# Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication

**Ekaterina Lapshinova-Koltunski** **Yuri Bizzoni**
**Heike Przybyl** **Elke Teich**
Saarland University, University Campus A.2.2, DE-66123 Saarbrücken
e.lapshinova@mx.uni-saarland.de yuri.bizzoni@uni-saarland.de
heike.przybyl@uni-saarland.de e.teich@mx.uni-saarland.de

## Abstract

We report on a study of the specific linguistic properties of cross-linguistically mediated communication, comparing written and spoken translation (simultaneous interpreting) in the domain of European Parliament discourse. Specifically, we compare translations and interpreting with target language original texts/speeches in terms of (a) predefined features commonly used for translationese detection, and (b) features derived in a data-driven fashion from translation and interpreting corpora. For the latter, we use n-gram language models combined with relative entropy (Kullback-Leibler Divergence). We set up a number of classification tasks comparing translations with comparable texts originally written in the target language and interpreted speeches with target language comparable speeches to assess the contributions of predefined and data-driven features to the distinction between translation, interpreting and originals. Our analysis reveals that interpreting is more distinct from comparable originals than translation and that its most distinctive features signal an overemphasis of oral, online production more than showing traces of cross-linguistically mediated communication.

## 1 Introduction

Interpreting has recently received increased attention in various scientific disciplines, from automatic and human language processing to corpus-based and experimental translatology. A common interest in these diverse fields is to get a good descriptive basis of the specific linguistic characteristics of interpreting output. In translatology, analysing interpreting output is the most direct way of tapping into the translation process (e.g. Chmiel, 2018). In the study of human language processing, interpreting offers a highly interesting experimental ground for observing the interplay of prediction, retrieval and working memory (e.g. Christoffels et al., 2006). And in automatic language processing, simultaneous interpreting by machine remains a challenging task with many interesting open research questions (e.g. Müller et al., 2016; Grissom II et al., 2014).

Here, we come from the perspective of "translationese", i.e. the observation that translations exhibit specific linguistic features that distinguish them from original, non-translated language due to simplification, normalization, shining-through of the source text etc. While well documented for written translation, there is only little work on "interpretese" (see Section 2). Specifically, we pursue the following hypotheses: (H1) Interpreting is a highly special type of communication and is therefore well distinguished from the other language products. (H2) Interpreting and translation are well distinguished from comparable original speech and text, respectively; at the same time, interpreting is more distinct from comparable originals than translation. (H3) While there are overlaps in the features distinguishing interpreting and translations from their comparable originals (general translationese effects), we also expect differences between interpretese and translationese (effects of spoken vs. written mode). H3 is motivated by insights from previous work observing that interpreting overemphasizes features of spoken production (Shlesinger and Ordan, 2012), such that the spoken signal is stronger than the translation signal, more than translations overemphasize features typical of written production.

The remainder of the paper is organized as follows. Section 2 discusses related work. In Section 3 we introduce data and methods, includ-

ing the features used for classification. Section 4 presents our results. We conclude with a summary and outlook (Section 5).

## 2 Related work

It has been shown in a number of **studies of translationese** that translated texts have certain linguistic characteristics in common which differentiate them from original, non-translated texts (Gellerstam, 1986; Baker, 1993; Toury, 1995). The differences are reflected in the distribution of lexico-grammatical, morpho-syntactic and textual language patterns that can be organised in terms of more abstract categories such as *simplification* (Toury, 1995), *explicitation* (Olohan and Baker, 2000), *normalisation*, *shining-through* (Teich, 2003) and *convergence* (Laviosa, 2002). The differences are of a statistical character and can be uncovered automatically, as it has been shown in several works. They all use an extensive set of (often overlapping) features to differentiate between translated and non-translated texts (Baroni and Bernardini, 2006; Volansky et al., 2015; Rubino et al., 2016; Kunilovskaya and Lapshinova-Koltunski, 2020).

On the one hand, there is a demand for easily-extractable and scalable features that can be of use for NLP applications (Freitag et al., 2020; Graham et al., 2020; Artetxe et al., 2020; Zhang and Toral, 2019). On the other hand, there is a need for human-interpretable features that would help to understand the linguistic behaviour of translators. Most existing studies meet either the first or the second requirement. In their first computational work on translationese, Baroni and Bernardini (2006) included abstract surface features, such as word form, lemma, part-of-speech (PoS) n-grams. Volansky et al. (2015) used easily extractable shallow features, such as sentence length or type-token ratio, and grouped them according to the translationese phenomena mentioned above. Rubino et al. (2016) also used surface features derived from studies on machine translation quality and enhanced them with information theory-inspired features based on n-gram log-probabilities and perplexities of words, delexicalised parts-of-speech and flattened syntactic trees. Syntactic tree features were also used by Kunilovskaya and Lapshinova-Koltunski (2020) who designed linguistically motivated features that can be automatically extracted from texts annotated with the Universal Dependency framework. Although their feature set is immediately linguistically interpretable as opposed to easily-extractable shallow patterns, it requires a fair amount of time and effort to engineer them.

The **study of interpretese** is a more recent endeavour. There are corpus-based studies showing that interpreted texts possess a number of linguistic features that differentiate them from other language products, including written translation (Kajzer-Wietrzny, 2012; Defrancq et al., 2015; Bernardini et al., 2016; Ferraresi and Miličević, 2017; Dayter, 2018). Computational approaches to study interpretese (He et al., 2016; Bizzoni and Teich, 2019; Lapshinova-Koltunski, 2021) frequently use features inspired by automatic analysis of translationese. He et al. (2016) distinguish translationese and interpretese using shallow, surface features as well as more linguistically motivated ones based on strategies such as segmentation, passivisation, generalisation, summarisation. Bizzoni and Teich (2019) explore differences between translation and interpreting using bilingual word embedding spaces. Lapshinova-Koltunski (2021) follows Shlesinger and Ordan (2012)'s idea that the difference between spoken and written texts exerts a stronger effect than the difference between translated and non-translated ones. However, the author applies hand-crafted, theoretically driven features to classify English-German interpretations and translations, as well as comparable spoken and written non-translations in German.

In the present study, we analyse the differences between translation/interpreting in relation to comparable, original productions with a focus on interpreting (see H1 above). Relying on the existing works above, we assume that we can automatically tease apart interpreting, spoken originals, translation and written originals (see H2 above). At the same time, as both translations and interpretations are products of transfer from a source to a target language, we expect them to exhibit commonalities (see H3 above). Importantly, we compare the effects of the most commonly used pre-defined translationese features from the literature and a set of features derived from corpus data using an information-theoretic measure of distinctivity (see Section 3.2 below). Our main interest here is to find those features that distinguish best between interpreting and translation and that

are human-interpretable at the same time.

## 3 Data and Methods

### 3.1 Data

As dataset we use the English subsets of the EPIC-UdS (Przybyl et al., forthcoming) and Europarl-UdS (Karakanta et al., 2018) corpora, see Table 1 for details. EPIC-UdS contains transcripts of original spoken discourse delivered at the European Parliament (EP), as well as the simultaneous interpretation of these speeches into selected target languages. Europarl-UdS is the written equivalent of EPIC-UdS, containing the officially published original speeches and translations. Written originals are based on the EP speeches delivered, however are modified to fulfil written conventions before being published (cf. Bernardini et al., 2016). The spoken data include typical features of spoken languages such as false starts, hesitations and truncated words, and includes metadata such as the delivery type of original speeches (read, impromptu or mixed). For this study, we use English spoken (ORGsp) and written (ORGwr) originals, simultaneous interpretations (SI) and translations (TR) into English with German as source language. Due to availability of data for the spoken dataset, the written and spoken mode differ greatly in size. However, this does not seem to have an negative impact on our results (see 3.3. Moreover, most of our analyses focus on a distinction within the written and spoken mode.

| subcorpus | tokens | texts |
|---|---|---|
| ORGwr | 8,693,135 | 1,071 |
| TR | 6,260,869 | 886 |
| ORGsp | 68,548 | 137 |
| SI | 59,100 | 326 |

Table 1: Corpus overview: English target data from German sources. ORGwr=Originals written, TR=Translation, ORGsp=Originals spoken, SI=Simultaneous Interpreting.

### 3.2 Features

Analysis is driven by two sets of features: (A) predefined features that are commonly used for translationese detection (cf. Section 2 above), (B) features derived from translation and interpreting as well as comparable target language corpora in a data-driven way (see Section 3.3 below).

(A) features include

- `Word/PoS n-grams`: word and part-of-speech n-grams – uni-, bi- and trigrams
- `LexDens`: lexical density – average number of lexical words per clause
- `STTR`: lexical diversity measured with standardized type-token ratio(s)
- `Mfw`: most frequent words

(B) features include

- hesitations (*euh, hum*)
- discourse markers/particles (*so, well*)
- intensifiers (*very, particularly, really*)
- conjunctions (*and, but, however, whether*)
- personal pronouns (*you, we, I, she*)
- deictics (*that, this, here*)
- prepositions (*of, for, to, by, as*)
- function words vs. lexical words

### 3.3 Methods

**Features derived with Kullback-Leibler Divergence (KLD)** We compute unigram models of the four corpora and apply KLD to compare them in terms of relative entropy. Concretely, KLD measures the number of additional bits needed per item (e.g. word) for encoding items distributed according to A when a using an encoding optmized for B (equation 1).

$$D(A||B) = \sum_i p(item_i|A) log_2 \frac{p(item_i|A)}{p(item_i|B)} \quad (1)$$

For example, we may note that modeling translation (A) based on original written language (B) needs fewer extra bits than modeling interpreting (A) based on original spoken language (B), which would mean that interpreting is more distinct from comparable spoken originals than translation from comparable written originals. A crucial feature of KLD is its asymmetry – e.g., modeling spoken on the basis of written will yield different results than written modeled on the basis of spoken. Also, for each linguistic unit (e.g., word), we know its contribution to the overall KLD score (pointwise KLD)

so that we can detect the words (or other kinds of units) that contribute most to the overall distinction. In addition, we assess the impact of individual features on the overall divergence by a t-test.

For an example, see Figure 1. Features derived by KLD form the basis for our feature set (B) (all features with p<0.05).

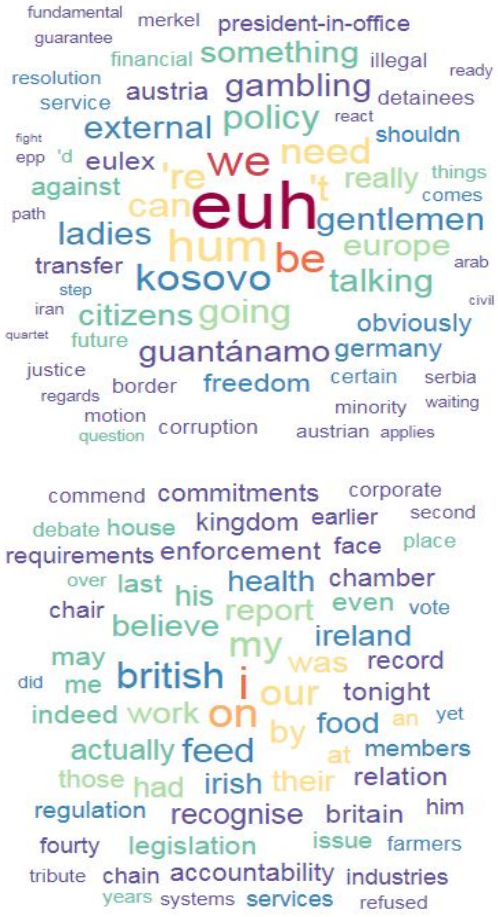


Figure 1: SI based on ORGsp (top), ORGsp based on SI (bottom) (source language: German). Item color denotes relative frequency (relF) (red=high relF, blue=low relF), item size denotes KLD score (large=high KLD, small=low KLD)

**Feature selection with Information Gain** As one of our aims includes comparison of interpretese and translationese features, we use several techniques to reduce the initial number of features to those relevant for a concrete prediction task. We use Information Gain (IG) along with frequency cuts to find an informative but also interpretable group of features. IG measures the expected reduction in entropy – uncertainty associated with a random feature (Roobaert et al., 2006, 464–465), or in other words, the feature's contribution to reduce the entropy. Given $S_X$ the set of training examples, $x_i$ the vector of $i^{th}$ variables in this set, $\mid S_{x_{i=v}} \mid / \mid S_X \mid$ the fraction of examples of the $i^{th}$ variable having value $v$, as shown in (2):

$$IG(S_{x,x_i}) = H(S_x) - \sum_{v=values(x_i)}^{\frac{S_{x_{i=v}}}{S_X}} H(S_{x_{i=v}}) \quad (2)$$

with entropy:

$$H(S) = -p_+(S)\log^2 p_+(S) - p_-(S)\log^2 p_-(S) \quad (3)$$

where $p \pm (S)$ is the probability of a training example in the set $S$ to be of the positive/negative class.

IG helps to select a feature set which is most suitable to distinguish interpreting from speech or translation from written text.

**Text classification** We perform text classification using Support Vector Machines (SVM, cf. Vapnik and Chervonenkis, 1974; Joachims, 1998) with a linear kernel. SVMs represent a learning algorithm that aims at classifying data points by maximizing the gap between classes in a hyperplane, making it particularly apt for feature-oriented machine learning approaches. For our study, we use SVM with a linear kernel, since we look for linearly classifiable features, and a 'one-vs-one' decision function.

We label our data with the information on classes represented in our case by mode (written, spoken) and translation type (translation, interpreting), collect the information on the feature frequencies from our corpus, and see if the corpus data support these classes.

We perform both a four-class classification task where each class is contrasted with all others, and two separate binary classification tasks to distinguish original and translated material within the same mode (interpreting vs. spoken originals, translation vs. written originals). The performance of the text classifiers are judged in terms of F1-measure. They are class-specific and indicate the results of automatic assignment of class labels to certain texts.

We also inspect the features that make the predefined classes distinct from one another. For this, the SVM weights (representing the hyperplane and corresponding to the support vectors) are judged – the magnitude of the weights provides

information on the importance of each feature: the higher the weight of a feature, the more distinctive it is for a particular class in the respective classification task.

For all our classification tasks, we used standard ten-fold cross-validation. Ten-fold cross-validation is a procedure used in classification processes to ensure that the classifier's results aren't due to a favorable or unfavorable distribution of the data in the test set – e.g., a test set containing only "easy" cases. It is performed by partitioning the dataset into ten equal parts and using each one in turn as a test set, with the remaining nine forming the training set. The final score is the average of the performance of the classifier on each test set. Another advantage of cross-validation is to partially counter the effect of class imbalance in our dataset, since all instances of every class will be used for validation once. Generally, anyway, we find that imbalance in our data is not a huge problem for this set of experiments. First, we mainly focus on binary classifications between balanced classes – original written texts vs. translations, or original speeches vs. interpreting transcripts. Second, our minority classes – originals speeches and interpreting transcripts – tend to return higher scores than the larger classes – original written texts and translations. Their performance is also consistent through cross-validation, as shown by the low standard deviations, confirming that they are not an artifact of small datasets.

## 4 Analysis and results

### 4.1 Feature Selection

We start our analyses with feature selection – the whole list of features is too long to be linguistically analysed for differences between interpreting and translation. Therefore, we test various settings with different groups of features. Table 2 presents their performance on the whole dataset in a multi-class classification.

We then select the three best performing groups of features (Word unigrams, Word+PoS n-grams and KLD) and perform filtering: with feature selection using IG – selecting top 400 and top 100 features within these three feature groups, and using a frequency cut (including only features of document frequency $>= 0.5$ – only features that occur in at least half of the documents). The filtering is an important step in our analysis, as we aim at an interpretable group of features that is also good in distinguishing interpreting and translation from the comparable originals in our data. Imposing a strong word-document frequency threshold helps filtering away content-specific lexical items, which reduces the risk that a topic imbalance in political speeches might help our classifiers.

| | F1 mean | F1 std |
|---|---|---|
| Word unigrams | .91 | .04 |
| Word+PoS n-grams | .89 | .04 |
| KLD (432) | .83 | .04 |
| STTR+LexDens | .36 | .07 |
| Mfw 100 | .77 | .9 |
| Word unigrams top 400 | .89 | .02 |
| Word+PoS n-grams top 400 | .83 | .04 |
| KLD top 400 | .82 | .04 |
| Word unigrams top 100 | .77 | .03 |
| Word+PoS n-grams top 100 | .76 | .07 |
| KLD top 100 | .68 | .09 |
| Word unigrams mf .5 | .71 | .05 |
| Word+PoS n-grams mf .5 | .83 | .06 |
| KLD mf .5 | .76 | .04 |

Table 2: Ten-fold cross-validation F1 mean and standard deviation for KLD-based features versus "classic" translationese features

The best performing feature sets, beyond the unfiltered sets, are those resulting from the IG top 400 selection. We also observe that the KLD features outperform word unigrams when we use a document frequency threshold of $>= 0.5$. This means that KLD brings up good classification measures if we want to reduce a feature set to a very short list. Our results show that if KLD gets scarce (in our case little more than ten words), then it works better than unfiltered word unigrams.

### 4.2 Hypothesis 1

We test the hypothesis that interpreting is clearly distinct from all other language products. For this, we perform a multi-class classification, where each subcorpus (SI, ORGsp, TR and ORGwr) is classified against the other subcorpora in our dataset. The results of automatic classification in Table 2 above shows that our classifiers achieve an overall good performance in recognising the classes in our data. Nonetheless, as appears in Figure 2, some settings can detect some classes better than others, which is not immediately evident in an overall multi-classification score.

To find out if interpreting is more distinct than the other subcorpora in our data, we inspect the resulting confusion matrix, visualised in Figure 2.

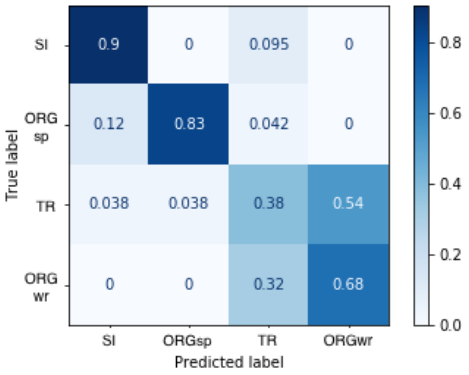


Figure 2: Confusion matrix using the 400 most informative words and PoS n-grams.

The accuracy numbers in the matrix confirm our assumption – interpreting is well distinguished from all other subcorpora in the data. It is never confused with either spoken or written originals and is rarely misclassified as translation. This is in line with the observations made in existing studies (Section 2) and confirms our hypothesis 1.

### 4.3 Hypothesis 2

To test if interpreting and translation are well distinguished from their comparable originals – speech for interpreting and text for translation – we perform two binary classification tasks. Tables 3 and 4 present an overview of the F1-measure values achieved with various groups of features.

| | F1 mean | F1 std |
|---|---|---|
| Word unigrams | .95 | .04 |
| Word+PoS n-grams | .94 | .05 |
| KLD | .91 | .06 |
| Word unigrams mf .5 | .7 | .05 |
| Word+PoS n-grams mf .5 | .8 | .06 |
| KLD mf .5 | .72 | .09 |
| Word unigrams top 400 | .91 | .04 |
| Word+PoS n-grams top 400 | .93 | .07 |
| KLD top 400 | .92 | .07 |

Table 3: Ten-fold cross-validation F1 mean and standard deviation in SI versus ORGsp.

As seen from the tables, both interpreting and translation can be automatically distinguished from the comparable originals with an F1-measure of up to 95%. The best results to identify interpreting are achieved with word unigrams, a combination of word and PoS ngrams, as well as the KLD features.

| | F1 mean | F1 std |
|---|---|---|
| Word unigrams | .91 | .02 |
| Word+PoS n-grams | .93 | .03 |
| KLD | .91 | .04 |
| Word unigrams mf .5 | .87 | .05 |
| Word+PoS n-grams mf .5 | .91 | .03 |
| KLD mf .5 | .9 | .01 |
| Word unigrams top 400 | .83 | .06 |
| Word+PoS n-grams top 400 | .86 | .07 |
| KLD top 400 | .84 | .07 |

Table 4: Ten-fold cross-validation F1 mean and standard deviation in TR versus ORGwr.

The F1-measure scores for these three groups of features are higher in Table 3 than in Table 4. This confirms our assumption that interpreting is more distinct from comparable speech than translation from comparable text.

### 4.4 Hypothesis 3

In the last step, we analyse the features that contribute to the distinction of interpreting against comparable speech (interpretese) and those that are distinctive for translation if classified against written texts (translationese). As this step is manual, we use the IG resulting selection of the three groups of features (Word unigrams top 400, Word+PoS n-grams top 400 and KLD top 400). We look into the overlap between the two lists of features (interpretese and translationese). Table 5 presents both absolute numbers and percent (calculated against the 400 items) of the overlaps. Interestingly, the KLD features have the biggest overlap (18.25%), whereas the word unigrams have 8.25% of overlapping features only.

| | **abs** | **in%** |
|---|---|---|
| Word unigrams top 400 | 33 | 8.25 |
| Word+PoS n-grams top 400 | 55 | 13.75 |
| KLD top 400 | 73 | 18.25 |

Table 5: The overlap of the the top 400 most relevant features per class vs. class - binary classification.

The overlapping KLD features are represented by various features that can be grouped according to the following categories: discourse markers (*again, already, because, just, obviously, particularly, therefore,* etc), specific verb types – verbs of activity (*come, get, react*), communica-

tion (*tell, talk*), mental processes (*think, remind* and existence (*represent*) – demonstrative pronouns (*this, that*), addressee reference (*ladies, gentlemen*), speaker reference (*we*) and various lexical items.

The overlapping word unigrams contain some of the features that occur in the KLD list too. However, the majority of the items in the KLD and word unigrams lists differ. For instance, the unigram list also contains the discourse marker *because*, but there are also *if* and *or* which were not contained in the overlapping KLD list. There are no demonstrative pronouns, but the personal pronoun *them*. Moreover, the *wh*-words *what* and *who* appear in the unigram list but there are fewer verbs (*be, think*). It also contains the addressee reference (*gentlemen, ladies*), but no speaker reference (*we*), like in the KLD list.

The overlapping word and PoS n-gram list lies in between – it contains fewer features than the overlapping KLD list, but more features than the word unigram list. The features contain n-grams with discourse markers (*conj, conj adp, conj adv, conj noun, noun conj det, noun conj*), addressee reference (*ladies and gentlemen, president ladies and, and gentlemen*), speaker reference (*we, our*), prepositional phrases (*adp adj, adp det*) and n-grams with pronouns (*det pron, noun pron noun, pron verb det, verb pron*) and various nominal and verbal phrases.

Comparing the overlapping lists, we observe that the weights of the same features are always higher in the interpretese lists than in the translationese lists. Besides that, there are some differences in the contextual use of the same features in interpretations and translations. Examples (1) and (2) illustrate such differences.

(1) a. SI: *and euh* ***obviously*** *fair trade is the foundation of Europe's prosperity.*
b. TR: *The material was* ***obviously*** *useful for both the preparation of the 2001 budget and for the 1999 discharge.*

In (1-a), the adverb *obviously* occurs at the utterance start, whereas the same adverb directly precedes the predicate *useful* in (1-b). The function of the adverb differs as well: In interpreting, *obviously* serves as a discourse marker, whereas in translation, it is a predicate modifying adverb.

(2) a. SI: *let me very briefly* ***remind*** *you about the short time span within which we reacted when banks in Europe were in trouble .*
b. TR: *I would* ***remind*** *people in this Parliament that it was not so long ago that this Parliament passed .*

In example (2), the verb *remind* is used in both interpreting and translation with the same purpose – to address the audience. However, we see in the corpus examples that the addressee reference differs – the second person pronoun *you* is used in interpreting (2-a) and a full nominal phrase (*people in this Parliament*) is used in translation (2-b). Further corpus analysis of our data reveals that the verb *remind* is followed by the pronoun *you* in 36.81% of all the cases in translation. By contrast, *you* follows this verb in 63.64% of the cases in our interpreting data.

The observed differences between the interpretese and the translationese features confirm our hypothesis (H3). They also go in line with the observations from previous work that interpreting emphasizes features of spoken production, still being distinct from the spoken originals. This latter distinction may have roots in the nature of the data, as some of original speeches are prepared and read out (see Section 3.1), whereas interpreting can be seen as spontaneous production.

## 5 Summary and conclusions

We have reported on a study of the specific linguistic properties of cross-linguistically mediated communication, comparing written translation and simultaneous interpreting in the domain of European Parliament discourse. To do so, we combined an exploratory, data-driven approach (KLD on unigram models) for detecting distinctive features with a supervised approach (SVM classification). Our initial hypotheses (H1 and H2, Section 1) that translation and interpreting are both clearly distinguished from comparable originals, but interpreting is more distinct than translation have been confirmed. We then inspected the features contributing to the distinctions and found that there is an overlap between the distinctive features of interpreting and translation, signalling the fact that both are instances of translated language, but there are also some unique features (cf. H3, Section 1). The unique features for interpreting are clearly signals of spoken, online

production, which confirms insights from previous work. Among the kinds of features considered, the features obtained by KLD typically come out with higher scores for interpreting than translation confirming that interpreting is the most distinctive kind of production (cf. H1, Section 1). Also, since another goal was to work with few but powerful features, KLD clearly supported this goal, e.g. compared to simply using n-gram frequency, we get fewer and better features.

In our ongoing work, we analyse in more depth the detected features by inspecting their linguistic properties and lexico-grammatical contexts. For instance, some of the interpretese effects will be related to the specific processing constraints of interpreting which have an impact on retrieval, working memory as well as prediction. To this end, we relate the features found to be typical of interpreting to indices of processing load, such as surprisal (Teich et al., 2020) or dependency length (Przybyl and Teich, forthcoming).

## 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.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In G. Francis Baker M. and E. Tognini-Bonelli, editors, *Text and Technology: in Honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.

Marco Baroni and Silvia Bernardini. 2006. 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.

Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28:61–86.

Yuri Bizzoni and Elke Teich. 2019. Analyzing variation in translation through neural semantic spaces. In *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC) at RANLP-2019*, Varna, Bulgaria. ACL.

Agnieszka Chmiel. 2018. In search of the working memory advantage in conference interpreting – training, experience and task effects. *International Journal of Bilingualism*, 22(3):371–384.

Ingrid K Christoffels, Annette MB 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.

Daria Dayter. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM*. To appear.

Bart Defrancq, Koen Plevoets, and Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo, editor, *Yearbook of Corpus Linguistics and Pragmatics*, pages 195–222. Springer International Publishing, New York.

Adriano Ferraresi and Maja Miličević. 2017. Phraseological patterns in interpreting and translation. Similar or different? In G. De Sutter, M.-A. Lefer, and I. Delaere, editors, *Empirical Translation Studies. New Methodological and Theoretical Traditions*, volume 300 of *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 157–182. Mouton de Gruyter.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

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.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, and 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 Association for Computational Linguistics: Human Language Technologies*, pages 971–976. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142, London, UK. Springer.

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 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, Marseille, France. European Language Resources Association.

Ekaterina Lapshinova-Koltunski. 2021. Analysing the dimension of mode in translation. In Mario Bisiada, editor, *Empirical Studies in Translation and Discourse*, Translation and Multilingual Natural Language Processing, pages 223–243. Language Science Press, Berlin.

Sara Laviosa. 2002. *Corpus-based Translation Studies, Theory, Findings, Application*. Rodopi, Amsterdam.

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Maeve Olohan and Mona Baker. 2000. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1:141–158.

Heike Przybyl, Alina Karakanta, Katrin Menzel, and Elke Teich. forthcoming. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Marta Kajzer-Wietrzny, Silvia Bernardinia, Adriano Ferraresi, and Ilmari Ivaska, editors, *Empirical investigations into the forms of mediated discourse at the European Parliament*, Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Heike Przybyl and Elke Teich. forthcoming. Dependency length minimization in simultaneous interpreting. In *Proceeding of the 3rd International Conference on Translation, Interpreting and Cognition*, Forli.

Danny Roobaert, Grigoris Karakoulas, and Nitesh V. Chawla. 2006. Information gain, correlation and support vector machines. In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh, editors, *Feature Extraction: Foundations and Applications*, pages 463–470. Springer Berlin Heidelberg, Berlin, Heidelberg.

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.

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

Elke Teich. 2003. *Cross-Linguistic Variation in System and 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*. John Benjamins, Amsterdam.

Vladimir Naumovich Vapnik and Alexey Yakovlevich Chervonenkis. 1974. *Theory of Pattern Recognition*. Nauka, Moscow.

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

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.