Information Divergence in Translation and Interpreting: Findings from Same-Source Texts

Maria Kunilovskaya[1], Sharid Loáiciga[2], Ekaterina Lapshinova-Koltunski[3]
[1] Department of Language Science and Technology, Saarland University
[2] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
[3] Department of Language and Information Sciences, University of Hildesheim
maria.kunilovskaya@uni-saarland.de
sharid.loaiciga@gu.se, lapshinovakoltun@uni-hildesheim.de

Abstract

The present paper reports on differences between written and spoken translation in English and German using Kullback-Leibler Divergence (KLD), an information-theoretic measure that captures the divergence of a given probability distribution from a reference distribution. While most studies comparing translation and interpreting focus on the target side only, we also take into account the source texts: we analyse differences between written and spoken translation renditions of the same source texts. The innovation of our approach is in the use of a source language model as an approximation of a cross-lingual modelling approach. Our results confirm the tendencies reported in previous work without same-source-text settings—namely, that interpreted targets differ more from their sources than written translations.

1 Introduction

In this study, we apply an information-theoretic approach to examine the differences between written and spoken modalities of translation. These differences have been widely documented from a translation studies perspective, highlighting the unique characteristics of each modality. However, previous studies have focused mostly on target-side comparisons between modalities. In contrast, our study employs parallel data with identical source-side texts, enabling a comprehensive analysis of both source- and target-side materials.

Information-theoretic measures such as surprisal or relative entropy are known to be predictive of language processing complexity, which involves not only the difficulty on the recipient's side but also translation production difficulty. The latter has been previously measured in a cross-lingual model setting only (Lim et al., 2024). We adopt two complementary approaches to assess modality differences: one relying on monolingual source and target language models, and a cross-lingual one relying on Neural Machine Translation (NMT) models. In both cases, we derive probability distributions from the models for each of the modalities and then measure the distance between them using Kullback-Leibler Divergence (KLD). While target language models provide a good estimate of the likelihood of the test set in the target language, we demonstrate that a source language model also serves as a good proxy for evaluating the differences between modes in the target language. So, while KLD scores on the target language models are indicative of the recipient's translation processing difficulty, the KLD scores obtained on the source language model give us an approximation of translation processing difficulty on the producer's side.

Analysing translation and interpreting involves not just two modalities, but two fundamentally different approaches to the translation process. Translators benefit from asynchronous access to the source text and supplementary resources such as dictionaries, whereas interpreters face the challenge of producing live, incremental translations under strict time constraints. The medium differences entail cognitive and processing challenges for interpreters.

By estimating KLD, we model the cognitive processing of translators and interpreters as producers. Also, unlike previous studies that only compared target-side differences, our approach introduces a novel approximation of the information transfer from source to target (see more details in Section 2.1). Taking into account the source text effectively models the interpreter's role in mediating the content. Furthermore, we analyse the linguistic cues that trigger these differences.

Our study is driven by three research questions:

RQ1 Is a source language model as sensitive in capturing differences across written and spoken modalities as a target language model?

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 192–202, 2025

RQ2 How do spoken and written target versions diverge from the same source texts across translation directions?

RQ3 What are the main qualitative differences observed between written and spoken targets with respect to their sources?

The remainder of the paper is organised as follows: in Section 2 we discuss the related work, Section 3 describes the methodology. Last, we present results in 4 and conclude in Section 5.

2 Related Work

2.1 Spoken and written translation

Differences between simultaneous interpreting and written translation have been shown in previous studies (see e.g. Kajzer-Wietrzny, 2012). From the studies focusing on the language of translation (translationese or interpretese), we know that translation and interpreting are both clearly different from comparable originals, i.e., texts originally authored in the target language. At the same time, Shlesinger and Ordan (2012) show that interpreted texts seem to reinforce the features of spoken language. As a result, interpreting exhibits more marked linguistic characteristics than other kinds of texts and, in this way, is the most distinctive type of language production. Information-theoretically inspired features used by Lapshinova-Koltunski et al. (2021) support this statement. The authors employ the features obtained by KLD (based on the study by Przybyl et al., 2022a) in addition to other linguistic features. Interpreting resulting in higher scores than translation confirms that interpreting is the most distinctive kind of language production.

In this paper, we also aim to use information-theoretic measures which allow for a link between language use and cognition. Translationese (and also interpretese) features are understood to reflect rational communication: translators tailor the language encoding so as to translate successfully while keeping their cognitive effort at a reasonable level. Such measures, e.g., surprisal (Shannon, 1948), perplexity, and (relative) entropy have been used in other studies of translationese (Yung et al., 2023; Bizzoni and Lapshinova-Koltunski, 2021; Rubino et al., 2016) with a focus on written translation.

Existing interpretese studies that use information-theoretic measures mostly focus on specific linguistic items, e.g., discourse connectives (Lapshinova-Koltunski et al., 2022). However, only a few of them (Pollkläsener et al., 2024; Kunilovskaya et al., 2023; Yung et al., 2023) consider not only the scores of the comparable target texts, but also the underlying sources. Kunilovskaya et al. (2023) compare the input and output differences in translation and interpreting focusing not on specific linguistic items, but on the overall score of segments. The authors measure surprisal monolingually for source and targets and looked into the surprisal correlation.

2.2 Monolingual modelling

Thanks to their scale, large language models (LLMs) can perform a wide range of tasks in zero- or few-shot settings. Even early LLMs like GPT-2 showed the ability to translate text despite being trained solely with the standard next-word prediction objective (Radford et al., 2019). Building on this, Han et al. (2021) show that monolingual LLMs can be used to generate synthetic training data, which can then be used to fine-tune other models for machine translation. Similarly, Garcia et al. (2023) demonstrate that decoder-only LLMs can achieve strong translation performance when trained on just a handful of examples.

2.3 Cross-lingual modelling

In the context of machine translation, information theoretical approaches have been used to capture subtle distinctions in the source texts. Xu et al. (2021), for instance, integrate bilingual mutual information into the objective function in order to generate target tokens that reflect their source tokens better than their default target distributions. Similarly, Zhang et al. (2022) uses mutual information between a target token and its source sentence conditioned on the target contexts to measure the importance of different target tokens by their dependence on the source sentence. Contrary to our approach, these works estimate mutual information at the token level, while we use it at the sentence-level. On its part, Bugliarello et al. (2020) use a statistical machine translation approach to use target side language model probabilities and translation table probabilities to compute cross-lingual mutual information to test whether it is easier to translate from English or into English.

Lim et al. (2024) define translation surprisal as $s_{mt}(y_j) = \sum_{i \in j} -\log \big(p_{mt}(y_i \mid x, y_{<i})\big)$, where x and y are parallel source-target segments, and probabilities $p_{mt}(y_i \mid x, y_{<i})$ are extracted for each target token y_i, given the entire source x and pre-

ceding target tokens $y_{<i}$ from a Neural Machine Translation (NMT). The show that this measure can serve as a complementary predictor of human translation difficulty. These model-derived scores correlate with behavioural indicators of processing difficulty, such as reading time during translation. They also find that NMT-based surprisal is the strongest individual predictor of production duration.

3 Methodology

In our approach, we model aligned source texts, written translations, and spoken translations as *probability distributions over sentences*. To quantify the differences between these sentence-level distributions, we use Kullback-Leibler Divergence (KLD), an information-theoretic measure that captures the divergence of a distribution Q from a reference distribution P. Specifically, it measures how many additional bits of information are required to encode samples from P using a code optimised for Q.

Mathematically, for discrete distributions P and Q over a set of aligned sentences $\mathcal{S}$, KLD is defined as:

$$D_{\mathrm{KL}}(P\,||\,Q) = \sum_{s\in\mathcal{S}} P(s)\log\left(\frac{P(s)}{Q(s)}\right) \tag{1}$$

We compare probability distributions represented by sentence-level probabilities of aligned source texts, written translations, and spoken translations, treating them as alternative encodings of the same message. Depending on the comparison type (see Table 1), we take either the source texts or the written translations as the reference distribution P, and the written/spoken translations or spoken translations as the comparison distribution Q, respectively. In this way, the resulting KLD values quantify, for example, the excess unpredictability of the spoken rendition relative to the written modality. In other words, KLD measures how much information (in bits) is required to encode the sentence-level distribution of the spoken version Q when the model is optimised for the written distribution P.

Our main hypothesis is that the interpreting process inherently leads to greater changes in the amount of conveyed information compared to the source text, due to the demanding cognitive conditions under which interpreters work. In contrast, translations are expected to be more faithful to the source text, as translators can work without the pressure of live and incremental input and output. Written translations, unlike online spoken translations (interpreting), can also be revised and edited at a later time. Therefore, we anticipate that

(i) all models will detect the difference between spoken and written encodings of the same message, including source language models when applied to text in the target language, and (ii) spoken targets will exhibit greater divergence from the source texts than written translations.

To compute KLD between sentence-level probability distributions in our dataset, we used SciPy's `entropy` function with a base-2 logarithm, yielding results in bits. In addition to the aggregate KLD (1), we also computed the *pointwise* contribution of each sentence pair to the divergence.

$$pw\_D_{kl}(s) = P(s)\log_2\left(\frac{P(s)}{Q(s)}\right) \tag{2}$$

This weighted log-ratio of sentence probabilities (which we refer to as *pointwise KLD* or *pseudo-KLD*) enables us to track local variation within the dataset, identify high-contributing instances, and better interpret the overall KLD values obtained from different models. Unlike the true KLD, which is always non-negative, the pointwise value can be negative when $Q(s) > P(s)$.

The next two sections detail our method for estimating sentence-level probabilities and explain how KLD was used to compare the two mediation modes (interpreting, translation), viewed here as different message encoding types.

3.1 Comparing sentence probabilities

The main goal of this study is to compare the differences across translation modalities while taking into account the source text. The source text is known to be a strong factor that influences the properties of the output in any cross-lingual transfer. Two alternative types of models were used to represent our data: monolingual and cross-lingual. The cross-lingual approach inherently conditions output probabilities on the source language input, but it does not allow for estimating the distance between sources and targets using KLD. As a proxy for a cross-lingual approach, we propose to use a source language model on target language texts in each mode and compare the resulting sentence probability distributions based on our set of sentences. If the source language model is approximately as good as the target language model in capturing the divergence of spoken targets from the written targets, then we can apply it to measure the divergence of targets from sources for each modality.

We use the dedicated German and English GPT2 models (Schweter, 2020; Radford et al., 2019) for monolingual processing and translation-direction-specific MarianMT machine translation models (Tiedemann and Thottingal, 2020) for cross-lingual inference. The language- and translation-direction-specific models are preferred over multilingual alternatives to avoid a potentially confounding factor of other languages in the model. The four selected models have approximately the same vocabulary size of 50-58 K tokens. Importantly, we aggregate sentence probabilities from the probabilities of word-tokens defined by Stanza (Qi et al., 2020) and not from the probabilities of subword-tokens as defined by tokenisers of the respective models. This allows us to keep German and English monolingual models comparable across the experiments (especially when applying a source language model to texts in the target language) and to introduce a common ground for comparison between monolingual and cross-lingual models. The sentence probabilities are obtained by summing up word probabilities in log space and then exponentiating the result.[1]

To sum up, our experimental setup includes three levels of comparative analysis in each translation direction. In Table 1, the Model Type column specifies the model from which the probabilities for the KLD calculations were obtained.

Level	Model Type	Input Pair (P vs. Q)
a	TL model SL model	Written target vs. Spoken target Written target vs. Spoken target
b	SL model SL model	Written source vs. Written target Spoken source vs. Spoken target
c	NMT model	Written target vs. Spoken target

Table 1: Three levels of comparative analysis using language and translation models. In each pair, the first text type is treated as the reference distribution P.

3.2 Data

Our multiparallel subset of written and spoken renditions of the same original texts is an intersection of the EuroParl-based corpora of written and spoken translations: Europarl-UdS (Karakanta et al., 2018) and EPIC (Przybyl et al., 2022b), respectively. Europarl-UdS contains edited transcripts of the speeches and their official translations as made available on the EP website, EPIC contains true manual transcripts of the debates in the European Parliament and their simultaneous interpretation as captured in the video recordings of the debates. All data is limited to English-to-German and German-to-English translation directions and only includes original speeches delivered by native speakers of English or German, respectively. For each translation direction, we identified source-target document pairs where the source language speeches were the same. Speaker, date and document size parameters were used as a coarse filter. The results were further refined using pair-wise cosine similarity of the document vectors (cutoff 0.8) derived from an off-the-shelf embedding model trained on a semantic textual similarity task (Reimers and Gurevych, 2019). However, due to the differences between written and spoken originals based on the same real-life speech (mostly due to dissimilar sentence alignment in each corpus), it was impossible to get written and spoken mediated versions for every source sentence in the multi-aligned documents. Therefore, the experimental data were further limited to include written and spoken sentence pairs where the source sentences had a Levenshtein distance of over 0.8. To account for possible rearrangement of sentences in the transcripts, the comparison was run in the context window of 10 sentences. To balance the dataset across translation directions, we selected five documents from the larger German-to-English set, each containing 70 reliably multi-aligned sentence pairs, ensuring that both directions had the same number of sentences. This approach preserved the topical coherence present in the English-to-German set, rather than taking a random sample of 70 sentences from across many documents.

The quantitative description of the final multi-parallel test dataset is given in Table 2.

Note that throughout the paper, we use ‘sentence’ and ‘segment’ interchangeably, but the true unit of analysis is the segment rather than the sentence, as it is the alignment unit: in parallel data, source and

[1] The aggregated probabilities from the English GPT2 model as well as for EN-DE NMT model for two German sentences, each containing more than 50 words, encountered a numerical underflow issue during the exponentiation step, even when using double-precision floating-point calculations. As a result, the values are set to a floor value of $1e - 250$.

direction	docs	segs	src_tokens
DE-EN	5	70	1,478
EN-DE	5	70	1,789

Table 2: Parameters of the datasets. Token counts are given for Stanza-tokenised sources (src), in multi-parallel data based on the written version of the source.

target sentences can align in a one-to-many manner, especially in spoken data.

4 Results and Discussion

4.1 RQ1: Differences across modalities

First, we investigate whether monolingual source language models can capture differences between spoken and written translations when measured using KLD, in the same way that target language models do. This provides us with a proxy of translation production, i.e., taking the source language differences into account. In this way, we approximate a cross-lingual approach but without using a cross-lingual or multilingual model.

Table 3 presents the KLD results based on the segment probabilities inferred using the out-of-the-box monolingual GPT models of the source and target languages applied to 70 pairs of written and spoken translations of the same sources in each translation direction.

direction	SL model	TL model
DE-EN	9.130 (de_gpt2)	7.468 (en_gpt2)
EN-DE	0.013 (en_gpt2)	0.697 (de_gpt2)

Table 3: KLD results on monolingual models of source and target languages as defined in Table 1 (row a): KLD shows divergence between written and spoken targets of the same source. Source and target language models are applied to the text in the target language.

As seen from Table 3, the source language model is as sensitive to the differences in written and spoken translation as the target language model: KLD values in rows are are of similar magnitude. This confirms that written and spoken translation differ, including when the source language is implicitly taken into account by applying the source language model to the translation in the target language. This also confirms the validity of the approach to use a source language model as a proxy for a cross-lingual approach to analyse differences between spoken and written translation.

As a sanity check, we calculate the perplexities of the same models on the same data. Perplexity provides an independent measure of how "surprised" a model is by the data: if a model assigns very low probability to the observed tokens, this is reflected in a high perplexity. Comparing KLD with perplexities allows us to confirm that observed differences between written and spoken translations are meaningful and not driven by degenerate probability distributions.

For each segment, we calculate the sentence-average negative $\log_2$ probabilities, also known as average sentence surprisal (AvS), and then average these across 70 instances in each sample. The result is exponentiated to yield the perplexity, as shown in Equation 3. Segment perplexities are combined using the geometric mean, which is more robust to outliers and better represents the central tendency.

$$\text{PPL} = 2^{-\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{T_s}\sum_{t=1}^{T_s}\log_2(p_{t,s})\right)} \tag{3}$$

As seen in Table 4, perplexities for subword tokens are consistently higher for spoken segments than for written segments across both source- and target-language models. This indicates that spoken translations are systematically harder for the models to predict, and, importantly for this study, are systematically treated as different from the written translation in absolute terms, validating our KLD-based approach, which is a relative difference measure.

	model	mode	subword	word
DE-EN	SL	written	56.18	331.07
		spoken	83.31	392.55
	TL	written	61.18	68.26
		spoken	75.85	86.13
EN-DE	SL	written	77.43	45165.76
		spoken	88.79	32002.15
	TL	written	52.04	85.80
		spoken	105.68	165.89

Table 4: Geometric mean segment perplexities of the source- and target-language GPT2 models (SL and TL model) for the written and spoken target versions (written and spoken mode) calculated for subword- and word-based segments.

Additionally, Table 4 reports perplexities for word-level tokens, since KLD values were calculated from Stanza-token probabilities, which in turn were obtained by aggregating probabilities of constituent GPT-4 subwords, and GPT-4 tokenisation strongly affects the results, especially in our unconventional inference setup. As expected, all written and spoken target language segments contain more subwords than words when processed by the source-language model.

The very low KLD between written and spoken German in the EN-DE direction (0.013) can be better understood in light of the perplexity results. From the perspective of the English source-language model, both written and spoken German segments are highly improbable (for this test set), which leads to very high perplexities. Because KLD measures relative divergence, the distributions of written and spoken German appear very similar to the English model, even though the absolute probabilities are extremely low. In other words, the low KLD reflects *relative* similarity between written and spoken segments, while the high perplexity highlights the poor adaptation of the source-language model to target-language text. The inclusion of a single very long segment (>75 words) further exaggerates the perplexity, but it does not affect the interpretation of KLD as a measure of relative differences.

We compare the KLD indices obtained with the source language model (SL model in Table 3) with the results from the cross-lingual NMT model, i.e., a cross-lingual setting. The KLD indices estimated from the language pair-specific NMT model are shown in Table 5.

direction	NMT model
DE-EN	35.543
EN-DE	5.175

Table 5: KLD for written and spoken target sentence probabilities from the NMT models as defined in Table 1 (row c).

The cross-lingual approach returns a similar result for the directions comparison: the divergence of spoken from written is greater for the German-to-English direction than for English-to-German (see SL model results in Table 3).

Overall, the KLD values reported for the two translation directions in our analysis are derived from language- or language-pair-specific models (e.g., en_gpt2, de_gpt2, and the DE–EN and EN–DE NMT systems). Even for structurally similar sentences, these models may assign systematically different probabilities, making direct comparisons of absolute KLD magnitudes across languages or directions unreliable. In particular, higher aggregated KLD values for English targets do not necessarily indicate a larger true divergence; they may reflect model-specific scaling effects.

To provide a more nuanced view, we plot the pointwise KLD values for paired written and spoken targets in our sample (Figure 1). These plots offer a comparable qualitative signal: for English as the target language (upper panel), the curve crosses zero more frequently, indicating greater variability in how the models assign probabilities to the spoken versus written versions. In other words, the pointwise differences are more volatile and less systematic, with frequent sign flips. By contrast, for German targets, probabilities for written versions tend to dominate more consistently, resulting in fewer zero crossings and lower variability in the curve.

4.2 RQ2: Source language divergence

Next, we examine how spoken and written target versions diverge from the same source texts. We measure the differences between these two modes using KLD against the shared source segments, applying source-language models that, as shown in Section 4.1, can capture mode differences similar to the target-language models.

direction	mode	SL model
DE-EN	written	44.881 (de_gpt2)
	spoken	63.949 (de_gpt2)
EN-DE	written	0.051 (en_gpt2)
	spoken	0.098 (en_gpt2)

Table 6: KLD results on monolingual models of source languages as defined in Table 1 (row b): cross-lingual divergence between sources and their targets depending on the mode. Source language model is applied to the text in the source language and its rendition in the target language.

Table 6 provides an overview of the KLD scores calculated for both translation directions. Higher scores indicate greater divergence between the probability distributions of the source text and its target rendition as estimated by the source-language model. In other words, the higher the KLD score, the more the statistical profile of the translation or interpreting output differs from that of its source.

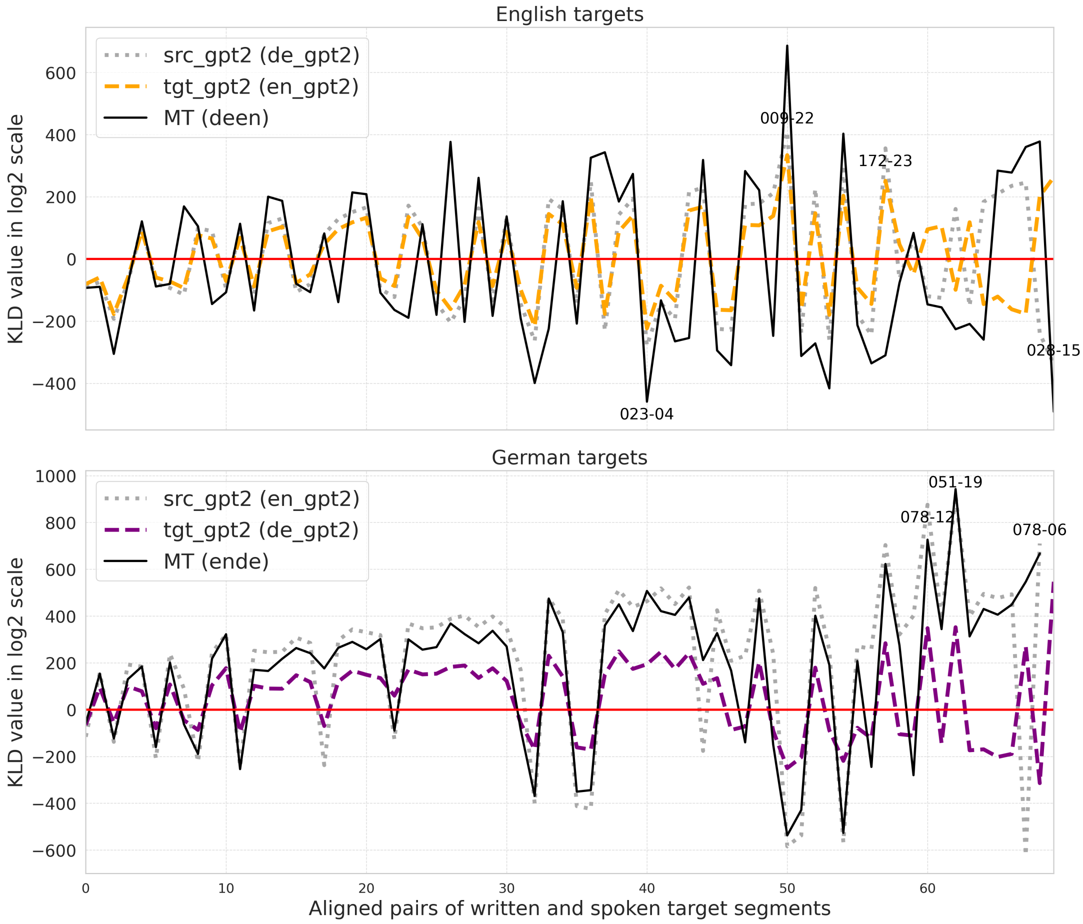

Figure 1: Comparison of GPT and MT-based KLD values (in log2 scale) across all instances in the sample. The instances are sorted in ascending order by the absolute KLD difference between source and target GPT2 models. The top three instances by KLD values based on source GPT2 and/or MT are shown on the x-axis.

Since we are comparing the same source texts and their two target-language versions, no differences in KLD scores would be expected if translation and interpreting were identical. However, the third-column scores in Table 6 do differ (44.881 vs. 63.949 and 0.051 vs. 0.098). Crucially, across both translation directions, the magnitude of divergence of target from source in spoken renditions is significantly higher than in written translations. For DE-EN, the median difference in pseudo-KLD magnitudes (i.e., absolute values) is 10.22 (Wilcoxon $p \approx 4.6 \times 10^{-7}$, permutation p = 0.0002), while for EN-DE, the median difference is 1.45 (Wilcoxon $p \approx 0.003$, permutation $p \approx 0.0032$), indicating that spoken translations consistently exhibit greater deviation from the reference distributions.

4.3 RQ3: Linguistic differences

And finally, we manually compare the spoken and the written translations of 70 input segments in both languages aiming to observe specific differences sentence by sentence. We do not collect any quantitative information on the type of differences, but report our observations only. We also visualise the contrasts between written and spoken versions of one segment from our data in Figure 2 which shows diverging word surprisal values for each word in the segment.

Overall, we observe that translators and interpreters apply different segmentation, with interpreting using more simple sentences instead of one long with subordinates. As a result, interpreting contains more repeated words or paraphrases than translation does.

DE-EN Observations:

- Interpreting uses more general words instead of specific ones: *to make use* vs. *to utilise*, *future generation* vs. *our children* (Figure 2). Further examples include (but not limited to):

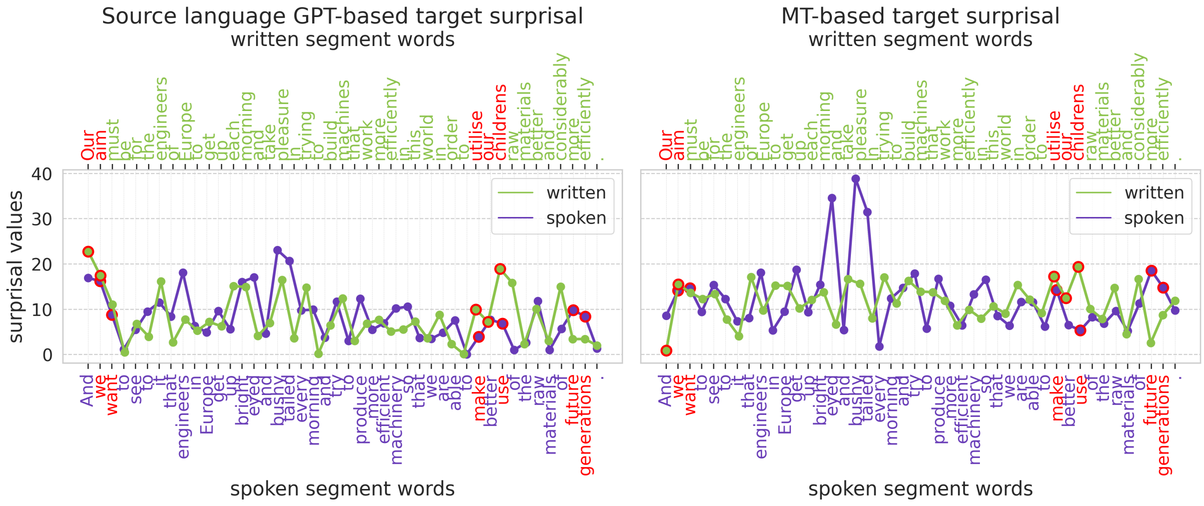

Figure 2: Word surprisals derived from GPT and MT models for DE_EN_028-15 instance. Contrastive written and spoken solutions are highlighted in red.

spoken	**written**
run up	*history*
Commissioner	*Mrs Fererro-Waldner*
people	*citizens*

- Interpreting prefer to use active instead of passive voice such as *we will be looking at proposals* vs. *proposals should be submitted.* Moreover, it has more verbal instead of nominal phrases: *we want* vs. *our aim* (Figure 2) or further cases:

spoken	**written**
the Commission's assessment	*the Commission's assessment*
people got to trust each other	*trust between the states*
we need	*an essential factor*
is very close to	*its geographical proximity*

- Furthermore, there is a difference in the speaker-addressee perspectives (3rd vs. 1st person or 2nd vs. 1st person reference) in both translation variants, although no systematic pattern could be observed:

source *Europa steht im internationalen Wettbewerb, in einem globalen Wettbewerb, in dem sich auch die europäischen Dienstleister positionieren und durchsetzen müssen.*

spoken *Europe is facing global competition, in particular in the area of services. And we need to defend our position.*

written *Europe faces international competition, global competition, in which European service providers also need to find their place and be successful.*

- At the same time, second person reference is used by an interpreter in a different case, where the translator prefers to use the first person reference (*you* instead of *we*).

source *Wenn wir uns die Vorgeschichte dieser Richtlinie...ansehen...*

spoken *If you look at the run up to this Directive.*

written *If we look at the history of this directive.*

EN-DE Observations:

- Similarly to the EN-DE pair, German interpreting contains more general words instead of specific ones. The written variant is not only more specific, but also more literal and closer to the source, while the spoken one is more general:

spoken	**written**
klappen	("to function")
("to work out")	*funktionieren*
diese Listen	*Kriterien für diese Barometer*
("these lists")	("criteria for these barometer")

Sometimes, details are omitted in interpreting and added in translation, and a general word is used for several terms:

source *...winter cereal crops susceptible to weeds and disease such as blight.*

spoken *...bei dem Winterweizen werden da Seuchen eingeführt.*

written *...Wintergetreide anfällig für Unkraut und Krankheiten, beispielsweise Pilzerkrankungen wie Mehltau oder Braunfäule.*

Another interesting example here is the translation of the word *pollination*. The translator used a specification with a prepositional phrase *Bestäubung von Pflanzen mit Pollen* ("pollination of plants with pollen") and the interpreter used simply the word *Polliniserung* instead, which is not common in German.

- German interpreting also prefers use of verbs over nominal phrases:

spoken	**written**
versuchen	*zur Aufgabe machen*
("try")	("to make a task")
aufpassen	*Gewissheit darüber haben*
("pay attention")	("to have certainty")

- Apart from this, we observe a more frequent use of relative clauses in interpreting, e.g. the source nominal phrase *a non-scientific-based definition* corresponds to *Definition, die nicht wissenschaftlich ist* in the spoken version, and to *wissenschaftlich nicht fundierte Definition* in the written one.

  Another case of use of relative clause is the translation of *adequate prevention strategies*: the written rendition is a literal translation. The corresponding interpreting *vorbeugende Strategien, die dagegen vorgehen* ("preventive strategies that work against it) contains the relative clause *die dagegen vorgehen* ("that work against it") which is repetitive and redundant as it conveys the same information as the modifying participle *vorbeugend* ("preventive").

- We also observe omission of some details like in the following example. The German interpreted version does not contain any information on skills and training, while the written rendition does contain this information that was left out in interpreting:

source *Secondly, the whole area of innovation about looking to see where jobs are going to come into future and making sure people have the skills and the training for that*

spoken *Dann innovative Maßnahmen, welche Arbeitsplätze für die Zukunft von Bedeutung sind und wie man die weiter ausbauen kann.* ("Then innovative measures, which jobs are important for the future and how they can be further expanded.")

written *Zweitens, der gesamte Bereich der Innovation und die Aufgabe, festzustellen, woher die Arbeitsplätze in der Zukunft kommen, und sicherzustellen, dass die Menschen die Fachkenntnisse und die Ausbildung dafür haben.*

The analysis of the concrete linguistic differences between spoken and written translation confirms the observations made in other existing studies (Lapshinova-Koltunski et al., 2021; Przybyl et al., 2022a; Shlesinger and Ordan, 2012). However, while most previous works were based on comparable corpora of translation and interpreting, our qualitative analysis is based on parallel segments with one underlying source.

5 Conclusion and Discussion

This study applies information-theoretic measures, known to be predictive of translation processing complexity on both producer and recipient's side, to analyse differences between translation and interpreting.

Our findings show that these differences can be effectively captured using both monolingual and multilingual (i.e., cross-lingual) approaches. We also demonstrate that interpreting diverges more strongly from the source texts than written translation does, a tendency observed consistently across both language pairs.

The main contributions of this study are as follows: (1) A comparative analysis of spoken and written translation renditions using identical source segments, allowing for a more controlled investigation of modality-based differences; (2) The application of information-theoretic measures to capture source-related divergences between interpreting and translation, extending previous work focused only on target-side comparisons; (3) The use of both cross-lingual and monolingual modelling approaches to assess translation difficulty, thereby offering complementary perspectives on cognitive and linguistic processing in human translation.

We believe that our approach can be useful for further applications. For instance, it can be used to measure other systemic divergences between language outputs, e.g. standard vs. simplified language, human vs. machine translations, etc.

6 Limitations

The KLD estimates used in this study are based on sentence-level probabilities aggregated from word-token outputs. Because these sentence probabilities can be extremely small, the calculations are susceptible to numerical underflow. Moreover, our sample—comprising 70 multiparallel sentence sets in each translation direction—may be too limited to approximate the true underlying probability distributions assumed by KLD with high reliability. As such, the present work should be regarded as proof of concept, and its findings interpreted withing this narrow scope.

7 Acknowledgments

At Saarland University and the University of Hildesheim, this research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – B7 "Translation as rational Communication" of SFB 1102 Information Density and Linguistic Encoding, Project-ID 232722074. Sharid Loáiciga was funded by the Swedish Research Council (VR) grant (2014-39) for the Centre for Linguistic Theory and Studies in Probability (CLASP). We thank Shalom Lappin for constructive discussions that helped improve this work. We also thank the anonymous reviewers for their insightful comments.

References

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Measuring translationese across levels of expertise: Are professionals more surprising than students? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 53–63, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever. 2021. Unsupervised neural machine translation with generative language models only. *Preprint*, arXiv:2110.05448.

Marta Kajzer-Wietrzny. 2012. *Interpreting universals and interpreting style*. Ph.D. thesis, Uniwersytet im. Adama Mickiewicza, Poznan, Poland. Unpublished PhD thesis.

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, Heike Przybyl, Ekaterina Lapshinova-Koltunski, and Elke Teich. 2023. Simultaneous interpreting as a noisy channel: How much information gets through. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 608–618, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ekaterina Lapshinova-Koltunski, Yuri Bizzoni, Heike Przybyl, and Elke Teich. 2021. Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 82–90, online. Association for Computational Linguistics.

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.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Christina Pollkläsener, Frances Yung, and Ekaterina Lapshinova-Koltunski. 2024. Capturing variation of discourse relations in english parallel data through automatic annotation and alignment. *Across Languages and Cultures*, 25(2):288 – 309.

Heike Przybyl, Alina Karakanta, Katrin Menzel, and Elke Teich. 2022a. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska, and Silvia Bernardini, editors, *Mediated discourse at the European Parliament*, number 19 in Translation and Multilingual Natural Language Processing, pages 191–218. Language Science Press, Berlin.

Heike Przybyl, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. 2022b. 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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California. Association for Computational Linguistics.

Stefan Schweter. 2020. German GPT-2 model. *Zenodo*.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisbon, Portugal. European Association for Machine Translation.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021. Bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 511–516, Online. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating explicitation of discourse connectives in translation using automatic annotations. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 21–30, Prague, Czechia. Association for Computational Linguistics.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022. Conditional bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2377–2389, Dublin, Ireland. Association for Computational Linguistics.