Simultaneous Interpreting as a Noisy Channel: How Much Information Gets Through

Maria Kunilovskaya
Saarland University, Saarbrücken
maria.kunilovskaya@uni-saarland.de

Heike Przybyl
Saarland University, Saarbrücken
heike.przybyl@uni-saarland.de

Elke Teich
Saarland University, Saarbrücken
e.teich@mx.uni-saarland.de

Ekaterina Lapshinova-Koltunski
University of Hildesheim
lapshinovakoltun@uni-hildesheim.de

Abstract

We explore the relationship between information density/surprisal of source and target texts in translation and interpreting in the language pair English-German, looking at the specific properties of translation ("translationese"). Our data comes from two bidirectional English-German subcorpora representing written and spoken mediation modes collected from European Parliament proceedings. Within each language, we (a) compare original speeches to their translated or interpreted counterparts, and (b) explore the association between segment-aligned sources and targets in each translation direction. As additional variables, we consider source delivery mode (read-out, impromptu) and source speech rate in interpreting. We use language modelling to measure the information rendered by words in a segment and to characterise the cross-lingual transfer of information under various conditions. Our approach is based on statistical analyses of surprisal values, extracted from n-gram models of our dataset. The analysis reveals that while there is a considerable positive correlation between the average surprisal of source and target segments in both modes, information output in interpreting is lower than in translation, given the same amount of input. Significantly lower information density in spoken mediated production compared to non-mediated speech in the same language can indicate a possible simplification effect in interpreting.

1 Introduction

In this study, we describe and explain linguistic choice in translation and interpreting from the point of view of rational communication, according to which language users strive to encode their messages effectively and efficiently, i.e. they attempt to ensure that their messages are transmitted successfully while at the same time, their cognitive effort stays at a reasonable level (see e.g. Crocker et al., 2015). Our approach stipulates that the behaviour of translators, while guided by effectiveness and efficiency, is severely constrained by the specific conditions of mediated communication, especially in interpreting (see studies on the cognitive effort in interpreting, e.g. Christoffels et al., 2006; Chmiel, 2021). Simultaneous interpreters have to balance allocating cognitive resources to overlapping comprehension and production processes in a way that allows them to complete the task and communication is not put at risk.

From empirical translatology we know that the coping mechanisms involved in translation/interpreting have an impact on the linguistic properties of the output, widely known as *translationese* (e.g. Baker, 1996; Teich, 2003; Shlesinger and Ordan, 2012, cf. Section 2). While there is a rich literature on trends in translational behaviour (e.g. simplification, explicitation, normalisation), a unifying explanation for the diverse linguistic phenomena is still lacking. This study is an attempt to fill this gap by adopting an information-theoretic approach. Our analysis is based on measuring *information density* (ID) aka *surprisal* of translation/interpreting outputs and contrasting them with non-mediated (i.e. original) speeches and between each other, as well as looking at the association between surprisal values of aligned source and target segments.

We interpret surprisal as the amount of information conveyed by a given linguistic event from the point of view of a given language model. In mediated communication, interpreters' and translators' output is expected to reflect the amount of information contained in the source. However, it may be expected that interpreters will not manage to encode the target to the same level of average surprisal (short: AvS) as observed in the source.

Apart from *mediation mode* (translation, inter-

Proceedings of Recent Advances in Natural Language Processing, pages 608–618
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_066

preting) and *translation direction*, further factors may have an impact on encoding. In simultaneous interpreting, where comprehension of the source text (ST) and production of the target text (TT) claim cognitive resources at the same time, the amount of information transmitted from ST to TT may vary according to *source delivery mode* (impromptu vs. read-out) and *source speech rate* (words per minute).

With regard to the various factors at play in cross-lingual mediation discussed above, we formulate the following hypotheses.

- **(H1)** While we expect a general, positive correlation between sources and targets in terms of AvS **(H1a)**, it can be hypothesised that interpreting will be lower in information output per same information input than translation (due to the specific on-line conditions of interpreting) **(H1b)**;

- **(H2)** AvS is expected to be lower in mediated texts relative to comparable non-mediated texts in the same language, irrespective of source/target language and mediation mode (cf. *simplification* trend in translation) **(H2a)**, the AvS and the range of surprisal values in interpreting are likely to be smaller than in translation due to *simplification* and reinforced features of spoken production **(H2b)**.

- **(H3)** AvS of interpreted texts should be less strongly associated with the source for read-out vs. impromptu delivery of the source (**H3a**) and also less associated for speeches with higher speed of the source delivery than for lower-speed delivery (due to increased processing cost) (**H3b**).

To address these hypotheses, we analyse surprisal in a bidirectional English-German corpus of European Parliament proceedings containing both mediation modes. The remainder of the paper is organised as follows. Section 2 provides an overview of related work and theoretical background. Section 3 describes our methodology and experimental setup. In Sections 4 and 5, we present the results and their interpretation. Section 6 gives a summary and conclusion.

2 Background and Related Work

2.1 Translation and Interpreting Studies

As mentioned above, mediated texts are known to carry *translationese* features, i.e. specific linguistic properties induced by the translation process that set translations apart from non-mediated originals in the target language. These features can be explained by simplification (see e.g. Laviosa, 1998; Toury, 1995) – the tendency to use simpler constructions (e.g. simpler syntactic structure or more general words), explicitation and implicitation (Blum-Kulka, 1986), often interpreted as an increased or decreased use of linking devices such as connectives, as well as normalisation and shining through (Baker, 1995; Teich, 2003), i.e. orientation of translations towards either target or source language, respectively. Due to their statistical character, these properties can be automatically uncovered (Baroni and Bernardini, 2005; Volansky et al., 2015; Kunilovskaya and Lapshinova-Koltunski, 2020) and have recently received increased attention in multilingual language processing (Dutta Chowdhury et al., 2020; Artetxe et al., 2020; Graham et al., 2020). However, simultaneous interpreting as a spoken mediation type tends to show different properties than translation (Kajzer-Wietrzny, 2012), *interpretese* being more pronounced overall and reinforcing spoken features (Shlesinger and Ordan, 2012).

Although there is a substantial bulk of work on translationese, the explanation for the mechanisms behind them is still missing. There exist studies attempting to explain translationese from the point of view of optimal communication using an information-theoretic framework. For instance, Bizzoni and Lapshinova-Koltunski (2021) and Rubino et al. (2016) use probabilistic measures (perplexity, entropy) to analyse morpho-syntactic differences between professional and student translations contrasting them to original non-mediated texts and relating them to shining through and normalisation. Martínez and Teich (2017) and Teich et al. (2020) focus on the lexical aspects of translationese and translation probability. However, while existing studies focus on the analysis of comparable corpora, i.e. mediated texts compared to non-mediated ones in the same language, we additionally investigate aligned source and target language segments, i.e. parallel texts. The only study on parallel data known to us is (Lapshinova-Koltunski et al., 2022), comparing translation and

interpreting with originals and the corresponding non-mediated texts in terms of explicitation and implicitation linking these phenomena to cognitive load measured with surprisal. However, while they look into surprisal of a restricted number of specific discourse connectives, we calculate surprisal at the level of aligned segments (typically sentences).

2.2 Information Theory as a Theoretical Premise

We apply *surprisal*, a measure based on Information Theory (Shannon, 1948) that quantifies the information content of a message in bits, to the contrastive analysis of spoken and written mediation (i) against their sources, (ii) against comparable originals in the target language, and (iii) between themselves. Surprisal is proportional to the cognitive effort required to process language units, high surprisal being indicated e.g. by a longer fixation time during reading and a larger N400 effect, a specific kind of brain response to visual or auditory stimuli observable in EEG (Lowder et al., 2018; Aurnhammer et al., 2021). Surprisal and other information-theoretic measures, such as entropy and perplexity mentioned above, are typically estimated with computational language models based on authentic language use (corpora) (Hale, 2001).

In this study, we use the (average) surprisal of translation/interpreting segments as a measure of the amount of information that gets transmitted between languages in various modes and conditions of mediated communication (as explained in Section 1).

3 Methodology

3.1 Data

This study relies on the document- and segment-aligned German-English (DE-EN) and English-German (EN-DE) subsets of EPIC-UdS (Przybyl et al., 2022) and Europarl-UdS (Karakanta et al., 2018). EPIC-UdS consists of speeches by members of the European Parliament (MEPs) and their simultaneous interpretation, both transcribed to reflect the spoken delivery features, whereas Europarl-UdS includes officially published speeches and their written translations. The materials in both corpora stem from the same communicative events — speeches made in the European Parliament — except that (i) they present the speeches either as transcripts of the spoken events or as documents adapted for reading (aka 'verbatim reports'); (ii) the target language side is either a transcript of simultaneous interpreting or a written translation. Both corpora only contain document pairs where the original speech is delivered by a person speaking in their mother tongue. The spoken corpora are enriched with the metadata on the delivery mode of source speeches (read-out, impromptu or mixed) as well as on speech rates ($slow$ ≤130w/m; medium = 131-160w/m; $high$ ≥161w/m).

		docs	segs	tokens	
				source	target
sp	DE-EN	165	3,247	56,142	49,265
	EN-DE	137	3,435	64,645	46,462
wr	DE-EN	170	2,796	67,726	77,427
	EN-DE	170	2,790	67,965	66,462

Table 1: Basic parameters of English-German parallel corpus by mode (sp and wr) and translation direction.

The general information about the datasets used in this study is given in Table 1. The counts are based on the annotated corpus, after filtering and pre-processing.

Importantly, the data was balanced across modes and translation directions to avoid biasing the models toward the properties of any over-represented test category, which is particularly important when working with smaller datasets. To that end, the amount of data available from Europarl-UdS was limited to a random set of 170 document pairs that were within one standard deviation (SD) of the average EPIC-UdS ST in terms of the number of segments per document. Care was taken to exclude Europarl-UdS speeches that appeared among EPIC-UdS transcripts. They accounted for about 90% in the German-English translation direction and could influence the model output.

Preprocessing steps included modifications that made the spoken and written documents more formally comparable. In particular, end-of-sentence (EoS) punctuation marks were added to transcribed sentences (EPIC-UdS) before linguistic annotation. With the view of reducing the n-gram model vocabulary and improving the modelling outcomes, all subcorpora were lemmatised using the default Stanza packages for German and English (Qi et al., 2020). The models' vocabularies went down by 22.2% and 20.4% for German and English, respectively (based on unigram types). For language modelling purposes, in written production (Europarl-UdS) EoS punctuation other than a full stop was

replaced with a full stop and mid-sentence punctuation was removed. In transcripts of spoken speech (EPIC-UdS), all indications of spoken phenomena (filled pauses, repetitions, repairs, etc) were removed.

3.2 Experimental Setup

An important modelling decision was to use all available balanced original and mediated data for each language, regardless of the mode, to obtain the frequency counts. We stipulate that this approach approximates the exposure to the original and mediated language experienced by European Parliament speakers and interpreters/translators and makes it possible to fairly estimate the information density of segments and individual tokens in context. Other training options — using all available written data, using only original speeches or limiting the training set to only written or spoken data to model respective subsets — reduce the comparability of modelling results across the text categories.

Our analysis relies on surprisal, an information-theoretic measure of (un)predictability of a word in context, calculated as the inverse probability of a word given its preceding context of three words measured in bits of information, see Equation (1).

$$S(w_i) = -log2(P(w_i|w_{i-3}, w_{i-2}, w_{i-1})) \quad (1)$$

The probability for each individual occurrence in a document was calculated based on the counts in the entire corpus, excluding the current document. The n-grams lists were generated with respect to sentence boundaries; *hapax legomena* tokens were replaced with a placeholder (UKN). The language models fell back to lower-order n-grams to estimate the probabilities in cases of zero evidence for higher-order n-grams.

To investigate the hypotheses put forward in Section 1, we used segment level surprisal from our 4-gram models and relied on linear regression and correlation analyses of AvS for aligned sources and target segments, as well as ran statistical significance tests to compare original and mediated sets of documents in each language, German or English.

4 Results and Analysis

4.1 Correlation Sources – Targets (H1)

First, we explore H1 to see if there is a positive association between sources and targets in terms of surprisal and if this correlation is stronger for translation compared to interpreting, given the selected modelling approach.

To quantify the relation between source and target surprisal values for each mode of mediation and each translation direction, we used the Spearman rank correlation coefficient. This measure was preferred over the Pearson correlation coefficient because we did not have enough evidence to assume a normal distribution of the surprisal values in paired sources and targets, and the variances of the respective samples were unequal (based on Shapiro-Wilk and Bartlett's tests) for some parallel corpora. Although the surprisal values for source and target segments were obtained from language-specific models, their correlation is still indicative of the strength and direction of a relation between sources and targets in terms of informativity. To ensure the comparability of results and to retain true alignment in each EPIC-UdS parallel corpus, we ignored segment pairs with zero surprisal on either side, i.e. segments that were either skipped or added in interpreting and were marked as NONE during alignment. They accounted for over 10% of all segment pairs in each translation direction.

direction	subcorpus	mode	r
DE-EN	Europarl-UdS	written	0.47
	EPIC-UdS	spoken	0.48
EN-DE	Europarl-UdS	written	0.51
	EPIC-UdS	spoken	0.44

Table 2: Spearman correlation coefficient between average surprisal for aligned source and target segments by mediation mode (for two translation directions). All results are statistically significant.

The results displayed in Table 2 show that there is a positive correlation between source and target irrespective of translation direction, which confirms our first hypothesis (**H1a**). Interestingly, there is no consistency across translation directions in the correlation levels between sources and targets in written and spoken data. The English-German data, in line with our expectations, demonstrated a higher correlation in written translation than in interpreting (0.51 for written vs 0.44 for spoken). However, in the German-English translation direction, the correlation is slightly higher in spoken than written mediation mode (0.47 for written vs 0.48 for spoken).

To visually explore the effect of mediation mode on the relation between AvS of sources and targets, we produced linear regression plots for aligned segments in each translation direction (see Figure 1). A linear regression model attempts to predict the

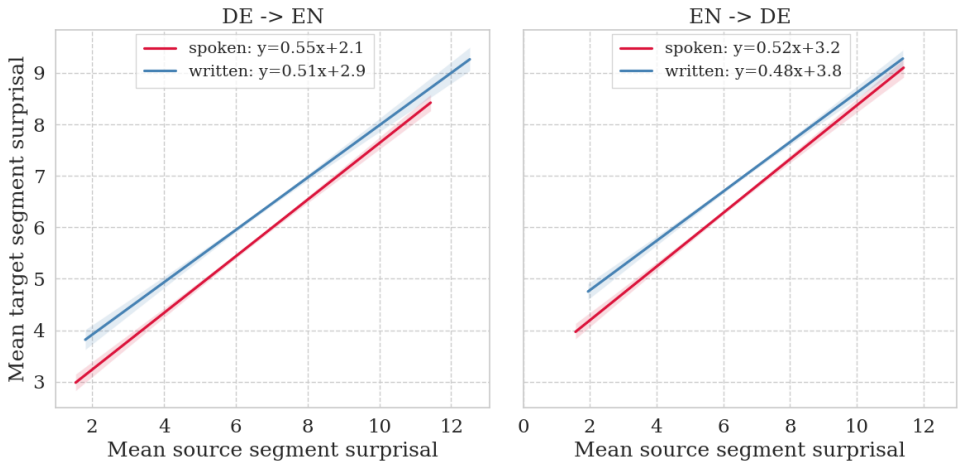

Figure 1: Linear regression based on AvS of aligned source and target segments by translation direction (DE-EN, EN-DE) and mediation type (spoken/interpreting, written/translation).

response variable (surprisal of targets, shown on y-axis) from values of the independent explanatory variable (surprisal of sources on x-axis), using a linear function. A linear relationship between the variables can be represented by Equation (2).

$$y = a * x + b \tag{2}$$

where a is the slope and b is the y-intercept.

The slope of each line indicates the amount of change in the response variable per unit of change in the explanatory variable. It can be seen that for both modes the slope is approximately the same.

The difference in y-intercept for the regression lines with almost the same slope (parallel lines) can be interpreted as the same value for the independent variable leading to different values in the response variable. Figure 1 shows that for the same level of informativity in the source (mean source segment surprisal) interpreters produce lower surprisal output than translators. This is true for both translation directions: red regression lines, representing the source-target association in interpreting, are located below the blue regression lines, representing written translation. This result confirms hypothesis **H1b**, stating that the information output in interpreting is lower than in translation for the same input.

4.2 Simplification in Mediated Texts (H2)

Next, we address the second hypothesis and analyse the expected simplification in mediated speech. For this, we compare the AvS of the mediated texts to that of comparable non-mediated texts in the same language, using statistical tests and looking at the parameters of respective distributions (the minimum and the maximum, as well as the interquartile range (IQR)). The comparison is extended to texts representing spoken and written modes in each language.

For this, we produced boxplots summarising the distribution of AvS across spoken and written modes in non-mediated (original) and mediated language production in English and German, see Figure 2. The boxes represent the spread of the middle 50% of observations. It can be seen that darker boxes representing mediated language are located lower than lighter boxes representing comparable non-mediated language, except for written German, where the surprisal values tend to be higher in translations than in non-translated documents. Given the long whiskers and a considerable number of outliers in the plots, the visual estimation of the differences between the categories might be misleading. The results from the Mann-Whitney-Wilcoxon test confirmed that the differences between the boxplotted categories are statistically significant at the confidence level of 5%, with p-values ranging from 1.41e-15 (for written non-mediated vs. written mediated in German) to 1.16e-83 (for spoken mediated vs. written mediated in German).

The Mann-Whitney-Wilcoxon significance test focuses on the rank ordering of the observations rather than the specific values themselves. The absolute values and comparisons between categories reveal some commonalities between the properties of the eight distributions shown in Figure 2. All distributions have similar parameters: the selected modelling approach results in a leptokurtic distri-

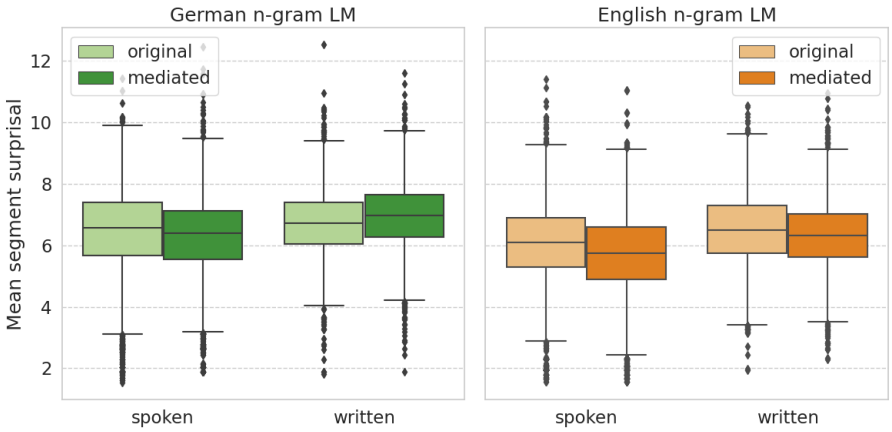

Figure 2: Average surprisal of segments across the subcorpora.

bution, with a higher and sharper peak compared to a normal distribution. The middle 50% of the data are hurdled within a narrow range, with the size of the box (interquartile range) being on average as low as 1.5 bits, while the entire range of values is from 1.54 to 12.52 bits, averaging at about 6.5 bits.

Our hypothesis that the AvS of the mediated texts is significantly lower than that of comparable non-mediated texts can be confirmed with the exception of the German written subcorpus. In the latter, written non-translated documents have lower mean segment surprisal values than translations (6.94 and 6.73 bits, respectively). **H2a** is confirmed for the spoken mode: interpreters produce less informationally dense output than original speakers. However, for the written mode this simplification effect is only seen in English.

The second part of the hypothesis, which expected the range of surprisal values to be smaller in interpreting than in translation, cannot be confirmed (**H2b**). The measures of spread employed in this analysis indicated that in both translation directions interpreted speeches had lower minimum, higher maximum, and higher standard deviation and IQR than translations. For example, interpreted documents into English had a SD = 1.42 and IQR = 1.68, while translations into English had CD = 1.13 and IQR = 1.41. Note that the same relation is seen between the respective non-mediated subsets.

4.3 Impact of Challenging Conditions (H3)

Now we test the hypothesis that the more challenging conditions of simultaneous interpreting such as read-out delivery and higher source speech rate would have a negative impact on the amount of information transmitted by an interpreter.

Figure 3 has the regression lines fitted to the datapoints annotated as 'impromptu' or 'read-out' source delivery. As before, the datapoints are defined by source segment surprisal values on the x-axis and target segment surprisal values on the y-axis. The plots do not show differences between the locations of regression lines for the two types of delivery for either language direction. Even though in the English-German direction the dark grey line for the read-out delivery condition appears below the impromptu line, both lines are within the shadowed area of the confidence interval. Interpreters seem to be able to encode the same level of information regardless of whether the original speaker reads out a prepared speech or speaks spontaneously. The differences in the association strength measured by a correlation coefficient are within the size of the statistical error. These experiments did not yield evidence to support **H3a**.

Figure 4 presents the outcomes of the regression analysis based on the word-per-minute speed of source speeches as the explanatory variable and target segment surprisal as the response variable. Although the regression lines appear to suggest a strong negative correlation between the variables, the Spearman coefficient returned low (but statistically significant) values: -0.06 and -0.09 for German-English and English-German directions. The slope suggests a modest drop of 0.004-0.005 bits for a considerable increase in speed of 100 words a minute. There are visible differences between speech rates in German and English as the source language: this measure might not be equally fair to capture the speed of information input for structurally different languages. Note that the speech rate is measured in words per minute

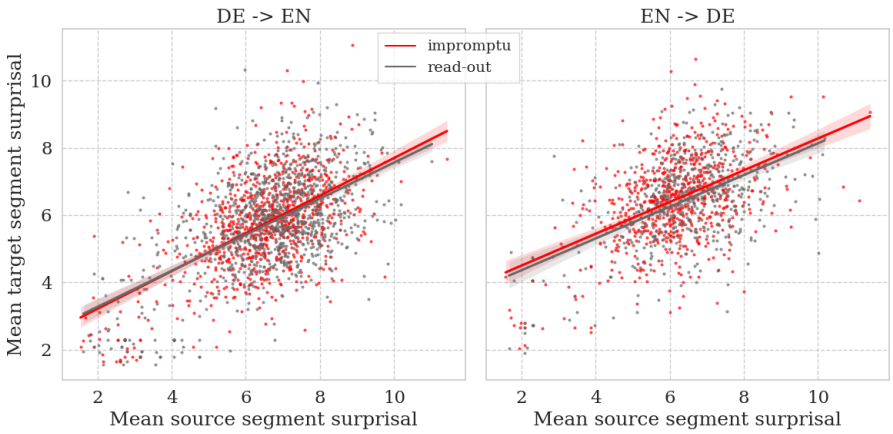

Figure 3: Association between AvS of sources and targets by source text delivery type (impromptu vs. read-out) and translation direction (DE-EN, EN-DE)

and words tend to be longer in German than English (e.g. due to compounding in German). Despite these limitations, both translation directions demonstrate that the higher the source speed, the lower the informativity of the target (confirming **H3b**).

The two parameters analysed in this section can be viewed as independent. The impromptu delivery is expected to display a wider range of spoken features, better aligned with interpreting and online processing. Although in our data, impromptu speeches were delivered at a higher average rate than read-out speeches, they had lower average segment surprisal and lower standard deviation in original speeches as well as in the associated interpreted segments than for the read-out speeches in both German and English.

The current experimental setup did not yield the theoretically expected results with regard to the special conditions in interpreting. It can be an indicator that the exploited language model lacked skill and subtlety or that some categories in this analysis are severely underrepresented. For example, the number of segment pairs in English originals annotated for slow speech rate (under 130 wpm) was only 104 (vs 2,313 segment pairs marked with 'high' speech rate).

5 Discussion

We have established that the information density of the target is strongly and positively correlated with the information density of the source in both mediation modes, spoken and written. However, the information output in interpreting is lower than in translation given the same input: the intercept of the regression lines for interpreting is lower in both translation directions (see the legends in Figure 1). To demonstrate the differences between translation and interpreting, we looked at the top and bottom segment pairs by target surprisal in EPIC-UdS and their translated alternatives from Europarl-UdS. Example (1) demonstrates that translation follows the German source more faithfully than the interpreted version, where the last coordination is omitted, making the output less informative.

(1) SOURCE: *Europa muss lernen, mit einer Stimme zu sprechen* ***und dann auch mit einer Position zu handeln.***
TRANSLATION: *Europe must learn to speak with one voice* **and to take united action**.
INTERPRETING: *Europe must learn to speak with one voice.* (AvS = 5.52)

In Example (2), the explicit description of an issue, given in the source and faithfully retained in translation, is replaced with a generic anaphoric phrase (*this sort of thing*), and the more specific word *Bürger (citizens)* is replaced with a general noun, *people*.

(2) SOURCE: ***Die Belastungen durch die stetig steigende Zahl illegaler – ich betone illegaler – Einwanderer****, sind auf Dauer für die EU- Bürger untragbar.*
TRANSLATION: ***The burden represented by the constantly growing number of illegal immigrants – I would like to emphasise the word 'illegal' here –*** *is becoming unbearable for the citizens of the EU.*

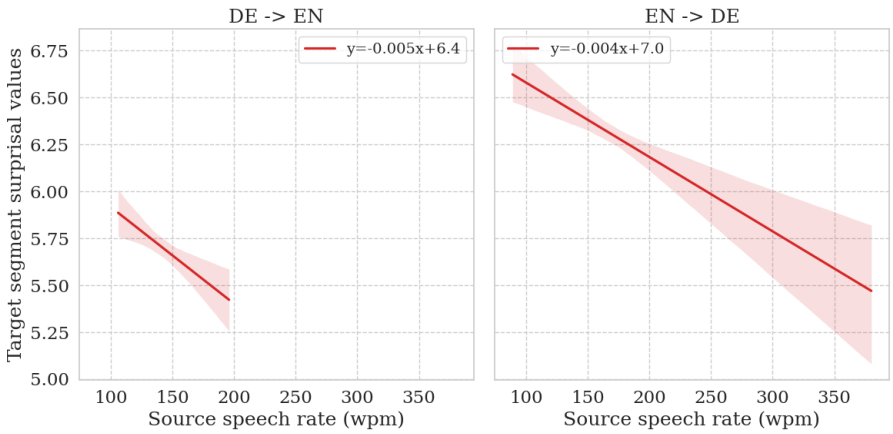

Figure 4: Regression plots: relation between target segment surprisal and source speech speed in words per minute (for two translation directions).

INTERPRETING: **And this sort of thing** *is an unsustainable situation in the EU and for people of the EU.* (AvS = 5.24).

The surprisal values for each token in the interpreted segment from Example (2) are shown in Figure 5. The lineplot demonstrates how simpler structural and lexical content in interpreting (as compared to translation) keeps the AvS low.

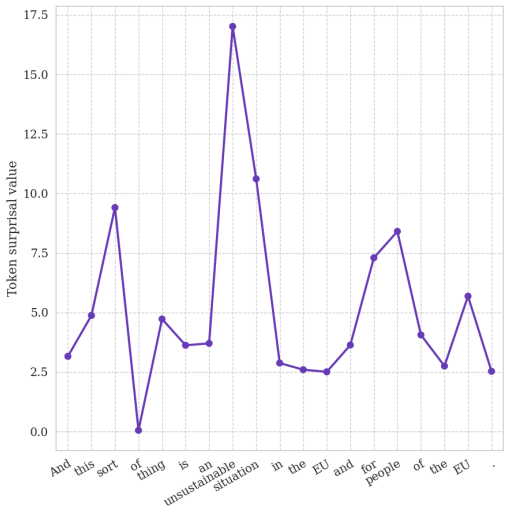

Figure 5: Token surprisal values in the interpreted segment with low AvS from Example (2).

The powerful simplification trend, which is reinforced by the spoken features of interpreting and which pulls the AvS in interpreting down, is counteracted by the tendency to follow source segment patterns, which generates a shining-through effect. It can be manifested in the use of cognates, unusual verb constructions, or as in Example (3) unexpected noun phrases.

(3) SOURCE: *One in four Europeans suffer from* ***mental health problems*** *at least once during their life.*
TRANSLATION: *Ein Viertel aller Europäer leidet mindestens einmal in dem Leben unter* ***psychischen Problemen****.*
INTERPRETING: *Jeder vierte Europäer leidet zumindest ein Mal in seinem Leben unter einer* ***geistigen Krankheit****.* (AvS = 9.76)

Similarly, the interpreted segment from Example (4) has a surprisal peak at the end of the sentence. It is generated by the word *complaints* in an unusual context, which was most likely an erroneous word choice.

(4) SOURCE: *Wollen wir den Chinesen mit WTO* ***Klagen*** *drohen.*
TRANSLATION: *Do we want to threaten the Chinese with World Trade Organisation (WTO)* ***sanctions****?*
INTERPRETING: *You know are we going to threaten the Chinese with WTO* ***complaints*** (AvS = 6.57).

Based on our results, rejection of **H2b** might be explained by the intensity of the two opposite trends that increase the spread of the surprisal values in interpreting. On the one hand, interpreters have a strong tendency to select simpler, more frequent vocabulary and fill pauses with highly expected phrases, which decreases mean segment surprisal.

On the other hand, interpreting can demonstrate more noticeable forms of interference and lack of fluency that would generate increased segment surprisal.

Finally, to ascertain that AvS values are aligned with intuition, we looked at the results for segments that were either omitted or added in interpreting. Typical segments that are skipped in our sample are the politeness formula and discourse organisation markers. For example, the interpreter omitted segments like the following: *Sehr geehrter Herr Präsident. (EN translation: Mister President.)* (AvS = 3.23), *Ich komme dann zu dem Ende. (EN translation: I am coming to the end.)* (AvS = 4.99), *Finally just to sum up very briefly an old saying.* (AvS = 5.42), *Let us be very clear.* (AvS = 4.17). A more curious case are additions, i.e. segments that were not aligned to any content on the source language side. These segments typically reiterated the speaker's emphasis and included short segments like *Aber was sollte man jetzt tun. (EN translation: But what should be done now.)* (AvS = 6.10), *Aber so ist es. (EN translation: But that's how it is)* (AvS = 5.14), *That is the thing.* (AvS = 3.88), *So here we have to speak out.* (AvS = 4.00). The AvS for omitted and added segments was lower than the average across all segments in both language directions in EPIC-UdS (6.31 and 5.69 for interpreted German and English, respectively). This means that the attempted modelling setup supports some theoretical expectations if not others.

Overall, a manual analysis of token surprisal values in various subsets of data demonstrated that an n-gram model trained on limited data might be too constrained by the amount of available corpus evidence to rely on its output for a fine-grained analysis of translational phenomena. However, surprisal contours are a good source for qualitative checks of statistical results. All else being equal, the German model returned higher surprisal values and perplexities, either suggesting a lower quality than that of the English model or simply a language-specific feature. Overall, the proposed modelling approach might be biased toward producing middle-range surprisal values (evidenced by a sharp-peak distribution with thin tails), partly because it assigns the same probability to all hapax legomena and uses a simple back-off to a lower-order n-gram to resolve the out-of-vocabulary issue.

6 Summary and Conclusion

The study demonstrated that mean segment surprisal values capture the distinction between non-mediated and mediated language for three out of four parallel subcorpora: mediated language has lower surprisal. Importantly, this difference can be interpreted as an indicator of simplification: mediated language is characterised by a lower information density than comparable non-mediated segments. It is particularly true for interpreting, as seen from our analysis of the association between sources and targets. This, however, does not affect the strong positive correlation between the information density of sources and targets, seen in this study for all parallel subcorpora. Contrary to our expectations, transcripts of interpreted documents had a higher variability of segment surprisal values than in translation, making their information density less predictable from that of the source segment.

The choice of the research method in this study was largely determined by the small size of the data available for modelling if we wanted to train on a balanced corpus (12 K segments, ca. 250 K tokens in each language). The parameters of the surprisal distributions suggest that the current modelling approach might be sub-optimal. In future work, we plan to explore other modelling approaches compatible with small-size datasets to obtain a more faithful representation of information density in a segment and across the segments. The ultimate goal of modelling surprisal is to apply information theory to the explanation of linguistic choice in mediated communication linking it to the availability of cognitive resources that can be more or less engaged depending on the properties of the source segment, context, mediation mode, and extralinguistic conditions of the information transfer. This goal calls for multilingual models, on the one hand, and for more fine-grained qualitative analysis, on the other. We believe that the interpreting data — represented by accurate transcripts of spoken sources and their targets, including disfluencies — is particularly suited for these purposes and for understanding the mechanisms of human speech generation, in general.

Acknowledgments

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1102 / Project-ID 232722074.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Christoph Aurnhammer, Francesca Delogu, Miriam Schulz, Harm Brouwer, and Matthew W. Crocker. 2021. Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLoS ONE*, 16(9):e0257430.

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–245.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers, editor, *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pages 175–188. John Benjamins, Amsterdam and Philadelphia.

Marco Baroni and Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Measuring translationese across levels of expertise: Are professionals more surprising than students? In *Proceedings of the 23rd NoDaLiDa*, pages 53–63, Online. ACL.

Shoshana Blum-Kulka. 1986. Shifts of Cohesion and Coherence in Translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication*, pages 17–35. Gunter Narr, Tübingen.

Agnieszka Chmiel. 2021. Effects of simultaneous interpreting experience and training on anticipation, as measured by word-translation latencies. *Interpreting*, 23(1):18–44.

Ingrid K. Christoffels, Annette M.B. de Groot, and Judith F. Kroll. 2006. Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language*, 54(3):324–345.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2015. Information density and linguistic encoding (ideal). *KI - Künstliche Intelligenz*, 30(1):77–81.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Stroudsburg, PA. Association for Computational Linguistics.

Marta Kajzer-Wietrzny. 2012. *Interpreting universals and interpreting style*. Ph.D. thesis, Adam Mickiewicz University, Poznan.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Europarl-UdS: Preserving metadata from parliamentary debates. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4102–4112. European Language Resources Association.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

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

Matthew W. Lowder, Wonil Choi, Fernanda Ferreira, and John M. Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, 42(S4):1166–1183.

José Manuel Martínez Martínez and Elke Teich. 2017. Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Larissa Cercel, Marco Agnetta, and Maria Teresa Amido Lozano, editors, *Kreativität und Hermeneutik in der Translation*, pages 403–427. Narr Francke Attempto Verlag.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022. EPIC-UdS - creation and applications of a simultaneous interpreting corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1193–1200, Marseille, France. ELDA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of NAACL HT 2006*, pages 960–970, San Diego, California.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Miriam Shlesinger and Noam Ordan. 2012. More spoken or more translated?: Exploring a known unknown of simultaneous interpreting. *Target. International Journal of Translation Studies*, 24(1):43–60.

Elke Teich. 2003. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Elke Teich, José Martînez Martînez, and Alina Karakanta. 2020. Translation, information theory and cognition. In Fabio Alves and Arnt Lykke Jakobsen, editors, *The Routledge Handbook of Translation and Cognition*, chapter 20. Routledge, London.

Gideon Toury. 1995. *Descriptive translation studies and beyond*. Benjamins translation library: 4. Benjamins.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.