How Human is Machine Translationese?
Comparing Human and Machine Translations of Text and Speech

Yuri Bizzoni[1], **Tom S Juzek**[1], **Cristina España-Bonet**[2], **Koel Dutta Chowdhury**[1],
Josef van Genabith[1,2], and **Elke Teich**[1]

[1] Department of Language, Science and Technology, Saarland University, Germany
[2] DFKI GmbH, Saarbrücken, Germany
{yuri.bizzoni,tom.juzek,koel.duttachowdhury}@uni-saarland.de
{cristinae,Josef.Van_Genabith}@dfki.de,{e.teich}@mx.uni-saarland.de

Abstract

Translationese is a phenomenon present in human translations, simultaneous interpreting, and even machine translations. Some translationese features tend to appear in simultaneous interpreting with higher frequency than in human text translation, but the reasons for this are unclear. This study analyzes translationese patterns in translation, interpreting, and machine translation outputs in order to explore possible reasons. In our analysis we (*i*) detail two non-invasive ways of detecting translationese and (*ii*) compare translationese across human and machine translations from text and speech. We find that machine translation shows traces of translationese, but does not reproduce the patterns found in human translation, offering support to the hypothesis that such patterns are due to the model (human vs. machine) rather than to the data (written vs. spoken).

1 Introduction

In recent years, a growing body of work has pointed to the presence, across different genres and domains, of a set of relevant linguistic patterns with respect to syntax, semantics and discourse of human translations. Such patterns make translations more similar to each other than to texts in the same genre and style originally authored in the target language. These patterns are together called "translationese" (Teich, 2003; Volansky et al., 2015).

Linguists classify translationese in two main categories: (i) source's interference, or *shining-though* as put forward by Teich (2003). For example, a translation replicating a syntactic pattern which is typical of the source language, and rare in the target language, displays a typical form of shining-through; (ii) aherence or over-adherence to the target language's standards, that is *normalisation*. For example, translating a sentence displaying marked order in the source with a sentence displaying standard order in the target it a typical example of over-normalization. Nonetheless, translationese's main causes remain unclear (Koppel and Ordan, 2011; Schmied and Schäffler, 1996).

Translationese displays different patterns depending on the translation's mode and register: a typical example is simultaneous interpreting, which shows translationese patterns distinct from those observed in written translations (Bernardini et al., 2016). We can interpret differences as either (*i*) an effect of the limitations of the human language apparatus that constrain translations, (*ii*) an inevitable effect of the structural and semantic differences between languages; or (*iii*) a combination of the two. To test these hypotheses, it is common to compare human translations (HT) produced under different circumstances, e.g. written translation versus simultaneous interpreting, following the assumption that, if translationese is a product of human cognitive limitations, translations produced under higher cognitive constrains should present more evident translationese symptoms. Machine translation (MT) does not have any human cognitive limitation, but current state-of-the-art systems learn to translate using human data, which is affected by translationese. On the other hand, unlike human translators, most standard modern MT systems still work at the level of individual sentences (rather than document or dialog) unaware of the surrounding co-text and context. Taking these observations into account, this paper explores the following research questions:

Q1*. *Does machine translation replicate the translationese differences observed between text and speech in human translation?

Q2*. *Do different machine translation architectures learn differently enough to display distinctive translationese patterns?

We study translationese in speech and writ-

Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT), pages 280–290
July 9-10, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

ten text by comparing human and machine translations to original documents. Specifically, we present a comparison between three MT architectures and the human translation of spoken and written language in German and English, exploring the question of whether MT replicates the differences observed in human data between text-based and interpreting-based translationese.

While some aspects of translationese tend to appear in simultaneous translation of speech with higher frequency than in translation of text, it is unclear whether such patterns are data- or model-dependent. Since machine translation does not have the cognitive limitations humans have (in terms of memory, online processing, etc.), MT systems should fail to replicate the differences between text and speech translationese observed in human translations, if such differences are due to the limits of human cognitive capacities.

Assuming that MT engines are trained only on translated texts and *not* on simultaneous interpreting, so that they cannot simply learn to mimic interpreting's translationese signal, the differences between text and speech translationese observed in human translation vs. interpreting are not expected to the same extent in MT if they are an effect of human cognitive limitations. If, on the other hand, translationese patterns in text and speech are due to characteristics inherent in the two modes of expression, MT systems' translationese patterns should mimic human translationese patterns.

The paper is organized as follows. Section 2 presents related work. Section 3 introduces the translationese measures and Section 4 the data and translation systems used in our study. Section 5 presents our results, comparing human with machine translations and comparing MT architectures among themselves. Section 6 concludes the paper.

2 Related Work

Translationese seems to affect the semantic as well as the structural level of text, but much of its effects can be seen in syntax and grammar (Santos, 1995; Puurtinen, 2003). An interesting aspect of translationese is that, while it is somewhat difficult to detect for the human eye (Tirkkonen-Condit, 2002), it can be machine learned with high accuracy (Baroni and Bernardini, 2006; Rubino et al., 2016). Many ways to automatically detect translationese have been devised, both with respect to textual translations and simultaneous interpreting (Baroni and Bernardini, 2006; Ilisei et al., 2010; Popescu, 2011). Simultaneous interpreting has shown specific forms of translationese distinct from those of textual translation (He et al., 2016; Shlesinger, 1995), with a tendency of going in the direction of simplification and explicitation (Gumul, 2006). Due to the particularly harsh constraints imposed on human interpreters, such particularities can be useful to better understand the nature and causes of translationese in general (Shlesinger and Ordan, 2012).

The features of machine translated text depend on the nature of both training and test data; and possibly also on the approach to machine translation, i.e. statistical, neural or rule-based. The best translation quality is achieved when the training parallel corpus is in the same direction as the test translation, i.e. original-to-translationese when one wants to translate originals (Lembersky et al., 2013). In this case, more human translationese features are expected, as systems tend to learn to reproduce human features. The effects of translationese in machine translation test sets is also studied in Zhang and Toral (2019a). In fact, texts displaying translationese features seems to be much easier to translate than originals, and recent studies advise to use only translations from original texts in order to (automatically) evaluate translation quality (Graham et al., 2019; Zhang and Toral, 2019b). By contrast, Freitag et al. (2019) show a slight preference of human evaluators to outputs closer to originals; in this case the translation is done from translationese input, because, as noted by Riley et al. (2019), the best of the two worlds is not possible: one cannot create original-to-original corpora to train bias-free systems. Aranberri (2020) analyzed translationese characteristics on translations obtained by five state-of-the-art Spanish-to-Basque translation systems, neural and rule-based. The author quantifies translationese by measuring lexical variety, lexical density, length ratio and perplexity with part of speech (PoS) language models and finds no clear correlation with automatic translation quality across different test sets. The results are not conclusive but translation quality seems not to correlate with translationese. Similar results are obtained by Kunilovskaya and Lapshinova-Koltunski (2019) when using 45 morpho-syntactic features to analyze English-to-Russian translations. Vanmassenhove et al. (2019) noted that statistical systems reproduce human lexical diversity better than neural systems, for English–Spanish and English–French even if transformer models, i.e. neural, are those with the highest BLEU score. However, their transformer model did not use subword segmentation putting a limit on the plausible lexical diversity it can achieve.

In our study, we focus on English (en) and German (de) texts and compare the presence of translationese in human translations and interpreting, and in machine translations obtained by in-house MT engines. Differently to Aranberri (2020), we develop the MT engines to have control on the training data, since the aim of this work is not to study how translationese features correlate with translation quality, but the presence of translationese features themselves. To measure them, we use two metrics: part-of-speech (PoS) perplexity and dependency parsing distances. We use them as complementary measures, perplexity to model PoS sequences of linguistics objects in sentences (linear dimension), and dependency distance to model their syntactic depth (hierarchical dimension).

3 Translationese Measures

3.1 Part-of-Speech Perplexity

Our first measure to detect translationese is based on the assumption that a language model is able to capture the characteristic PoS sequences of a given language . Since grammatical structures between languages differ, a language model trained on Universal Part of Speech sequences of English will on average display less perplexity if exposed to an English text than if exposed to a German text. Following the same logic, if two models, one trained on English and one on German, were to progress through the PoS sequences of an English translation showing strong German interference, we could expect the English model's perplexity scores to rise, while the German model's perplexity would stay relatively low (Toral, 2019). On the other hand, if the English translation were displaying normalisation, we would expect the English model to display a lower perplexity than the German one. Perplexity is defined as the exponentiation of the entropy $H(p)$:

$$2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)} \tag{1}$$

where $p(x)$ is the probability of a token x (possibly given its context), and $-\log_2 p(x)$ is the surprisal of x. While surprisal measures in bits the uncertainty in a random variable taking a certain value x, entropy measures the weighted average surprisal of the variable.

3.2 Universal Dependencies

A syntactic analysis examines how the elements of a linguistic sequence relate to each other; for our purposes, those elements are words of a sentence. We employ the framework of Universal Dependencies (UD), which expresses syntactic relations through dependencies: each element depends on another element, its head (Nivre et al., 2019). In UD, in contrast to most other dependency frameworks, the head is the semantically more salient element and the dependent modifies the head. The top-level head is the root of a sequence, which is typically the main verb of the matrix clause. For instance, in the sentence *The great sailor is waiving*, *the* and *great* modify and depend on *sailor*, while *sailor* and *is* modify and depend on *waiving*, which is the root of the sentence. Figure 1 illustrates a dependency analysis of this example. The UD framework aims to be universal, i.e. suitable for all of the world's languages, and there are a large number of resources and tools available.[1]

An analysis of dependency lengths could help to identify translation artifacts. If a source language is shining-through, the translation's dependency lengths will be closer to the source language's average; and vice versa for normalisation. Explicitation will lead to longer, and simplification to shorter distances compared to originals.

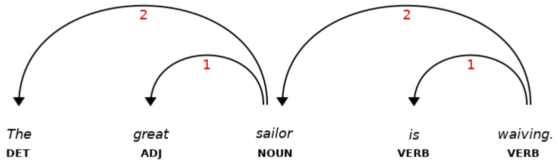


Figure 1: Visualizing a simple sequence in the Universal Dependencies framework, incl. dependency distances (red numerals below a dependency arch/edge).

We use spaCy's parser[2] to parse our corpora. An evaluation of 2000 head tags taken from randomly sampled sentences gives an accuracy rate of 91.2% and 93.6% for the German spoken and written originals, respectively, and 95.0% and 95.8% for English, as evaluated by a senior linguist. Sentences shorter than 3 tokens were excluded, as such sequences typically lack a verb and thus a root.

[1] See https://universaldependencies.org/ for details.

[2] https://spacy.io

We then use those parses for a macro-analysis of the syntactic structures, viz. an analysis of *average summed distances*. This analysis measures for each word the distance in words to its head. In Figure 1, the distance from *the* to its head is 2, from *great* to its head, it is 1. The sum of distances for the example is 6. In the following, we average the summed distances per sentence length. The measure can be interpreted as a proxy of (cumulative) processing difficulty/complexity of a sequence, where long distance structures are regarded as more complex than structures with shorter distances. Particle verbs illustrate this. In *Mary picked the guy that we met the other day up*, the particle *up* has a long distance to its head and the sequence is relatively hard to process. This contrasts to *Mary picked up the guy that we met the other day*, where the distance between particle and verb is short, which reduces the cognitive load. This dependency-based measure is taken from Gibson et al. (2019) and Futrell et al. (2015) and builds on work by Liu (2008).

4 Experimental Settings

4.1 Corpora

Originals and Human Translations. We used human written texts and speech transcriptions to train and test our language models and to extract part of our syntactic distance measures. We use datasets that belong to the same genre and register but to different modalities: transcriptions of European Parliament speeches by native speakers and their interpreted renditions (EPIC-UdS, spoken), the written speeches and their official translations (Europarl-UdS, written) (Karakanta et al., 2018). Table 1 summarizes the number of sentences and words for each of the six categories:

1. Original written English
2. Original written German
3. Original spoken English (transcript)
4. Original spoken German (transcript)
5. English to German translations
6. English to German interpreting (transcript)

For each corpus, we train the PoS models on 3000 random sentences and evaluate on the remaining data. We tokenized our data using NLTK (Bird and Loper, 2004) and performed universal PoS tagging via spaCy. We train our language models using a one-layer LSTM with 50 units (Chollet et al., 2015). Due to the small dimensions of the vocabulary (17 PoS), 5 iterations over 3000 sentences suffice to converge. In our experiments, we measure the average perplexity of each model on *unseen* human data from of each category, and on the translations produced by three MT architectures in two different settings (see Section 4.2).

MT Training Data. In order to adapt our machine translation engines to the previous modalities as much as possible, we gather two different corpora from OPUS (Tiedemann, 2012), one of them text-oriented (C_t) and the other speech-oriented (C_s). The distribution of their sub-corpora is shown in Table 2. Note that we do not include Europarl data here so that there is no overlap between MT training and our analysis data.

Note also that our speech data (TED talks and subtitles) is still made up from translations and not simultaneous interpreting. This is important since it prevents MT systems from simply mimicking interpreting's pronounced translationese.

All datasets are normalised, tokenized and truecased using standard Moses scripts (Koehn et al., 2007) and cleaned for low quality pairs. Duplicates are removed and sentences shorter than 4 tokens or with a length ratio greater than 9 are discarded. We also eliminate sentence pairs which are not identified as English/German by *langdetect*[3] and apply basic cleaning procedures. With this, we reduce the corpus size by more than half of the sentences. In order to build balanced corpora we limit the number of sentences we used from ParaCrawl to 5 million and from Open Subtitles to 10 million. With this, both C_t and C_s contain around 200 million tokens per language. Finally, a byte-pair-encoding (BPE) (Sennrich et al., 2016) with 32 k merge operations trained jointly on en–de data is applied before training neural systems. After shuffling, 1,000 sentences are set aside for tuning/validation.

4.2 Machine Translation Engines

We train three different architectures, one statistical and two neural, on the corpora above.

Phrase-Based Statistical Machine Translation (SMT). SMT systems are trained using standard freely available software. We estimate a 5-gram language model using interpolated Kneser–Ney discounting with `SRILM` (Stolcke, 2002). Word alignment is done with `GIZA++` (Och and Ney, 2003)

[3] https://pypi.org/project/langdetect/

Europarl-UdS			EPIC-UdS		
	lines	tokens		lines	tokens
German Translation	137,813	3,100,647	German Interpreting	4,080	58,371
Written German	427,779	7,869,289	Spoken German	3,408	57,227
Written English	372,547	8,693,135	Spoken English	3,623	68,712

Table 1: Corpus collections used to train our language models: Europarl-UdS (written) and EPIC-UdS (spoken). German translation and interpreting are both from English.

	lines	*de* tokens	*en* tokens	C_t	C_s
CommonCrawl	2,212,292	49,870,179	54,140,396	✓	✓
MultiUN	108,387	4,494,608	4,924,596	✓	✓
NewsCommentary	324,388	8,316,081	46,222,416	✓	✓
academia career limiting moves Rapid	1,039,918	24,563,476	148,360,866	✓	✓
ParaCrawl-5M	5,000,000	96,262,081	103,287,049	✓	✗
TED	198,583	3,833,653	20,141,669	✗	✓
OpenSubtitles-10M	10,000,000	85,773,795	93,287,837	✗	✓
Total clean Speech	13,379,441	187,551,444	197,175,542	✗	✓
Total clean Text	9,121,710	198,340,602	207,434,038	✓	✗

Table 2: Text-oriented (C_t) and speech-oriented (C_s) corpora used for training the MT systems.

and both phrase extraction and decoding are done with the `Moses` package (Koehn et al., 2007). The optimization of the feature weights of the model is done with Minimum Error Rate Training (MERT) (Och, 2003) against the BLEU (Papineni et al., 2002) evaluation metric. As features, we use the language model, direct and inverse phrase probabilities, direct and inverse lexical probabilities, phrase and word penalties, and lexicalized reordering.

The neural systems are trained using the `Marian` toolkit (Junczys-Dowmunt et al., 2018) in a bidirectional setting $\{en,de\}\leftrightarrow\{de,en\}$:

RNN-Based Neural Machine Translation (RNN). The architecture consists of a 1-layer bidirectional encoder with complex GRU units (4 layers) and a decoder also with complex GRUs (8 layers). The tied embeddings have a dimension of 512 and hidden states with a size of 1024, using the Adam optimizer (Kingma and Ba, 2015) with β_1=0.9, β_2=0.98 and ϵ=1e-09 and a growing learning rate from 0 to 0.0003. Label (0.1) and exponential smoothing, dropout of 0.1 and layer normalisation are also applied.

Transformer Base Neural Machine Translation (TRF). We use a base transformer architecture as defined in Vaswani et al. (2017), that is, a 6-layer encoder–decoder with 8-head self-attention, a 2048-dim hidden feed-forward, and 512-dim word vectors. Optimization algorithm, dropout, smoothings and learning rate (with warmup till update 16,000) are the same as for RNN.

5 Experimental Results

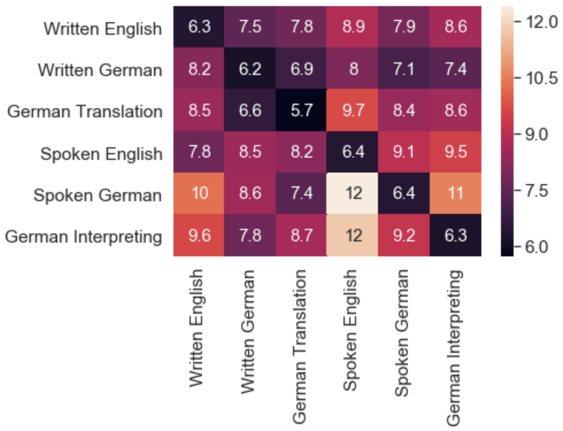
Below we present the translationese characteristics in the different modalities found in our study.

5.1 Human Translationese

5.1.1 Perplexity

The PoS perplexity scores of our language models on human data shows that, as expected, each model's perplexity is at its lowest when confronted with data from the same category. Since the amount of data differs among modalities, we checked that 10 independent partitions of the data lead to the same results. Translation and interpreting models are least perplexed by unseen instances of their own categories, and are not confused by original written or spoken data: translation and interpreting display indeed detectable and idiosyncratic patterns.

Figure 2 shows perplexity scores in a matrix where the x-axis reflects the data on which the language models have been trained, and the y-axis reflects the partition of data on which the PoS mod-



	Written English	Written German	German Translation	Spoken English	Spoken German	German Interpreting
Written English	6.3	7.5	7.8	8.9	7.9	8.6
Written German	8.2	6.2	6.9	8	7.1	7.4
German Translation	8.5	6.6	5.7	9.7	8.4	8.6
Spoken English	7.8	8.5	8.2	6.4	9.1	9.5
Spoken German	10	8.6	7.4	12	6.4	11
German Interpreting	9.6	7.8	8.7	12	9.2	6.3

Figure 2: Perplexity of each universal PoS language model (x-axis) on unseen data from each category (y-axis).

els are tested. The diagonal corresponds to training and testing in the same modality and language and, consequently, shows the lowest perplexities as said before. Leaving the diagonal of the matrix aside, we see that German translations are least perplexing for the model trained on written German, and vice-versa, written German sequences are least perplexing for the model trained on German translation. The German translation model displays its highest perplexity on English, and the written English model is more perplexed by German translations than by German originals. These observations seem to point away from shining-through, and to point instead towards the presence of normalization in German translation and interpreting.

Interpreting differs from translation. While German translation sequences are of low-perplexity for the written German model, German interpreting sequences are quite perplexing for the spoken German model, and in general present higher levels of perplexity than translation for all German models. Unlike German translation, the interpreting model returns high perplexity on all datasets except its own. This particularity of interpreting data was previously noted (Bizzoni et al., 2019) and ascribed to structural over-simplification, as a possible effect of the cognitive pressure on interpreters.

5.1.2 Syntax: Dependency Length

The corresponding analysis for our syntax measure is presented in Figure 3, the top left and top center plots. For written spoken data in both German and English, the summed dependency distances increase as sentence lengths increase, as one would expect. Translations from English into German are slightly less complex than German originals. The same applies to English–German interpreting. This is in contrast with translations from German into English, that are somewhat more complex than the English originals. For German–English interpreting, there is no difference to the originals. Arguably, the fact that English-to-German translations are less complex than German originals and that German-to-English translations are more complex than English originals is an artifact of source language shining-through. Notice that discrepancies between curves are more evident for long sentences, and that sentences in spoken texts are systematically shorter than in written text.

5.2 Human vs. Machine Translationese

Effects of translationese are expected in MT outputs as long as the systems are trained with human data which is rich in translationese artifacts. In this section, we compare the translationese effects present in human translations with those present in TRF translations, the state-of-the-art MT architecture. For comparison, our bidirectional TRF trained with text-oriented data achieves a BLEU score of 35.5 into English and 38.1 into German on Newstest 2018, and 33.0 and 36.2 respectively when trained with speech-oriented data. MT systems have been trained using the corpora described in Table 2. Data has not been classified according to translationese, but one can assume[4] that most of the corpora will be original English. This would create a bias towards human-like translationese in German translations.

The single translationese feature that seems to be most characteristic of the output of our MT model is structural shining-through, a characteristic present in human translations, but that appears to become particularly relevant in various MT outputs, specially the ones translating German written modality into English. Figure 4 shows low perplexities when the written German language model is used on the translated text into English (5.9 vs. 7.3 for the English model on the English translation). According to the MT training data distribution, this feature cannot be due to data biases but to the MT architecture itself. At the same time,

[4] Riley et al. (2019) quantified the amount of original German texts in the WMT18 data with a 22%. Their corpora only differ from ours by lacking Europarl and by including in addition TED talks and OpenSubtitles. The authors report an F_1=0.85 for their classification.

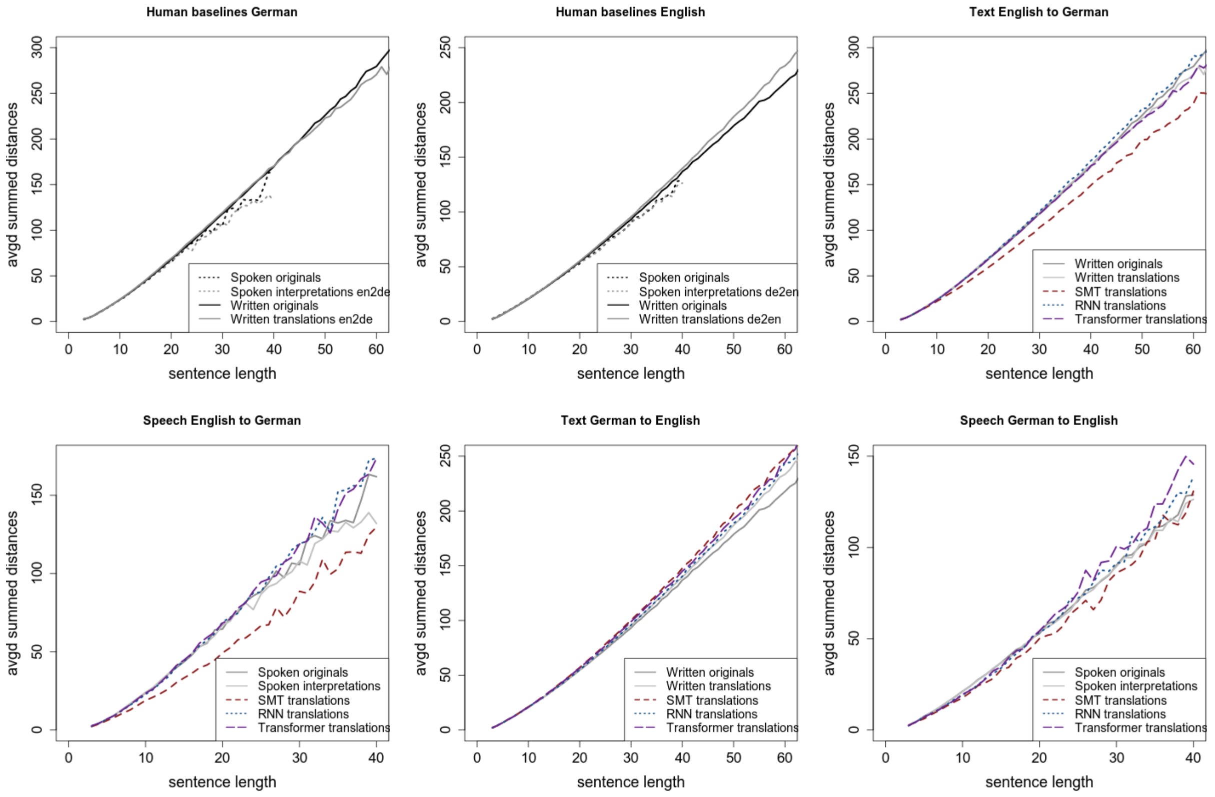


Figure 3: Averaged summed dependency distances, y-axis, per sentence length, x-axis, for German HTs (top left) and for English HTs (top center). The same incl. HTs and MTs for translations of written English to German (top right), spoken English to German (bottom left), written German to English (bottom center), and for spoken German to English (bottom right).

target language normalisation, which is a prominent translationese feature in human translations, appears less evident in our MT output. In this case, normalisation is only slightly more prominent in translations into German with a perplexity of 6.9 in translated text and 7.0 in translated speech, where translations into English score 7.3 and 6.9 respectively. We also checked if the content of the MT training data is relevant for the conclusions and might bias the comparison with interpretings and translations for instance. To this end, we use the text-oriented MT engine to translate speech, and the speech-oriented MT engine to translate text. In our experiments, we reach equivalent perplexities for written text data whichever engine we use but, for speech, perplexities are in general higher when the text-oriented MT is used. This could be expected, but we attribute the noticeable effect only on spoken data due to its completely different nature.[5]

Possibly due to the larger presence of shining-through, machine translations from English to German appear to behave quite differently from machine translations from German to English. While differences due to the source and target language naturally exist in human translations, the MT output appears more sensitive to such variations. Summed parse tree distance Figure 3 show how TRF outputs are more complex than English originals and translations/interpretings but have a similar degree in the German counterparts. We found that machine translation seems to *over-correct* translationese effects, again not following the characteristics of training data.

5.3 Translationese across MT Architectures

The previous section summarizes the differences between a state-of-the-art MT system and human translations/interpretings, but one could expect different behaviors for other architectures. In the following, we present a detailed analysis of each ar-

[5] Whereas clean transcriptions as TED talks are used to train our speech-oriented engines, Europarl data include many speech-distinctive elements such as hesitations, broken sentences and repetitions: *but / is / euh what/ what's most difficulty bringing in prisoners from Guantánamo because we see them to be a security euh risk.*

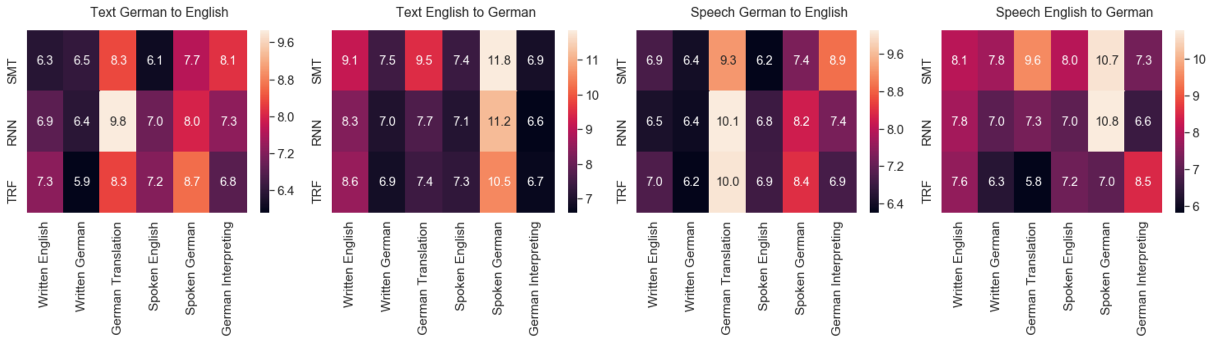


Figure 4: Perplexity per language and mode. Language: German to English and English to German. Mode: written and spoken.

chitecture for our two measures independently.

5.3.1 Perplexity

The results for PoS perplexities are illustrated in Figure 4.

Results on SMT. German language models return on average higher perplexities on SMT than English models, and translations into English are on average less perplexing than translations into German, which hints at English to German shining-through. The observed absence of shining-through in the human data confirms this hypothesis, since our German translation and interpreting models are highly perplexed by SMT data, indicating that this kind of translation presents structural patterns that differ from human translationese. It also differs from neural systems in the sense that SMT seems to better reproduce the structure of the MT training corpus. SMT is acknowledged to translate more literally, while NMT tends to be more inspirational, and this trend is observed in our setting too.

Results on RNN. English models show higher perplexities when tested on any translated German than when tested on translated English, and German models show more perplexity for German translations than for English ones. German translation and interpreting models reverse the pattern, showing lower perplexity on German translations than on English translations. This seems to point to a more complex phenomenon, possibly a mix of shining-through and normalization, but it seems to be less perplexing for the models trained on translation and interpreting. RNN's translation patterns are clearly closer to human translationese than SMTs.

Results on TRF. While RNN data attains the lowest *average* perplexity, our Transformer manages to attain the lowest single scores —reaching in some cases lower perplexities than the ones elicited by human equivalents. Consistently, the highest TRF perplexities are lower than the highest perplexities of the other systems. The single model displaying the lowest perplexity on this data is German Translation, followed by written German, while spoken German shows the highest perplexity, followed by German Interpreting. Written German shows low-level perplexities on German translations, in accordance with the German translation model; but also lower perplexities on English translations, which instead spark significant perplexities in the German translation model. This behavior seems to point to a "stronger than human" shining-through of German into English, which makes English translations less surprising for German than for English but perplexes a model trained on the patterns of human translations. Models trained on spoken data also appear more perplexed than models trained on written data, hinting at a presence of "written patterns" in speech translation that does not appear in human data.

5.3.2 Syntax: Dependency Length

The results for the syntactic macro-analysis of dependency lengths are illustrated in Figure 3 (top right and all bottom plots).

Results on SMT. Similarly to the trends observed with perplexities, the SMT translations diverge the most of all models. Translations into German of English speech and text lack the complexity of the German originals and are even considerably less complex than human interpreting/translation. Translations into English of German speech are also less complex than the HTs, in contrast to other models. The model is only in line with human translations and other models when it comes to translations of German text.

Results on RNN. Overall, RNN translations come out really well for this measure, meaning that they are able to mimic human outputs. Translations of English texts into German are in line with German originals, translation of English speech is slightly more complex than the HTs. Translations of German speech and texts into English are very close to the HTs.

Results on TRF. The TRF translations of English texts into German come out really well, i.e. close to the HTs – the TRF comes closest of all models here. Translations of English speech also comes close to the HTs. However, TRF translations of German speech into English is overly complex in comparison to the HTs. Translations of German texts are slightly more complex than the HTs.

6 Conclusions

Crossing PoS perplexity scores and syntactic dependency lengths, we draw some conclusions about the difference between human and machine structural translationese in text and speech. We summarize our findings as preliminary answers to the questions posed in the Introduction.

Q1. The structural particularity displayed by human interpreting is not replicated by MT models: machine translation of spoken language is not as marked as human interpreting. Human written translations are similar to comparable originals, while interpreting transcripts appear different from all other datasets, at least based on our measures. This feature fails to appear in any of the machine translations. If we look at MT models' outputs going from SMT to TRF, we see that written language perplexity for the target decreases, without getting smaller than the human "lower bound". For spoken language, perplexity in MT seems to outdo human levels with ease: for example, TRF reaches 7.2 where human perplexity is 12 (English to German Interpreting). This apparent improvement on speech-based data could confirm the idea that the special features of interpreting depend on the cognitive overload of human translators rather than to a systematic difficulty of translating speech.

Q2. Overall, we see a decrease of perplexity and an increase in syntactic complexity when moving from SMT to RNN to TRF, hinting at a tendency for SMT to produce over-simplified structures, while neural systems seem able to deal better with the complexity of their source. Figures 3 and 4 show these trends. With respect to the syntactic measures, we see clear tendencies in the human translations. The MTs, however, are more heterogeneous: translations by the SMT often over-simplify in contrast to the HTs, translations by the neural systems come out reasonably close to the HTs, i.e. close to originals, but they are sometimes more complex. In general the statistical engines show more evident signs of syntactic simplification than both human and neural translations. This can be the result of having a phrase-based translation (SMT) in contrast to sentence-based translations (humans and neural systems). The difference between architectures is less strong for the grammatical perplexities, where machine translation displays in general higher levels of structural shining-through than human translation, and lower levels of normalization, presenting more sensibility to the source language than human translation.

In general, while we find evident differences in translationese between human and machine translations of text and speech, our results are complex to analyze and it would be wise not to overinterpret our findings. In future work, the impact of the difference in the amount of available data and languages involved for written and spoken texts should also be analysed. Machine translations do present symptoms of structural translationese, but such symptoms only in part resemble those displayed by human translators and interpreters, and often follow independent patterns that we still have to understand in depth. This understanding can help in improving machine translation itself with simple techniques such as reranking of translation options, or more complex ones such as guiding the decoder to follow the desired patterns or rewarding their presence. For this, a complementary study on the correlation of translation quality with our translationese measures is needed.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 Information Density and Linguistic Encoding.

References

Nora Aranberri. 2020. Can translationese features help users select an MT system for post-editing? *Proce-*

samiento del Lenguaje Natural, 64:93–100.

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 EPTICExploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1):61–86.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Yuri Bizzoni, Elke Teich, Tom S Juzek, Katherine Menzel, and Heike Przybyl. 2019. Found in Interpreting: Detection and analysis of translationese using computational language models. Poster at SFB Colloquium, Saarbrücken.

François Chollet et al. 2015. Keras. `https://keras.io`.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-Scale Evidence of Dependency Length Minimization in 37 Languages. In *Proceedings of the National Academy of Sciences*, volume 112, pages 10336–10341.

Edward Gibson, Richard Futrell, Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *ArXiv*, abs/1906.09833.

Ewa Gumul. 2006. Explicitation in simultaneous interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures*, 7(2):171–190.

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.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 503–511. Springer.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates. *ParlaCLARIN: Creating and Using Parliamentary Corpora*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9:159–191.

Joakim Nivre, Mitchell Abrams, Zeljko Agić, and et al. 2019. Universal Dependencies 2.4.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318.

Marius Popescu. 2011. Studying translationese at the character level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 634–639, Hissar, Bulgaria. Association for Computational Linguistics.

Tiina Puurtinen. 2003. Features of translationese contradicting translation universals? *Corpus-based approaches to contrastive linguistics and translation studies*, 20:141.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2019. Translationese as a Language in ”Multilingual” NMT. *ArXiv*, abs/1911.03823.

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.

Diana Santos. 1995. On grammatical translationese. In *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*, pages 59–66.

Josef Schmied and Hildegard Schäffler. 1996. Approaching translationese through parallel and translation corpora. *Language and Computers*, 16:41–58.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (ACL 2016), Volume 1: Long Papers*, pages 1715–1725, Berlin, Germany.

Miriam Shlesinger. 1995. Shifts in cohesion in simultaneous interpreting. *The translator*, 1(2):193–214.

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.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 901–904.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Sonja Tirkkonen-Condit. 2002. Translationese –a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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. 2019a. The effect of translationese in machine translation test sets. *arXiv preprint arXiv:1906.08069*.

Mike Zhang and Antonio Toral. 2019b. 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.