreting seems to be more spoken than originals. One explanation could be that, although all of the originals are true transcripts of original speeches held in the European Parliament, some of the interventions had been prepared by the members of Parliament, and therefore typical spoken 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 because 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.

(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.

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 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.

(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.

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 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. For our results to be comparable regardless of corpus size, we compute the 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 $= 3$, $p = 4.429 \times 10^{-5}$; DE: $t = 25.299$, df $= 3$, $p = 0.0001354$). 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 shows 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 the word level, a check of mean token length reveals that SI tends towards using shorter words (see Table 3, EN: $t = 99.46, \mathrm{df} = 3, p = 2.241 \times 10^{-6}$; DE: $t = 62.285, \mathrm{df} = 3, p = 9.119 \times 10^{-6}$). The 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, $p = 1.91 \times 10^{-6}$; DE: t = 74.007, df = 3, $p = 5.437 \times 10^{-6}$). 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.

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

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 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: $\chi^2 = 26662$, df = 21, $p < 2.2 \times 10^{-16}$; EN: $\chi^2 = 14266$, df = 21, $p < 2.2 \times 10^{-16}$). Figure 9 plots the part-of-speech distributions in terms of relative frequencies (the 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 are for the spoken ones. Determiners are less frequently used in SI, and pronouns seem to be compensating for 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. 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 confirms 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

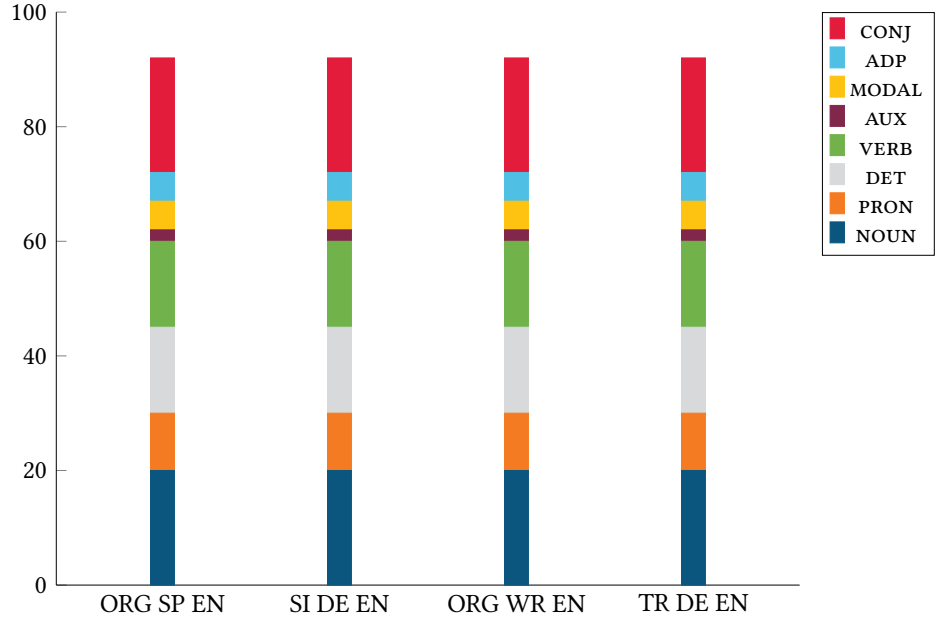
[3]DE: $\chi^2 = 3269.4$, df = 9, $p < 2.2 \times 10^{-16}$; EN $\chi^2 = 2022.6$, df = 9, $p < 2.2 \times 10^{-16}$.

[4]ORG SP DE vs. ORG WR DE: $\chi^2 = 1.1987$, df = 3, $p = 0.7533$.

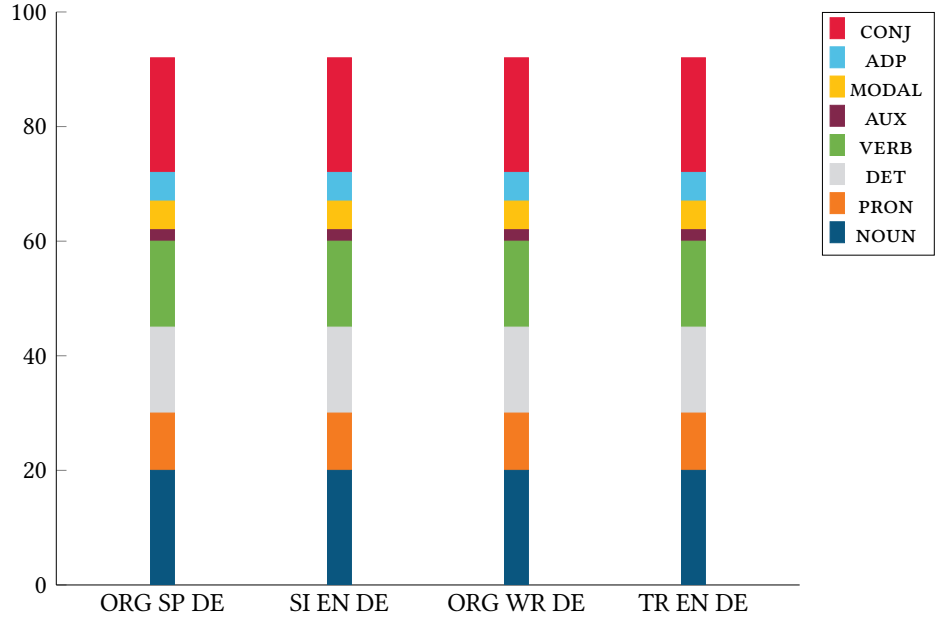
[5]ORG SP DE vs. SI EN DE: $\chi^2 = 248.73$, df = 3, $p < 2.2 \times 10^{-16}$; ORG WR DE vs. TR EN DE: $\chi^2 = 2625.8$, df = 3, $p < 2.2 \times 10^{-16}$.

[6]SI EN DE vs. TR EN DE: $\chi^2 = 912.86$, df = 3, $p < 2.2 \times 10^{-16}$.

[7]ORG SP EN vs. ORG WR EN: $\chi^2 = 314.39$, df = 3, $p < 2.2 \times 10^{-16}$; ORG SP EN vs. SI DE EN: $\chi^2 = 248.73$, df = 3, $p < 2.2 \times 10^{-16}$; ORG WR EN vs. TR DE EN: $\chi^2 = 1381.2$, df = 3, $p < 2.2 \times 10^{-16}$; SI DE EN vs. TR DE EN: $\chi^2 = 303.46$, df = 3, $p < 2.2 \times 10^{-16}$.

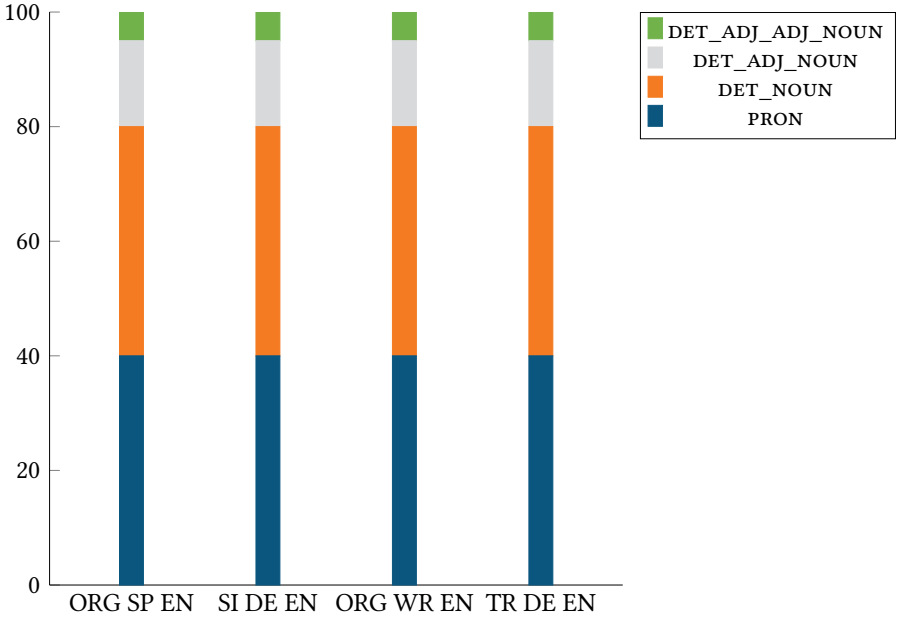


(a) English

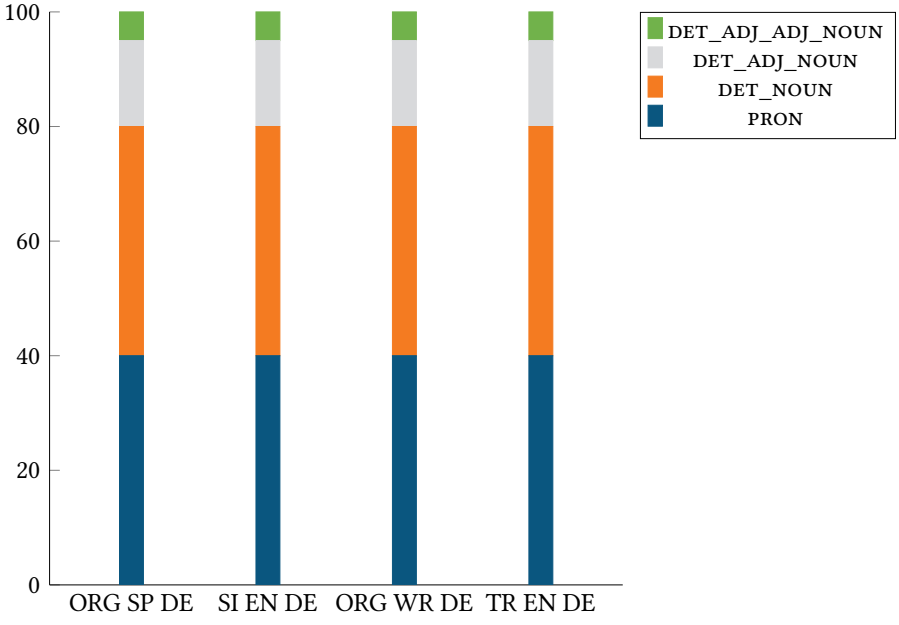


(b) German

Figure 9: POS distribution for selected POS for English and German.



(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 because 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 mechanism for reducing processing effort.

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

A more detailed observation of the most frequently used lexical types in the patterns reveals that SI seems to use more fixed, standardised phrases (e.g. *eine wichtige Rolle* (*an important role*)) that do 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 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. Therefore, we compare the use of ver-

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).

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

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.

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

[8] ORG SP EN vs. ORG WR EN: $\chi^2 = 48.11$, df = 2, $p = 3.572 \times 10^{-11}$; ORG SP EN vs. SI DE EN: $\chi^2 = 34.381$, df = 2, $p = 3.422 \times 10^{-8}$; ORG WR EN vs. TR DE EN: $\chi^2 = 5318.5$, df = 2, $p < 2.2 \times 10^{-16}$; SI DE EN vs. TR DE EN: $\chi^2 = 37.131$, df = 2, $p = 8.65 \times 10^{-9}$.

[9] ORG SP DE vs. ORG WR DE: $\chi^2 = 0.29177$, df = 2, $p = 0.8643$.

[10] ORG SP DE vs. SI EN DE: $\chi^2 = 9.9219$, df = 2, $p = 0.007006$; ORG WR DE vs. TR EN DE: $\chi^2 = 1030.4$, df = 2, $p < 2.2 \times 10^{-16}$; SI EN DE vs. TR EN DE: $\chi^2 = 38.291$, df = 2, $p = 4.843 \times 10^{-9}$.

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.

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

See some examples in (1) 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*.

(1) a. ORG SP DE: "... es ist gut dass wir ihn haben / *aber* er soll eben Maßnahmen regeln die..."

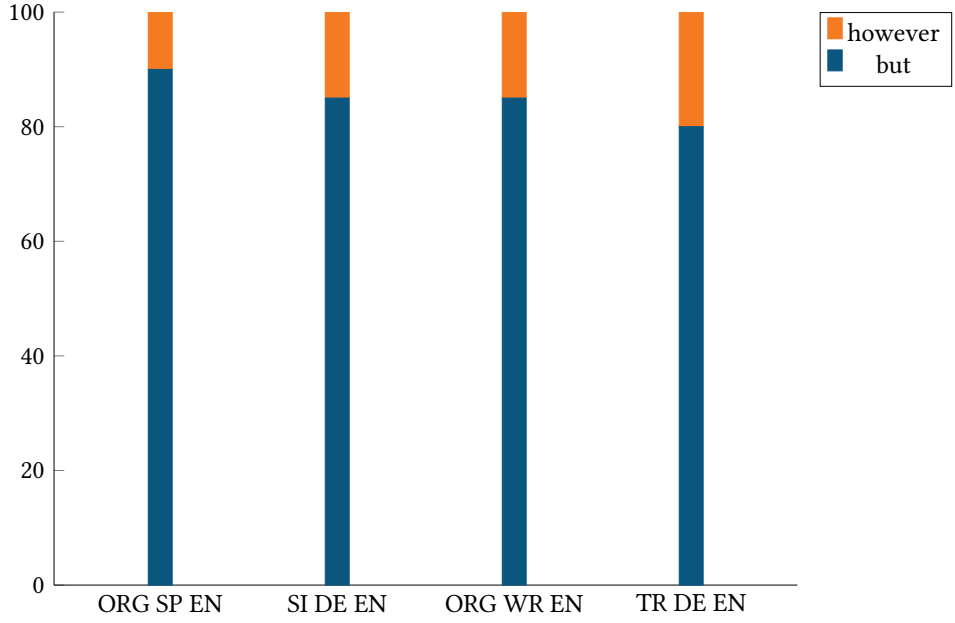
b. SI DE EN: "...it is very good that we have it / *but* its rule should apply to..."

c. 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 ..."

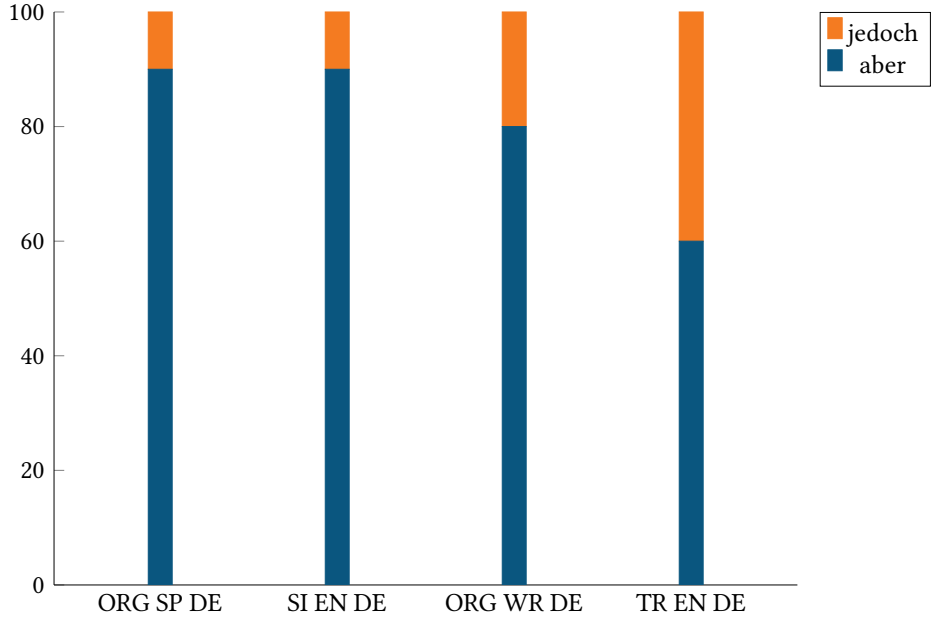
6 Conclusion and outlook

We have presented a data-driven, exploratory method to analyse 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.

[11] ORG SP EN vs. SI DE EN: $p = 0.1627$; ORG SP DE vs. SI EN DE: $p = 0.7176$.
[12] ORG WR EN vs. TR DE EN: $p < 2.2 \times 10^{-16}$, ORG WR DE vs. TR EN DE: $p < 2.2 \times 10^{-16}$.



(a) English



(b) German

Figure 11: Distribution of *but/however* and *aber/jedoch*.

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) through 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, Viktor. 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*, 149–161. https : / / www . euroconferences . info / proceedings / 2005 _ Proceedings/2005_proceedings.html.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 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*, 21–42. Singapore: Springer. DOI: 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: 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: De Gruyter Mouton. DOI: 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 As-*