MoTra 2021

Proceedings of the First Workshop on Modelling Translation - Translatology in the Digital Age

31 May, 2021
Saarland University
Saarbrücken, Germany

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-91-7929-613-1

Introduction

Translatology is the theoretical and practical study of translation. It combines insights from linguistics, the humanities, cognitive and computer science to understand the process of translating between languages and the particular features characterizing language in translation. Central concepts of contemporary translatology are translationese, linguistic patterns that tend to make translations more similar to each other than to texts originally written in their target language; and variation, which refers to the fact that different types of translations, such as written translations vs. interpreting, display systematic linguistic differences.

The **Workshop on Modelling Translation: Translatology in the Digital Age** seeks to facilitate collaboration and knowledge exchange between researchers in linguistics, AI, CL, NLP, translation studies, cognitive and computer science focusing on modeling translation from diverse angles, such as variation in translation, machine translation, translation quality assessment and translationese. Specifically, the workshop aims to foster innovative research at the intersection between machine and human translation modeling by applying concepts from translation studies to machine translation or using machine translation techniques to explore research questions in translatology. We encourage research on modeling aspects of translation, including word embeddings, neural or statistical machine translation, feature-based text classification, syntactic and semantic parsing, monolingual or multilingual language models, text generation, and stylometry. Our Call for Papers elicited contributions from a heterogeneous group of researchers. We are very happy to present 11 papers from diverse fields such as computational linguistics, computer science, and translation studies.

The papers cover topics ranging from the creation of more reliable interpreting corpora to the study of sentiment intensity in alternative translations. Major themes include a focus on methods to evaluate and explain linguistic variation in translations, new quantitative and experimental approaches, the creation of tools for translators and translation research and the need for data and corpora to better study translators' choices in all their aspects.

This workshop would not have been possible without the contributions of both authors and reviewers. We would like to thank everyone who submitted their work to this workshop and the program committee for their extensive and helpful reviews.

We would also like to thank our invited speakers, Jörg Tiedemann (University of Helsinki) and Markus Freitag (Google), for sharing their insights on this fascinating topic. Finally, we would like to thank all the attendees of the workshop. All of this contributes to a truly enriching event!

Yuri Bizzoni, Josef van Genabith, Cristina España i Bonet and Elke Teich

Saarbrücken

May 2021

Organisers:

Yuri Bizzoni
Elke Teich
Cristina España i Bonet
Josef van Genabith

Program Committee:

Silvia Bernardini, Yuri Bizzoni, Michael Carl, Cristina España i Bonet, Josef van Genabith, Alina Karakanta, Ekaterina Lapshinova-Koltunski, Antoni Oliver, Serge Sharoff, Antonio Toral, Elke Teich, Carl Vogel, Shuly Wintner

Invited Speakers:

Jörg Tiedemann, University of Helsinki
Markus Freitag, Google

Table of Contents

Do not Rely on Relay Translations: Multilingual Parallel Direct Europarl

Kwabena Amponsah-Kaakyire[1,2], **Daria Pylypenko**[1], **Cristina España-Bonet**[2], and **Josef van Genabith**[1,2]
[1]Saarland University, [2]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Saarbrücken, Germany
s8kwampo@stud.uni-saarland.de
daria.pylypenko@uni-saarland.de
{cristinae, Josef.Van_Genabith}@dfki.de

Abstract

Translationese data is a scarce and valuable resource. Traditionally, the proceedings of the European Parliament have been used for studying translationese phenomena since their metadata allows to distinguish between original and translated texts. However, translations are not always direct and we hypothesise that a pivot (also called "relay") language might alter the conclusions on translationese effects. In this work, we (*i*) isolate translations that have been done without an intermediate language in the Europarl proceedings from those that might have used a pivot language, and (*ii*) build comparable and parallel corpora with data aligned across multiple languages that therefore can be used for both machine translation and translation studies.

1 Introduction

Original text and text translated from another language differ in several characteristics (Gellerstam, 1986). The differences are assumed to be systematic and referred to as *translationese*. Translationese includes language independent characteristics like simplification, normalization, explicitation and avoiding repetitions (e.g., Baker et al. (1993)), as well as language-pair specific features, e.g. shining-through of source language patterns in target text (Toury, 1979; Teich, 2003).

In order to be successfully used for studying translationese phenomena, corpora need to be equipped with additional meta-information: whether the text is original or translated, the direction of translation, production mode of the source text (spoken/written) to give some examples. It is also useful to know whether the original text has been produced by a native speaker, as it has been shown that texts produced by non-native speakers can be quite easily separated from the texts produced by native speakers and translated texts (Nisioi et al., 2016). Information about native language and qualifications of the translator is also relevant.

For this reason, collecting multilingual (same-domain) data suitable for studying translationese is a challenging task. The proceedings of the European Parliament (*Europarl*) have often been used previously for this purpose (Koppel and Ordan, 2011; Rabinovich and Wintner, 2015; Lembersky et al., 2011), as they cover a lot of languages and provide relevant metadata. However, one problem with this data is that translation in the European Parliament sometimes happens indirectly, through pivot (also called "bridge" or "relay") languages. With 24 official languages, there are 552 possible direct translation combinations, therefore translations are often made first into one of the most frequently used languages: English, French or German, and then into other languages (Parliament, c; Katsarova, 2011). This can be problematic for studies that compare translations coming from different source languages. Unfortunately, there are no meta-annotations for the European Parliament proceedings that would indicate whether the translation has been indirect, and exactly which pivot languages have been used. According to Bogaert (2011); Parliament (a), the system of relay languages was introduced in 2004, when a number of states joined the EU, and the number of official languages grew from 9 to 20. We use this date for our main separation of the data.

The contributions of our paper are twofold: (*i*) we extract the unequivocally direct translations and (*ii*) we align the corpus paragraph-wise across seven languages: English (EN), French (FR), Spanish (ES), German (DE), Dutch (NL), Italian (IT) and Portuguese (PT), and provide scripts for extracting comparable and parallel subcorpora

from it.

The rest of the paper is organised as follows. Section 2 presents previous work done on building corpora for translationese research, and, in particular, corpora based on the proceedings of the European Parliament. Section 3 describes the procedure of creating the corpora. In Section 4, we compare the "reliable" and "unreliable" parts of the corpus on the task of translationese classification. Lastly, in Section 5 we present our conclusions and ideas for future work.

2 Related Work

2.1 Data available for translationese research

There are only a few multilingual corpora for translationese research. The UN parallel corpus (Ziemski et al., 2016) consists of multilingual parliamentary documents of the United Nations in 6 languages, organized into bilingual parallel corpora. From this corpus Tolochinsky et al. (2018) derived 5 parallel corpora from English into other languages and annotated them for translation direction.

The Canadian Hansard corpus[1] consists of transcriptions of the Canadian parliament in English and French and their translations, and has metadata indicating the original language.

Rabinovich et al. (2015) compile a parallel English–French corpus from TED talks, annotated for translation direction. They also provide aligned English–French and English–German book corpora, collected from public domain books, and an English–German corpus of political news and commentary collected from the Project Syndicate[2] and Diplomatisches Magazin[3].

2.2 Corpora based on Europarl proceedings

Many projects have focused on creating corpora based on the proceedings of the European Parliament, available in 24 languages. According to Nisioi et al. (2016), the proceedings are transcribed, edited and then translated by professional translators who are required to be native speakers of the target language (Pym et al., 2011). Koehn (2005) compiled the *Europarl corpus*: monolingual corpora and parallel corpora for 10 languages with English, and provided a sentence alignment tool. However their parallel corpora do not contain any meta-information, and the monolingual corpora have information that is not always consistent and also scarce, according to Karakanta et al. (2018). Graën et al. (2014) attempted to clean and correct some errors in the Europarl corpus of Koehn (2005). Islam and Mehler (2012); Lembersky et al. (2011); Rabinovich et al. (2015) and Cartoni and Meyer (2012) employed the Europarl corpus of Koehn (2005) for translation studies, relying on its metadata ("language tags"). Ustaszewski (2019) created the EuroparlExtract toolkit that allows extraction of bilingual parallel corpora and monolingual comparable corpora from the Europarl corpus of Koehn (2005) with explicit annotation of translation direction and source language. They also rely on the metadata present in the Europarl corpus of Koehn (2005). Nisioi et al. (2016) additionally crawl the information about the Members of the European Parliament (MEPs) from the European Parliament's website in order to identify native or non-native speakers.

Karakanta et al. (2018), in contrast to the previous approaches, do not use the Europarl corpus of Koehn (2005), but provide a pipeline (Europarl-UdS) for re-crawling the European Parliament proceedings from the official website of the European Parliament[4], as well as MEP meta-information, and compiling comparable corpora annotated with information about the original language and the status of the speaker (native/non-native). We build upon their approach and enable multilingual paragraph-level parallelization of texts, as well as add metadata about direct/possibly indirect translation.

2.3 Pivot languages

The issue with relay languages in translation of the European Parliament proceedings has been raised previously by researchers in linguistics and translation studies.

Cartoni and Meyer (2012); Cartoni et al. (2013) claim that a corpus that contains indirect translations cannot be reliable for studies aiming to analyze a translation from a specific source language into a specific target language, however it could still be used for comparison between the original and translated texts in general.

Rabinovich (2018) use Europarl of Koehn (2005) spanning from years 1999 to 2011, and

[1] https://www.english-corpora.org/hansard/

[2] https://www.project-syndicate.org/

[3] http://www.diplomatisches-magazin.de/

[4] http://www.europarl.europa.eu/

iid	src	ns	dir	org	de	en	es	fr	it	nl	pt
199907...	de	1	1	Frau Präs...	Frau Präs...	Madam P...	Señora P...	Madame l...	Signora P...	Mevrouw ...	Senhora ...
201006...	en	1	0	I would al...		I would al...		J'invite ég...	Invito inol...	Ik dring er...	
200612...	fr	0	0	Tout au lo...	In der Wo...	Through....	Durante t...	Tout au lo...	Durante l'...	De week ...	
200302...	it	1	1	L'ultima c...		My last p...	Por últim...		L'ultima c...		O meu últ...
200204...	nl	0	1	Ten eerst...			En primer...	Première...		Ten eerst...	Em prime...

Figure 1: Sample lines from the initial corpus extracted from the xml files and aligned across 7 languages. The columns from left to right: paragraph id, source language, native/non-native speaker, direct/undefined translation, the originally produced paragraph in its original language, translations into all of the languages. The initial aligned corpus contains blank cells where the translations are missing.

treat all the translations into languages other than English as indirect. They perform source language identification and phylogenetic tree construction on English and French translations from various languages, and report that the translationese signal seems to weaken due to the pivot translation, however it is still identifiable.

Ustaszewski (2021) use corpora extracted with the EuroparlExtract toolkit (Ustaszewski, 2019), and treat the translations from 2004 onwards as English-mediated. They perform classification between direct and indirect translations, whereas we classify translations vs. original texts.

3 Multilingual Parallel Direct Europarl

This section describes how we build the multilingual corpus with parallel data for both machine translation and translation studies from the Europarl proceedings. Our corpus has originals and translations available in 7 languages: Dutch, English, French, German, Italian, Portuguese and Spanish.

We firstly use the code[5] provided by Karakanta et al. (2018) to extract the Europarl proceedings from the official website into metadata-rich xml files. Subsequently, we align the data across the 7 languages. Figure 1 visualizes a sample of the aligned dataset. The alignment is done on a paragraph basis[6]. On average, a paragraph has 78 words. In aligning the segments, we take into consideration the number of paragraphs in each speech (intervention). In the different parallel interventions, the different translations are sometimes organised into different number of paragraphs. We only consider interventions whose translations are aligned paragraph-wise.

According to Parliament (a,b,c) and Bogaert (2011), since 2004 translations, especially for less widely-used languages, are *mostly* made through pivot languages. Due to the lack of meta-annotations, it is not possible to ascertain which translations from 2004 onwards are direct translations and which are not. Since the information about whether translations are direct or through a relay language is important for studying translationese, we annotate all translations up to 2003 as *direct* to separate them from the data that might possibly contain pivot translations, which we denote as $undefined$.

In addition to this, we also use annotations from the xml files from which the data is extracted, based on the nationality of a speaker to annotate which texts were produced by native speakers and which were not. This however is not guaranteed to be a perfect annotation as people sometimes naturalise to become citizens of other countries; speakers may also have a minority language in the country of origin as their mother tongue, and finally, the writers of a speech may not be identical to the MEPs who gave the speech. This however helps, to a large extent, to distinguish a greater portion of non-native from native-speaker text for studies where this is required or desired.

We provide scripts[7] to extract parallel and comparable corpora of all possible combinations of the

[5] https://github.com/hut-b7/europarl-uds

[6] This is due to the fact that the translations of paragraphs are not aligned sentence-wise. While the original paragraph may have n sentences, one translation may have m sentences and another k.

[7] https://github.com/UDS-SFB-B6-Datasets/Multilingual-Parallel-Direct-Europarl

	Direct	Undefined	All
Native	119k	245k	364k
Non-native	118k	313k	431k
All	237k	558k	795k

Table 1: Number of aligned paragraphs in the 7-language initial corpora extracted from the xml proceedings with different filtering options.

	Direct	Undefined	All
Native	51k	66k	138k
Non-Native	11k	15k	26k
All	73k	99k	196k

Table 2: Number of aligned paragraphs in the fully parallel 7-language datasets, balanced by the source language.

7 languages, and filtering options i.e. native/non-native speaker and direct/undefined translations.

Tables 1, 2 and 3 show statistics for these extractions for all 7 languages. Table 1 shows the number of aligned paragraphs for the initial corpora extracted from the xml parliamentary proceedings, as depicted on Figure 1. In this case, not all the entries have the translations into all 7 languages, but the scripts allow to select fully aligned parallel subsets for any combination of languages. Table 2 corresponds to the most restrictive case, the fully parallel 7-language datasets, i.e. the entries where translations into any one of the languages are missing have been removed. Additionally, each of these datasets has been balanced to have the same number of entries per source language (second column in Figure 1). Finally, Table 3 shows statistics for the translationese comparable corpora. All of the comparable corpora mentioned in this table have structure as shown in Figure 2. We extract original and translations paragraphs in equal proportions. The *originals* part contains texts in 7 languages and the the *translationese* part contains translated texts in 7 languages in equal proportions, where for each language these are translations from 6 languages also in equal proportions.

4 Translationese Classification

In order to see if the purity of the resulting corpus affects distinguishability of translations and originals, we perform a first naïve translationese classification task on both direct and undefined translations for a subset of languages (English, German and Spanish), but leave a deep analysis of the topic for future work. The classification is done on the balanced subsets of direct (up to 2003) and undefined (after 2003) data using both native and non-native speaker data. We perform classification on monolingual comparable corpora, which have an analogous structure to the multilingual corpus shown in Figure 2, however there is only one target and only one source language. These corpora were extracted with the scripts that we provide, since they allow extraction of the corpora for any combinations of the 7 languages. Thus half of each corpus is made up of original texts, and the

		Direct	Undef.	All
Native	Orig.	52k	82k	162k
	Trans.	52k	82k	162k
Non-native	Orig.	53k	92k	160k
	Trans.	53k	92k	160k
All	Orig.	136k	354k	490k
	Trans.	136k	354k	490k

Table 3: Paragraph count in the 7-language comparable corpora for translationese classification: originals (Orig.) and translations (Trans.).

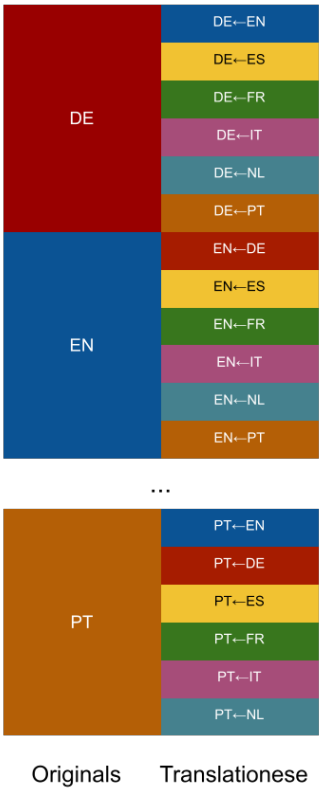

Figure 2: Structure of a 7-language comparable corpus for translationese classification.

Language		Accuracy		Δ
Text	**Source**	**Direct**	**Undefined**	**D-U**
DE	EN	71.08	69.46	+1.62
DE	ES	74.93	72.46	+2.47
EN	DE	69.55	66.46	+3.09
EN	ES	70.34	66.79	+3.55
ES	DE	70.12	70.80	-0.68
ES	EN	67.04	69.90	-2.86
Average		70.51	69.31	+1.20

Table 4: Translationese classification results (accuracy) and difference between direct (D) and undefined (U) accuracies (Δ).

other half consists of translations from a certain language, e.g. English originals vs. translations from Spanish into English. We perform classification on 6 possible combinations of 3 languages: German, English and Spanish. For each combination, the training set contains 29k paragraphs, test and validation set contain 6k paragraphs each.

We train a Support Vector Machine classifier with a linear kernel. The *INFODENS* toolkit (Taie et al., 2018) is used to extract features and to train and evaluate the classifier. We tune the regularization parameter C on the validation set. We use a subset of the features provided by the toolkit inspired by the optimised feature selection approach in Rubino et al. (2016), and add custom backward language modelling features[8]. In particular, we use 108 features divided as:

- surface features: average word length, syllable ratio, sentence length;
- lexical features: lexical density, type-token ratio;
- unigram bag of PoS;
- language modelling features: log probabilities and perplexities, according to the forward and backward n-gram language models ($n \in [1;5]$) built on tokens and PoS-tags;
- n-gram frequency distribution features: percentages of n-grams in the paragraph occurring in each quartile ($n \in [1;5]$).

The n-gram language models are estimated with SRILM (Stolcke, 2002) and spaCy[9] is used for tokenizing and PoS-tagging the texts.

[8] https://github.com/daria-pylypenko/B6-SFB1102

[9] https://spacy.io/

Our results are reported in Table 4. We observe that accuracy for direct translations only is higher than for undefined in most cases, but not always. We assume that only the direct translations provide us with the reliable results, since for the undefined part we do not know the exact proportion of direct and pivot translations. For the undefined part, we also hypothesize that accuracy will depend on the distance between the pivot and the source language: it will determine whether translationese features of the original source will be amplified, overridden or left intact during the second translation and this is why the accuracy in the classification might be changing with respect to the direct translation texts. However, due to the fact that we do not have pivot language annotations, the hypothesis cannot be confirmed or rejected. According to our results, translationese effects are more evident in German text (highest accuracy, therefore easiest text to classify), whereas Spanish text coming from English is the most difficult to detect (accuracy of 75% vs. 67%). Undefined translations, however, diminish the difference (72% vs. 70%).

5 Conclusions and Future Work

We have presented a corpus based on the proceedings of the European Parliament, aligned across 7 languages on a paragraph level, and scripts for extracting parallel and comparable subcorpora for all combinations of these languages. We have also enabled subsampling the corpus to extract the part of the data that consists only of direct translations, as opposed to data with unknown status. The corpus is suitable for translation studies and machine translation.

Future work could involve extending the paragraph-level alignment to sentence level. Moreover, indirect translation is a multi-faceted research topic (Pieta, 2019), and it would be interesting to examine it in the context of translationese. Since the pivot language annotations for the Europarl proceedings are not available, another future work direction could be to study influence of pivot languages in machine translationese.

Acknowledgements

This research is funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft*) under grant SFB 1102: Information Density and Linguistic Encoding.

References

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, page 233–, Netherlands. John Benjamins Publishing Company.

Caroline Bogaert. 2011. Is absolute multilingualism maintainable? The language policy of the European Parliament and the threat of English as a lingua franca. Master's thesis, UGent. Faculteit Letteren en Wijsbegeerte.

Bruno Cartoni and Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2132–2137, Istanbul, Turkey. European Language Resources Association (ELRA).

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. 2013. Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27:23–42.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Johannes Graën, Dolores Batinic, and Martin Volk. 2014. Cleaning the Europarl Corpus for Linguistic Applications. In *Conference Proceedings of the 12th Konvens*, pages 222–227.

Zahurul Islam and Alexander Mehler. 2012. Customization of the Europarl corpus for translation studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2505–2510, Istanbul, Turkey. European Language Resources Association (ELRA).

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Europarl-uds: Preserving and extending metadata in parliamentary debates. In *Proceedings of the LREC 2018*, Miyazaki, Japan.

Ivana Katsarova. 2011. The EU and multilingualism. `http://www.europarl.europa.eu/RegData/bibliotheque/briefing/2011/110248/LDM_BRI(2011)110248_REV1_EN.pdf`.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, pages 79–86, Phuket, Thailand.

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.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. 2016. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4197–4201, Portorož, Slovenia. European Language Resources Association (ELRA).

European Parliament. a. EP Translators. `https://www.europarl.europa.eu/pdf/multilinguisme/EP_translators_en.pdf`.

European Parliament. b. European parliament - never lost in translation. `https://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT%20IM-PRESS%2020071017FCS11816%200%20DOC%20XML%20V0//EN`.

European Parliament. c. Which languages are in use in the Parliament? `https://www.europarl.europa.eu/news/en/faq/21/which-languages-are-in-use-in-the-parliament`.

Hanna Pieta. 2019. Indirect translation: Main trends in practice and research. *Slovo.ru: Baltic accent*, 10:21–36.

Anthony Pym, François Grin, Claudio Sfreddo, and A.L.J. Chan. 2011. The status of the translation profession in the european union. *The Status of the Translation Profession in the European Union*, pages 1–182.

Ella Rabinovich. 2018. *A Computational Approach to the Study of Multilingualism*. Ph.D. thesis.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2015. The Haifa Corpus of Translationese. *CoRR*, abs/1509.03611.

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-HLT 2016, Association for Computational Linguistics*, pages 960–970, San Diego, California.

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

Ahmad Taie, Raphael Rubino, and Josef van Genabith. 2018. INFODENS: An Open-source Framework for Learning Text Representations. *arXiv preprint arXiv:1810.07091*.

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

Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. 2018. The UN parallel corpus annotated for translation direction. *CoRR*, abs/1805.07697.

Gideon Toury. 1979. Interlanguage and its manifestations in translation. *Meta*, 24(2):223–231.

Michael Ustaszewski. 2019. Optimising the europarl corpus for translation studies with the europarlextract toolkit. *Perspectives*, 27(1):107–123.

Michael Ustaszewski. 2021. Towards a machine learning approach to the analysis of indirect translation. *Translation Studies*, 0(0):1–19.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

HeiCIC: A simultaneous interpreting corpus combining product and pre-process data

Kerstin Kunz
Ruprecht-Karls-Universität Heidelberg
kerstin.kunz@iued.uni-heidelberg.de

Christoph Stoll
Ruprecht-Karls-Universität Heidelberg
christoph.stoll@iued.uni-heidelberg.de

Eva Klüber
Ruprecht-Karls-Universität Heidelberg
eva.klueber@iued.uni-heidelberg.de

Abstract

This paper presents HeiCIC, a simultaneous interpreting corpus that comprises audio files, time-aligned transcripts and corresponding preparation material complemented by annotation layers. The corpus serves the pursuit of a range of research questions focusing on strategic cognitive load management and its effects on the interpreting output. One research objective is the analysis of semantic transfer as a function of problem triggers in the source text which represent potential cognitive load peaks. Another research approach correlates problem triggers with solution cues in the visual support material used by interpreters in the booth. Interpreting strategies based on this priming reduce cognitive load during SI.

1 Motivation

The aim of this paper is twofold: We present the architecture and on-going collation of a series of simultaneous interpreting (SI) subcorpora, integrated in the Heidelberg Conference Interpreting Corpus (HeiCIC): HeiCIC contains authentic speeches from LSP domains with simultaneous interpretations by learners and professionals in eight languages. The English-German core corpus is aligned with pre-process data that visualize the established conference interpreting workflow.

The pre-process data we are interested in is the visual support material which is used by interpreters to cope with expected problem triggers (PT) in a speech and to avoid peaks in cognitive processing.

Conference Interpreters are trained to condense the logical structures and PT of source texts as cues to target text solutions using a special note-taking technique for consecutive interpreting. The result of the pre-process PT analysis for simultaneous interpreting is noted in a similar fashion: as an amalgamation of source text logic, terminology and cues for cognitive load relief in a visio-spatial structure or 'map' of the thought processes (Gile, 1995; Stoll, 2009).

More precisely, this map combines expected source language macrostructures, conceptual relations and terminology with cues to trigger target language structures with cognitive load-relieving interpreting strategies. These include memory relief, listening analysis and comprehension relief, patterns for target text production and strategies for easier output monitoring using top-down and bottom-up plausibility checks (Gile, 1995; Stoll, 2009).

Furthermore, we introduce the research in progress to be done on the core corpus: Our on-going research has two objectives: a) analysing semantic transfer from source to target text in relation to expected problem triggers in the source text and b) correlating semantic transfer with pre-process data to determine which features reflect high performance SI strategies. In this way, our empirical research combines product- and process-related studies.

There are several aspects that set the corpus apart from other SI corpora: To date, no large, comparative learner/professional LSP corpus exists for SI, least for the language combinations in focus here. There are some learner corpora for Chinese <-> English, such as the learner corpus from Leung and Yip containing interpretations of nine trainees (Bendazzoli, 2018; Leung and Yip, 2013; Zhang, 2017), which are however rather limited in size. Existing professional interpreter corpora are larger but differ in terms of metadata: For instance, EPIC, EPTIC and EPICG (Bernardini et al., 2018) focus on interpreting in the institutional setting of the European Union and therefore are rather heterogeneous in terms of topic, register and level of technicality. NAIST (Japanese - English) (Neubig et al., 2018), (387,000 word and comparable to HeiCIC in size) reflects interpreting environments for a general/non-expert audience. Other SI corpora incorporate other forms of interpreting such as SIREN, which includes simultaneous interpreting with text and television interpreting in English and Russian in its 33.55h (235,040 words) of records (Dayter, 2018).

HeiCIC is designed to map authentic professional settings, where the highly technical nature of LSP and scientific conferences requires a structured, partially automated workflow for knowledge acquisition, content organization and terminology management. Our corpus design is unique in that it aligns this pre-process data with both original speeches and interpreting output. This permits insights into advanced interpreting strategies used in LSP settings and thus process-related phenomena, while other corpora typically focus on product data (Gile, 2002; Díaz Galaz, 2015).

2 Data collection and corpus design

HeiCIC is collated mainly at the Heidelberg Conferences: scientists and experts present their research in a variety of LSP domains and send preparation material, which is used by interpreters with different levels of expertise (students at MA level from the second to the final semester, young and seasoned professionals) to prepare and then interpret from, into and between German, English, French, Italian, Spanish, Portuguese, Russian and Japanese. Subcorpora differ in terms of formats available, languages included, LSP domains covered and level of interpreter expertise.

The core corpus is a homogeneous subpart containing several parallel interpretations by students, professionals with different levels of interpreter expertise, and transcripts (English <-> German) in selected LSP domains such as electrical engineering in car manufacturing, astronomy, investor relations and annual general meetings (AGMs) of international corporations. It currently contains recorded speeches and interpretations of around 83 hours with transcripts comprising around 400,000 tokens and is constantly expanded as new recordings, transcripts and annotation layers are added.

We seek to follow basic principles of corpus compilation (Bernardini et al., 2018; Hansen-Schirra et al., 2012). Metadata are stored in a separate file for each transcript. They are structured as follows: information about speaker (e.g. gender, role, native language and language variety), interpreter (e.g. gender, level of expertise, native language and language combination) and text (e.g. setting, language, register, topic and mode, text length in seconds and tokens) and allow for filtering according to these criteria.

In addition, transcripts, recordings and annotation layers are aligned with strategic pre-process data of interpreters. Pre-process data, which includes visual preparation material created by interpreters, is available in an electronic format and attributed to the individual interpreter, target and source text combination.

2.1 Transcription

The transcription process used to provide the transcripts as a basis for analysis includes several steps and is partially automated. Transcripts are generated automatically using automatic speech recognition and corrected by manual revision.

We apply transcription guidelines which are a slightly modified version of those for the GECCo Corpus (Kunz et al., 2011; Lapshinova et al., 2012). They include tags accounting for spoken language features (such as non-standard language, truncated or repeated words), tags related to cognitive load in general (such as filled and silent pauses), and tags related to SI in particular (such as interpreter turns, incomplete sentences and grammatical errors), (Plevoets and Defrancq, 2016). For instance, Example 1 shows tags for truncated words and phrases and fillers.

> […] In case of a mosquito bite, [t=or a malaria] malaria [t=is] [ehm] [t=can] is supposed to be the case. […]

Example 1. Transcription tags for spoken language features.

Revised transcripts are automatically time-aligned with the audio signal using WebMAUS (Kisler et al., 2017). The resulting files are further processed with EXMARaLDA. In combination with the time-aligned transcripts, this allows for alignment of several interpretations with one original speech (Schmidt and Wörner, 2014).

2.2 Annotation and alignment

The core part of the corpus contains automatic basic level annotations, such as tokenization, lemmatization and POS tagging. The performance of the latter is improved via additional renderings during transcription (see above example). In addition, semi-automatic and manual annotation layers are added in alignment with the current research objectives (see more details below). Main annotations include information on problem triggers in the source text and on semantic transfer between source and target text. Manual annotation steps are performed by several annotators. Each source text is currently annotated by two skilled student annotators. We regularly evaluate annotator agreement to ensure high annotation quality and to improve detailed annotation guidelines.

In order to analyse correlations between process and product data, we include several alignments: Problem triggers in the source text are aligned with respective renderings in the visual support material and corresponding expressions in the target texts. Moreover, solution cues marked in the visual support material are related to indicators of interpreting strategies in the target text.

3 Problem triggers

In a first step we annotate source texts for problem triggers representing potential cognitive load peaks in original texts (Gile, 2009). We focus on "problem triggers pertaining to the message", as classified by Mankauskienė (2016: 146). This type is structured further into categories such as numbers, proper nouns, collocations, terminology and complex phrases. Sender-related problem triggers (e.g. accent) or technical problem triggers can be integrated at later stages of the project. We currently implement procedures for semi-automatic extraction and manual post-correction for some of these categories (e.g. terminology and numbers). Other categories, e.g. complex phrases are annotated manually. Double annotation is possible, meaning that one source text element can incorporate several problem triggers.

4 Semantic transfer

In a second step, (non-)renderings corresponding to problem triggers are identified in the respective target texts and grouped into transfer categories specifying their relation to the source text problem trigger. Transfer categories focus on semantic relations with category options determined by the problem trigger category.

This serves as a basis for the analysis of semantic transfer from source to target text, i.e. the reproduction of a message uttered in one language into another (Schjoldager, 1995). Problem trigger renderings are not analysed in isolation, but within the units of meaning in which they occur to allow for a more comprehensive analysis of semantic transfer from source to target text. For this purpose,

interpreting units are identified in both source and target text based on functional, semantic and syntactical information (Alves et al., 2019; Christoffels and de Groot, 2005).

Semantic transfer is defined as the relation between source and target interpreting units on a scale from omission and implicitation to explicitation and addition and analysed by assessing features contained in the interpreting units in terms of their structure and their semantic content (Becher, 2011; Hansen-Schirra et al., 2012). The semantic content is categorised in terms of explicitness: Words (or expressions) that can potentially encode a higher semantic range than others are classified as less explicit than words (or expressions) that have a narrower semantic range (Gumul, 2017). Semantic transfer may be encoded using different means, for example substitution such as pronouns or hyponyms or hypernyms in the target text in relation to the source text segment. Examples 2 and 3 show instances of the semantic transfer categories implicitation by substitution and omission of part of a segment.

	source text	target text	semantic transfer
47	Why is this tiredness warning system useful?	Wieso ist dies hilfreich?	implicitation

Example 2. Semantic transfer: implicitation.

	source text	target text	semantic transfer
43	In other words, you can remain in the navigation system or rate your list view and still change the driving mode for the car at the push of a button.	Man kann beispielsweise während des Navigationsmodus den Effizienzmodus einschalten auf Tasterdruck.	omission

Example 3. Semantic transfer: omission.

The focus of analysis lies on interpreting units that contain problem triggers as they potentially provide insights into the effect of cognitive load peaks on semantic transfer (Mankauskienė, 2016). Shifts in the position of interpreting units within sentence and text structures are analysed as well.

Previous studies on SI have focused either on individual transfer phenomena such as explicitation or on linguistic features such as cohesion markers (Kajzer-Wietrzny, 2012). To our knowledge, a comprehensive analysis of the semantic content of interpreting units and transfer categories in combination with the analysis of information structure has not been attempted so far.

5 Visual support material

In a third step, the properties of the interpretation output are correlated with pre-process data: visual support materials prepared by interpreters as a substantial part of the interpretation workflow.

As widely agreed in research on simultaneous interpreting, conference preparation goes beyond the bilingual organization of terminology and glossaries, notably in alphabetical order (Rütten, 2007; Will, 2009). Visual support material ideally combines information on expected content with organizations of concepts and terminology (Stoll, 2009 and 2019). It contains chronological renderings of expected macrotopics reflecting textual function and skopos. Macrotopics are complemented on the microlevel as ontological representations of concepts (i.e. semantic relations and semantic roles) and mapped onto terminological expressions.

Furthermore, these visio-spatial maps integrate simultaneous interpreting strategies, i.e. strategy cues relating predictions of source language problem triggers such as cognitive load conflicts and overruns (Seeber, 2011-17) to efficient target language solutions (Stoll, 2019). Some examples are structures related to listening comprehension enhancing anticipation/priming of collocations, complex syntactic structures and terminology.

For instance, the source text cue revenue in an earnings release event semantically primes the hypernym, KPI (key performance indicators for corporations) and other co-hyponyms such as earnings and profit. The target language solutions (“Umsatz, Absatz, Ertrag”) are directly

linked to the semantic priming by the cue 'revenue' in the visual support material (cue map). Shortcuts from consecutive note taking are used to indicate such semantic relations.

Speech production and monitoring effort relief strategies in the visual map use domain specific jargon compression, e.g. Luftwiderstands-Beiwert (“aerodynamic drag coefficient”) is rendered as “drag”. Other strategies replace complex syntactic structures by prosodic and cohesive elements.

These electronic maps of pre-process thoughts are mind-map-like multidimensional structures that tap into the interpreter skillset: layout patterns and symbols from consecutive note-taking in relational databases, xml structures, spread sheets, and multi-layered documents bear tangible and - correlatable testimony to the categories of cognition moved upstream in the interpreting workflow in several dimensions: In keeping with professional practice, conceptual and terminological information is combined into a single structure with different views for pre- and in-process phases (Stoll, 2009; Fantinuoli, 2012): While the pre-process view shapes terminology and expert knowledge into an ontological hierarchy (Rütten, 2007; Will, 2009), the in-process view lists macrotopics, semantic relations, terminology and strategy cues in chronological order. Thus, visual support material used in the booth is a condensed in-process version of the pre-process map (Stoll, 2009). The level of condensation may vary, depending on the level of expertise and familiarity with the topic and register.

6 Correlating product and process data

Our approach aims to determine which features in visual support materials used in the booth can be identified as solution cues and therefore indicators of deliberate high-performance SI strategies as they correlate with the interpreter’s output, thus proving process in product features. Correlating problem triggers in the source text with semantic transfer categories and thus interpreting output on the one hand, and with entries in the support material on the other hand, should yield information as to how predictions of source language problem triggers are marked and strategically related to efficient target language solution cues. They may then be assigned to individual types of cognitive load, as mentioned above. Moreover, our analyses may reveal whether and how these entries in the visual support material relate to solutions in the interpretation output. In this, we invert the traditional errors-and-omissions-based approach to establish an evidence-based, hierarchical typology of verifiable strategies of semantic, conceptual, lexical and strategic priming.

Insights obtained may serve to optimize the organization of electronic visual support material in general and improve CAI tools for in-process use, contributing to augmented interpretation.

We plan to make our corpus accessible for corpus-querying via a web interface such as CQPWeb for independent validation, validity and reliability of our research. The corpus is well documented to permit research beyond our current focus in the future.

References

Fabio Alves, Adriana Pagano, Stella Neumann, Erich Steiner and Sylvia Hansen-Schirra. 2019. Translation units and grammatical shifts. David B. Sawyer, Frank Austermühl and Vanessa Enríquez Raído. (Eds.) *The Evolving Curriculum in Interpreter and Translator Education, American Translators Association Scholarly Monograph* Series, XV. John Benjamins Publishing Company, Amsterdam. 109–142.

Viktor Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts.* Hamburg, Universität Hamburg.

Claudio Bendazzoli. 2018. Corpus-based Interpreting Studies: Past, Present and Future Developments of a (Wired) Cottage Industry. Mariachiara, Claudio Bendazzoli, Bart Defrancq (Eds.) *Making way in corpus-based interpreting studies. New Frontiers in Translation Studies*. Springer, Singapore. 1–20.

Silvia Bernardini, Adriano Ferraresi, Mariachiara Russo, Camille Collard and Bart Defrancq. 2018. Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task. Mariachiara Russo,

Claudio Bendalozzi, and Bart Defrancq. (Ed.) *Making Way in Corpus-based Interpreting Studies*. Singapore, Springer.

Ingrid K. Christoffels and Annette M. B. de Groot. 2005. Simultaneous interpreting: A cognitive perspective. Judith F. and Annette M. B. de Groot. (Ed.). *Handbook of Bilingualism: Psycholinguistic Approaches*. New York, Oxford University Press. 454–479.

Daria Dayter. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM* 16:2. 241–264.

Stephanie Díaz-Galaz. 2015. *La influencia del conocimiento previo en la interpretación simultánea de discursos especializados: Un estudio empírico*. PhD thesis, Universidad de Granada.

Claudio Fantinuoli. 2012. *InterpretBank - Design and Implementation of a Terminology and Knowledge Management Software for Conference Interpreters*. Berlin, epubli GmbH.

Daniel Gile. 2002. The Interpreter's Preparation for Technical Conferences: Methodological Questions in Investigating the Topic. *Conference Interpretation and Translation* 4:2. 7-27.

Ewa Gumul. 2017. Explicitation and directionality in simultaneous interpreting. *Linguistica Silesiana*, 2017. 311-329.

Sylvia Hansen-Schirra, Stella Neumann and Erich Steiner. 2012. *Cross-linguistic corpora for the Study of Translation: Insights from the Language-Pair English German*. Berlin, de Gruyter.

Marta Kajzer-Wietrzny. 2012. *Interpreting universals and interpreting style*. PhD dissertation, Adam Mickiewicz University.

Cynthia Jane Mary Kellett Bidoli. 2016. Methodological challenges in Consecutive Interpreting Research: Corpus analysis of notes. Claudio Bendazzoli and Claudia Monacelli. (Ed.) *Addressing methodological challenges in Interpreting Studies Research*. Newcastle upon Tyne, Cambridge Scholars. 141-169.

Thomas Kisler, Uwe Reichel and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–347.

Ekaterina Lapshinova-Koltunski. Kerstin Kunz and Marilisa Amoia. 2012. Compiling a Multilingual Spoken Corpus. *Spoken Corpora and Annotation; Speech Technology and Data Bases. Proceedings of the VIIth GSCP International Conference*. Firence, Firence University Press.

Leung, S.M. Ester and Leonard Yip. 2013. *A bilingual corpus of interpreting students' performance*. http://arts.hkbu.edu.hk/~engester/main.html.

Dalia Mankauskienė. 2016. Problem trigger classification and its applications for empirical research. *Procedia - Social and Behavioral Sciences* 231. 143 – 148

Graham Neubig, Hiroaki Shimizu, Sakriani Sakti, Satoshi Nakamura and Tomoki Toda. 2018. The NAIST Simultaneous Translation Corpus. Mariachiara Russo, Claudio Bendalozzi and Bart Defrancq. (Ed.) *Making Way in Corpus-based Interpreting Studies*. Singapore, Springer.

Koen Plevoets and Bart Defrancq. 2016. The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. *Translation and Interpreting Studies*, 11 (2): 202-224.

Anja Rütten. 2007. *Information and Knowledge Management in Conference Interpreting (in German)*. Frankfurt, Lang.

Anne Schjoldager. 1995. An Exploratory Study of Translational Norms in Simultaneous Interpreting: Methodological Reflections. *Hermes, Journal of Linguistics*, 8 (14): 65-88.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. *Handbook on Corpus Phonology*. Oxford, Oxford University Press. 402-419.

Kilian Seeber. 2011. Cognitive load in simultaneous interpreting. Existing theories – new models. *Interpreting*, 13 (2): 176-204.

Kilian Seeber. 2013. Cognitive load in simultaneous interpreting: Measures and methods. *Target*, 25 (1): 18-32.

Kilian Seeber. 2017. Multimodal processing in simultaneous interpreting. John W. Schwieter and Aline Ferreira. (Ed.) *The Handbook of translation and cognition*. New Jersey, Wiley Blackwell.

Christoph Stoll. 2002. Dolmetschen und neue Technologien. Joanna Best and Sylvia Kalina. (Ed.) *Übersetzen und Dolmetschen. Eine Orientierungshilfe*. Tübingen, Francke. 307-312.

Christoph Stoll. 2009. *Jenseits simultanfähiger Terminologiesysteme. Methoden der Vorverlagerung von Kognition im Arbeitsverlauf professioneller Konferenzdolmetscher*. Trier, WVT.

Christoph Stoll. 2019. Terminology Systems and Workflow Automation for Simultaneous Interpreters: CAI tools and Research within the HeiCiC Corpus (in German). *edition*, 1/2019. 25-33.

Martin Will. 2009. *Interpreting-Oriented Terminology Work (in German).* Tübingen, Narr.

Wei Zhang. 2017. Chinese interpreting learner corpus construction and research: Theory and practice (in Chinese). *Chinese Translators Journal*, 38 (1): 53-60

Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods

Lifeng Han[1], **Gareth J. F. Jones**[1], and **Alan F. Smeaton**[2]
[1] ADAPT Research Centre
[2] Insight Centre for Data Analytics
School of Computing, Dublin City University, Dublin, Ireland
lifeng.han@adaptcentre.ie

Abstract

To facilitate effective translation modeling and translation studies, one of the crucial questions to address is how to assess translation quality. From the perspectives of accuracy, reliability, repeatability and cost, translation quality assessment (TQA) itself is a rich and challenging task. In this work, we present a high-level and concise survey of TQA methods, including both manual judgement criteria and automated evaluation metrics, which we classify into further detailed sub-categories. We hope that this work will be an asset for both translation model researchers and quality assessment researchers. In addition, we hope that it will enable practitioners to quickly develop a better understanding of the conventional TQA field, and to find corresponding closely relevant evaluation solutions for their own needs. This work may also serve inspire further development of quality assessment and evaluation methodologies for other natural language processing (NLP) tasks in addition to machine translation (MT), such as automatic text summarization (ATS), natural language understanding (NLU) and natural language generation (NLG). [1]

1 Introduction

Machine translation (MT) research, starting from the 1950s (Weaver, 1955), has been one of the main research topics in computational linguistics (CL) and natural language processing (NLP), and has influenced and been influenced by several other language processing tasks such as parsing and language modeling. Starting from rule-based methods to example-based, and then statistical methods (Brown et al., 1993; Och and Ney, 2003; Chiang, 2005; Koehn, 2010), to the current paradigm of neural network structures (Cho et al., 2014; Johnson et al., 2016; Vaswani et al., 2017; Lample and Conneau, 2019), MT quality continue to improve. However, as MT and translation quality assessment (TQA) researchers report, MT outputs are still far from reaching human parity (Läubli et al., 2018; Läubli et al., 2020; Han et al., 2020a). MT quality assessment is thus still an important task to facilitate MT research itself, and also for downstream applications. TQA remains a challenging and difficult task because of the richness, variety, and ambiguity phenomena of natural language itself, e.g. the same concept can be expressed in different word structures and patterns in different languages, even inside one language (Arnold, 2003).

In this work, we introduce human judgement and evaluation (HJE) criteria that have been used in standard international shared tasks and more broadly, such as NIST (LI, 2005), WMT (Koehn and Monz, 2006a; Callison-Burch et al., 2007a, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020), and IWSLT (Eck and Hori, 2005; Paul, 2009; Paul et al., 2010; Federico et al., 2011). We then introduce automated TQA methods, including the automatic evaluation metrics that were proposed inside these shared tasks and beyond. Regarding Human Assessment (HA) methods, we categorise them into traditional and advanced sets, with the first set including intelligibility, fidelity, fluency, adequacy, and comprehension, and the second set including task-oriented, extended criteria, utilizing post-editing, segment ranking, crowd source intelligence (direct assessment), and revisiting traditional criteria.

Regarding automated TQA methods, we classify these into three categories including simple n-gram based word surface matching, deeper lin-

[1] authors GJ and AS in alphabetic order

guistic feature integration such as syntax and semantics, and deep learning (DL) models, with the first two regarded as traditional and the last one regarded as advanced due to the recent appearance of DL models for NLP. We further divide each of these three categories into sub-branches, each with a different focus. Of course, this classification does not have clear boundaries. For instance some automated metrics are involved in both n-gram word surface similarity and linguistic features. This paper differs from the existing works (Dorr et al., 2009; EuroMatrix, 2007) by introducing recent developments in MT evaluation measures, the different classifications from manual to automatic evaluation methodologies, the introduction of more recently developed quality estimation (QE) tasks, and its concise presentation of these concepts.

We hope that our work will shed light and offer a useful guide for both MT researchers and researchers in other relevant NLP disciplines, from the similarity and evaluation point of view, to find useful quality assessment methods, either from the manual or automated perspective, inspired from this work. This might include, for instance, natural language generation (Gehrmann et al., 2021), natural language understanding (Ruder et al., 2021) and automatic summarization (Mani, 2001; Bhandari et al., 2020).

The rest of the paper is organized as follows: Sections 2 and 3 present human assessment and automated assessment methods respectively; Section 4 presents some discussions and perspectives; Section 5 summarizes our conclusions and future work. We also list some further relevant readings in the appendices, such as evaluating methods of TQA itself, MT QE, and mathematical formulas.[2]

2 Human Assessment Methods

In this section we introduce human judgement methods, as reflected in Fig. 1. This categorises these human methods as Traditional and Advanced.

2.1 Traditional Human Assessment

2.1.1 *Intelligibility and Fidelity*

The earliest human assessment methods for MT can be traced back to around 1966. They include the intelligibility and fidelity used by the automatic language processing advisory committee (ALPAC) (Carroll, 1966). The requirement that a translation is intelligible means that, as far as possible, the translation should read like normal, well-edited prose and be readily understandable in the same way that such a sentence would be understandable if originally composed in the translation language. The requirement that a translation is of high fidelity or accuracy includes the requirement that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.

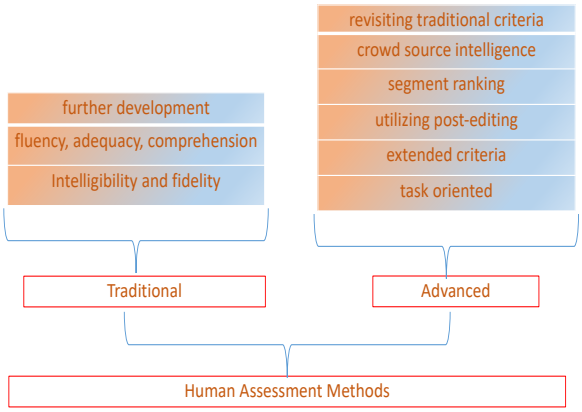

Figure 1: Human Assessment Methods

2.1.2 *Fluency, Adequacy and Comprehension*

In 1990s, the Advanced Research Projects Agency (ARPA) created a methodology to evaluate machine translation systems using the adequacy, fluency and comprehension of the MT output (Church and Hovy, 1991) which adapted in MT evaluation campaigns including (White et al., 1994).

To set upp this methodology, the human assessor is asked to look at each fragment, delimited by syntactic constituents and containing sufficient information, and judge its adequacy on a scale 1-to-5. Results are computed by averaging the judgments over all of the decisions in the translation set.

Fluency evaluation is compiled in the same manner as for the adequacy except that the assessor is to make intuitive judgments on a sentence-by-sentence basis for each translation. Human assessors are asked to determine whether the translation is good English without reference to the correct translation. Fluency evaluation determines whether a sentence is well-formed and fluent in context.

[2]This work is based on an earlier preprint edition (Han, 2016)

Comprehension relates to "Informativeness", whose objective is to measure a system's ability to produce a translation that conveys sufficient information, such that people can gain necessary information from it. The reference set of expert translations is used to create six questions with six possible answers respectively including, "none of above" and "cannot be determined".

2.1.3 *Further Development*

Bangalore et al. (2000) classified accuracy into several categories including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracy. Reeder (2004) found the correlation between fluency and the number of words it takes to distinguish between human translation and MT output.

The "Linguistics Data Consortium (LDC)" [3] designed two five-point scales representing fluency and adequacy for the annual NIST MT evaluation workshop. The developed scales became a widely used methodology when manually evaluating MT by assigning values. The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a translation hypothesis; the second five point scale indicates how fluent the translation is, involving both grammatical correctness and idiomatic word choices.

Specia et al. (2011) conducted a study of MT adequacy and broke it into four levels, from score 4 to 1: highly adequate, the translation faithfully conveys the content of the input sentence; fairly adequate, where the translation generally conveys the meaning of the input sentence, there are some problems with word order or tense/voice/number, or there are repeated, added or non-translated words; poorly adequate, the content of the input sentence is not adequately conveyed by the translation; and completely inadequate, the content of the input sentence is not conveyed at all by the translation.

2.2 Advanced Human Assessment

2.2.1 *Task-oriented*

White and Taylor (1998) developed a task-oriented evaluation methodology for Japanese-to-English translation to measure MT systems in light of the tasks for which their output might be used. They seek to associate the diagnostic scores assigned to the output used in the DARPA (Defense Advanced Research Projects Agency) [4] evaluation with a scale of language-dependent tasks, such as scanning, sorting, and topic identification. They develop an MT proficiency metric with a corpus of multiple variants which are usable as a set of controlled samples for user judgments. The principal steps include identifying the user-performed text-handling tasks, discovering the order of text-handling task tolerance, analyzing the linguistic and non-linguistic translation problems in the corpus used in determining task tolerance, and developing a set of source language patterns which correspond to diagnostic target phenomena. A brief introduction to task-based MT evaluation work was shown in their later work (Doyon et al., 1999).

Voss and Tate (2006) introduced tasked-based MT output evaluation by the extraction of *who*, *when*, and *where* three types of elements. They extended their work later into event understanding (Laoudi et al., 2006).

2.2.2 *Extended Criteria*

King et al. (2003) extend a large range of manual evaluation methods for MT systems which, in addition to the earlir mentioned accuracy, include *suitability*, whether even accurate results are suitable in the particular context in which the system is to be used; *interoperability*, whether with other software or hardware platforms; *reliability*, i.e., don't break down all the time or take a long time to get running again after breaking down; *usability*, easy to get the interfaces, easy to learn and operate, and looks pretty; *efficiency*, when needed, keep up with the flow of dealt documents; *maintainability*, being able to modify the system in order to adapt it to particular users; and *portability*, one version of a system can be replaced by a new version, because MT systems are rarely static and they tend to improve over time as resources grow and bugs are fixed.

2.2.3 *Utilizing Post-editing*

One alternative method to assess MT quality is to compare the post-edited correct translation to the original MT output. This type of evaluation is, however, time consuming and depends on the skills of the human assessor and post-editing performer. One example of a metric that is designed in such a manner is the human translation error rate (HTER) (Snover et al., 2006). This is based on

[3] https://www.ldc.upenn.edu

[4] https://www.darpa.mil

the number of editing steps, computing the editing steps between an automatic translation and a reference translation. Here, a human assessor has to find the minimum number of insertions, deletions, substitutions, and shifts to convert the system output into an acceptable translation. HTER is defined as the sum of the number of editing steps divided by the number of words in the acceptable translation.

2.2.4 *Segment Ranking*

In the WMT metrics task, human assessment based on segment ranking was often used. Human assessors were frequently asked to provide a complete ranking over all the candidate translations of the same source segment (Callison-Burch et al., 2011, 2012). In the WMT13 shared-tasks (Bojar et al., 2013), five systems were randomised for the assessor to give a rank. Each time, the source segment and the reference translation were presented together with the candidate translations from the five systems. The assessors ranked the systems from 1 to 5, allowing tied scores. For each ranking, there was the potential to provide as many as 10 pairwise results if there were no ties. The collected pairwise rankings were then used to assign a corresponding score to each participating system to reflect the quality of the automatic translations. The assigned scores could also be used to reflect how frequently a system was judged to be better or worse than other systems when they were compared on the same source segment, according to the following formula:

$$\frac{\#\text{better pairwise ranking}}{\#\text{total pairwise comparison} - \#\text{ties comparisons}} \quad (1)$$

2.2.5 *Crowd Source Intelligence*

With the reported very low human inter-agreement scores from the WMT segment ranking task, researchers started to address this issue by exploring new human assessment methods, as well as seeking reliable automatic metrics for segment level ranking (Graham et al., 2015).

Graham et al. (2013) noted that the lower agreements from WMT human assessment might be caused partially by the interval-level scales set up for the human assessor to choose regarding quality judgement of each segment. For instance, the human assessor possibly corresponds to the situation where neither of the two categories they were forced to choose is preferred. In light of this rationale, they proposed continuous measurement scales (CMS) for human TQA using fluency criteria. This was implemented by introducing the crowdsource platform Amazon MTurk, with some quality control methods such as the insertion of *bad-reference* and *ask-again*, and statistical significance testing. This methodology reported improved both intra-annotator and inter-annotator consistency. Detailed quality control methodologies, including statistical significance testing were documented in direct assessment (DA) (Graham et al., 2016, 2020).

2.2.6 *Revisiting Traditional Criteria*

Popović (2020a) criticized the traditional human TQA methods because they fail to reflect real problems in translation by assigning scores and ranking several candidates from the same source. Instead, Popović (2020a) designed a new methodology by asking human assessors to mark all problematic parts of candidate translations, either words, phrases, or sentences. Two questions that were typically asked of the assessors related to *comprehensibility* and *adequacy*. The first criteria considered whether the translation is understandable, or understandable but with errors; the second criteria measures if the candidate translation has different meaning to the original text, or maintains the meaning but with errors. Both criteria take into account whether parts of the original text are missing in translation. Under a similar experimental setup, Popović (2020b) also summarized the most frequent error types that the annotators recognized as misleading translations.

3 Automated Assessment Methods

Manual evaluation suffers some disadvantages such as that it is time-consuming, expensive, not tune-able, and not reproducible. Due to these aspects, automatic evaluation metrics have been widely used for MT. Typically, these compare the output of MT systems against human reference translations, but there are also some metrics that do not use reference translations. There are usually two ways to offer the human reference translation, either offering one single reference or offering multiple references for a single source sentence (Lin and Och, 2004; Han et al., 2012).

Automated metrics often measure the overlap in words and word sequences, as well as word order and edit distance. We classify these kinds of

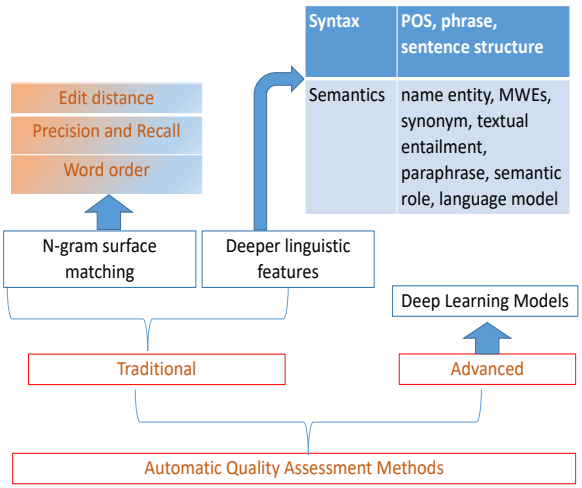

Figure 2: Automatic Quality Assessment Methods

metrics as "simple n-gram word surface matching". Further developed metrics also take linguistic features into account such as syntax and semantics, including POS, sentence structure, textual entailment, paraphrase, synonyms, named entities, multi-word expressions (MWEs), semantic roles and language models. We classify these metrics that utilize the linguistic features as "Deeper Linguistic Features (aware)". This classification is only for easier understanding and better organization of the content. It is not easy to separate these two categories clearly since sometimes they merge with each other. For instance, some metrics from the first category might also use certain linguistic features. Furthermore, we will introduce some recent models that apply deep learning into the TQA framework, as in Fig. 2. Due to space limitations, we present MT quality estimation (QE) task which does not rely on reference translations during the automated computing procedure in the appendices.

3.1 N-gram Word Surface Matching

3.1.1 *Levenshtein Distance*

By calculating the minimum number of editing steps to transform MT output to reference, Su et al. (1992) introduced the word error rate (WER) metric into MT evaluation. This metric, inspired by Levenshtein Distance (or edit distance), takes word order into account, and the operations include insertion (adding word), deletion (dropping word) and replacement (or substitution, replace one word with another), the minimum number of editing steps needed to match two sequences.

One of the weak points of the WER metric is the fact that word ordering is not taken into account appropriately. The WER scores are very low when the word order of system output translation is "wrong" according to the reference. In the Levenshtein distance, the mismatches in word order require the deletion and re-insertion of the misplaced words. However, due to the diversity of language expressions, some so-called "wrong" order sentences by WER also prove to be good translations. To address this problem, the position-independent word error rate (PER) introduced by Tillmann et al. (1997) is designed to ignore word order when matching output and reference. Without taking into account of the word order, PER counts the number of times that identical words appear in both sentences. Depending on whether the translated sentence is longer or shorter than the reference translation, the rest of the words are either insertion or deletion ones.

Another way to overcome the unconscionable penalty on word order in the Levenshtein distance is adding a novel editing step that allows the movement of word sequences from one part of the output to another. This is something a human post-editor would do with the cut-and-paste function of a word processor. In this light, Snover et al. (2006) designed the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step. The shift option is performed on a contiguous sequence of words within the output sentence. For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution.

3.1.2 *Precision and Recall*

The widely used evaluation BLEU metric (Papineni et al., 2002) is based on the degree of n-gram overlap between the strings of words produced by the MT output and the human translation references at the corpus level. BLEU calculates precision scores with n-grams sized from 1-to-4, together multiplied by the coefficient of brevity penalty (BP). If there are multi-references for each candidate sentence, then the nearest length as compared to the candidate sentence is selected as the effective one. In the BLEU metric, the n-gram precision weight λ_n is usually selected as a uniform weight. However, the 4-gram precision value can be very low or even zero when the test corpus is small. To weight more heavily those n-grams that are more informative, Doddington (2002) pro-

poses the NIST metric with the information weight added. Furthermore, Doddington (2002) replaces the geometric mean of co-occurrences with the arithmetic average of n-gram counts, extends the n-gram into 5-gram ($N = 5$), and selects the average length of reference translations instead of the nearest length.

ROUGE (Lin and Hovy, 2003) is a recall-oriented evaluation metric, which was initially developed for summaries, and inspired by BLEU and NIST. ROUGE has also been applied in automated TQA in later work (Lin and Och, 2004).

The F-measure is the combination of precision (P) and recall (R), which was firstly employed in information retrieval (IR) and latterly adopted by the information extraction (IE) community, MT evaluations, and others. Turian et al. (2006) carried out experiments to examine how standard measures such as precision, recall and F-measure can be applied to TQA and showed the comparisons of these standard measures with some alternative evaluation methodologies.

Banerjee and Lavie (2005) designed METEOR as a novel evaluation metric. METEOR is based on the general concept of flexible unigram matching, precision and recall, including the match between words that are simple morphological variants of each other with identical word stems and words that are synonyms of each other. To measure how well-ordered the matched words in the candidate translation are in relation to the human reference, METEOR introduces a penalty coefficient, different to what is done in BLEU, by employing the number of matched chunks.

3.1.3 *Revisiting Word Order*

The right word order plays an important role to ensure a high quality translation output. However, language diversity also allows different appearances or structures of a sentence. How to successfully achieve a penalty on really wrong word order, i.e. wrongly structured sentences, instead of on "correct" different order, i.e. the candidate sentence that has different word order to the reference, but is well structured, has attracted a lot of interest from researchers. In fact, the Levenshtein distance (*Section 3.1.1*) and n-gram based measures also contain word order information.

Featuring the explicit assessment of word order and word choice, Wong and yu Kit (2009) developed the evaluation metric ATEC (assessment of text essential characteristics). This is also based on precision and recall criteria, but with a position difference penalty coefficient attached. The word choice is assessed by matching word forms at various linguistic levels, including surface form, stem, sound and sense, and further by weighing the informativeness of each word.

Partially inspired by this, our work LEPOR (Han et al., 2012) is designed as a combination of augmented evaluation factors including n-gram based *word order penalty* in addition to *precision*, *recall*, and *enhanced sentence-length penalty*. The LEPOR metric (including *h*LEPOR) was reported with top performance on the English-to-other (Spanish, German, French, Czech and Russian) language pairs in ACL-WMT13 metrics shared tasks for *system level* evaluation (Han et al., 2013d). The n-gram based variant *n*LEPOR (Han et al., 2014) was also analysed by MT researchers as one of the three best performing *segment level* automated metrics (together with METEOR and sentBLEU-MOSES) that correlated with human judgement at a level that was not significantly outperformed by any other metrics, on Spanish-to-English, in addition to an aggregated set of overall tested language pairs (Graham et al., 2015).

3.2 Deeper Linguistic Features

Although some of the previously outlined metrics incorporate linguistic information, e.g. synonyms and stemming in METEOR and part of speech (POS) in LEPOR, the simple n-gram word surface matching methods mainly focus on the exact matches of the surface words in the output translation. The advantages of the metrics based on the first category (simple n-gram word matching) are that they perform well in capturing translation fluency (Lo et al., 2012), are very fast to compute and have low cost. On the other hand, there are also some weaknesses, for instance, syntactic information is rarely considered and the underlying assumption that a good translation is one that shares the same word surface lexical choices as the reference translations is not justified semantically. Word surface lexical similarity does not adequately reflect similarity in meaning. Translation evaluation metrics that reflect meaning similarly need to be based on similarity of semantic structure and not merely flat lexical similarity.

3.2.1 *Syntactic Similarity*

Syntactic similarity methods usually employ the features of morphological POS information,

phrase categories, phrase decompositionality or sentence structure generated by linguistic tools such as a language parser or chunker.

In grammar, a **POS** is a linguistic category of words or lexical items, which is generally defined by the syntactic or morphological behaviour of the lexical item. Common linguistic categories of lexical items include noun, verb, adjective, adverb, and preposition. To reflect the syntactic quality of automatically translated sentences, researchers employ POS information into their evaluations. Using the IBM model one, Popović et al. (2011) evaluate translation quality by calculating the similarity scores of source and target (translated) sentences without using a reference translation, based on the morphemes, 4-gram POS and lexicon probabilities. Dahlmeier et al. (2011) developed the TESLA evaluation metrics, combining the synonyms of bilingual phrase tables and POS information in the matching task. Other similar work using POS information include (Giménez and Márquez, 2007; Popovic and Ney, 2007; Han et al., 2014).

In linguistics, a **phrase** may refer to any group of words that form a constituent, and so functions as a single unit in the syntax of a sentence. To measure an MT system's performance in translating new text types, such as in what ways the system itself could be extended to deal with new text types, Povlsen et al. (1998) carried out work focusing on the study of an English-to-Danish MT system. The syntactic constructions are explored with more complex linguistic knowledge, such as the identification of fronted adverbial subordinate clauses and prepositional phrases. Assuming that similar grammatical structures should occur in both source and translations, Avramidis et al. (2011) perform evaluation on source (German) and target (English) sentences employing the features of sentence length ratio, unknown words, phrase numbers including noun phrase, verb phrase and prepositional phrase. Other similar work using phrase similarity includes (Li et al., 2012) that uses noun phrases and verb phrases from chunking, (Echizen-ya and Araki, 2010) that only uses the noun phrase chunking in automatic evaluation, and (Han et al., 2013c) that designs a universal phrase tagset for French to English MT evaluation.

Syntax is the study of the principles and processes by which sentences are constructed in particular languages. To address the overall goodness of a translated **sentence's structure**, Liu and Gildea (2005) employ constituent labels and head-modifier dependencies from a language parser as syntactic features for MT evaluation. They compute the similarity of dependency trees. Their experiments show that adding syntactic information can improve evaluation performance, especially for predicting the fluency of translation hypotheses. Other works that use syntactic information in evaluation include (Lo and Wu, 2011a) and (Lo et al., 2012) that use an automatic shallow parser and the RED metric (Yu et al., 2014) that applies dependency trees.

3.2.2 *Semantic Similarity*

As a contrast to syntactic information, which captures overall grammaticality or sentence structure similarity, the semantic similarity of automatic translations and the source sentences (or references) can be measured by employing semantic features.

To capture the semantic equivalence of sentences or text fragments, **named entity** knowledge is taken from the literature on named-entity recognition, which aims to identify and classify atomic elements in a text into different entity categories (Marsh and Perzanowski, 1998; Guo et al., 2009). The most commonly used entity categories include the names of persons, locations, organizations and time (Han et al., 2013a). In the MEDAR2011 evaluation campaign, one baseline system based on Moses (Koehn et al., 2007) utilized an Open NLP toolkit to perform named entity detection, in addition to other packages. The low performances from the perspective of named entities causes a drop in fluency and adequacy. In the quality estimation of the MT task in WMT 2012, (Buck, 2012) introduced features including named entity, in addition to discriminative word lexicon, neural networks, back off behavior (Raybaud et al., 2011) and edit distance. Experiments on individual features showed that, from the perspective of the increasing the correlation score with human judgments, the named entity feature contributed the most to the overall performance, in comparisons to the impacts of other features.

Multi-word Expressions (MWEs) set obstacles for MT models due to their complexity in presentation as well as idiomaticity (Sag et al., 2002; Han et al., 2020b,a; Han et al., 2021). To investigate the effect of MWEs in MT evaluation (MTE),

Salehi et al. (2015) focused on the *compositionality* of noun compounds. They identify the **noun compounds** first from the system outputs and reference with Stanford parser. The matching scores of the system outputs and reference sentences are then recalculated, adding up to the Tesla metric, by considering the predicated compositionality of identified noun compound phrases. Our own recent work in this area (Han et al., 2020a) provides an extensive investigation into various MT errors caused by MWEs.

Synonyms are words with the same or close meanings. One of the most widely used synonym databases in the NLP literature is WordNet (Miller et al., 1990), which is an English lexical database grouping English words into sets of synonyms. WordNet classifies words mainly into four kinds of POS categories; Noun, Verb, Adjective, and Adverb, without prepositions, determiners, etc. Synonymous words or phrases are organized using the unit of synsets. Each synset is a hierarchical structure with the words at different levels according to their semantic relations.

Textual entailment is usually used as a directive relation between text fragments. If the truth of one text fragment TA follows another text fragment TB, then there is a directional relation between TA and TB (TB $\Rightarrow$ TA). Instead of the pure logical or mathematical entailment, textual entailment in natural language processing (NLP) is usually performed with a relaxed or loose definition (Dagan et al., 2006). For instance, according to text fragment TB, if it can be inferred that the text fragment TA is *most likely* to be true then the relationship TB $\Rightarrow$ TA is also established. Since the relation is directive, it means that the inverse inference (TA $\Rightarrow$ TB) is not ensured to be true (Dagan and Glickman, 2004). Castillo and Estrella (2012) present a new approach for MT evaluation based on the task of "Semantic Textual Similarity". This problem is addressed using a textual entailment engine based on WordNet semantic features.

Paraphrase is to restate the meaning of a passage of text but utilizing other words, which can be seen as bidirectional textual entailment (Androutsopoulos and Malakasiotis, 2010). Instead of the literal translation, word by word and line by line used by meta-phrases, a paraphrase represents a dynamic equivalent. Further knowledge of paraphrases from the aspect of linguistics is introduced in the works by (McKeown, 1979; Meteer and Shaked, 1988; Barzilay and Lee, 2003). Snover et al. (2006) describe a new evaluation metric TER-Plus (TERp). Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in the TERp phrase table.

Semantic roles are employed by researchers as linguistic features in MT evaluation. To utilize semantic roles, sentences are usually first shallow parsed and entity tagged. Then the semantic roles are used to specify the arguments and adjuncts that occur in both the candidate translation and reference translation. For instance, the semantic roles introduced by Gim*é*nez and M*á*rquez (2007); Gim*é*ne and M*á*rquez (2008) include causative agent, adverbial adjunct, directional adjunct, negation marker, and predication adjunct, etc. In a further development, Lo and Wu (2011a,b) presented the MEANT metric designed to capture the predicate-argument relations as structural relations in semantic frames, which are not reflected in the flat semantic role label features in the work of Gim*é*nez and M*á*rquez (2007). Furthermore, instead of using uniform weights, Lo et al. (2012) weight the different types of semantic roles as empirically determined by their relative importance to the adequate preservation of meaning. Generally, semantic roles account for the semantic structure of a segment and have proved effective in assessing adequacy of translation.

Language models are also utilized by MT evaluation researchers. A statistical language model usually assigns a probability to a sequence of words by means of a probability distribution. Gamon et al. (2005) propose the LM-SVM, language model, and support vector machine methods investigating the possibility of evaluating MT quality and fluency in the absence of reference translations. They evaluate the performance of the system when used as a classifier for identifying highly dis-fluent and ill-formed sentences.

Generally, the linguistic features mentioned above, including both syntactic and semantic features, are combined in two ways, either by following a machine learning approach (Albrecht and Hwa, 2007; Leusch and Ney, 2009), or trying to combine a wide variety of metrics in a more simple and straightforward way, such as (Gim*é*ne and M*á*rquez, 2008; Specia and Gim*é*nez, 2010; Comelles et al., 2012).

3.3 Neural Networks for TQA

We briefly list some works that have applied deep learning and neural networks for TQA which are promising for further exploration. For instance, Guzmán et al. (2015); Guzmn et al. (2017) use neural networks (NNs) for TQA for pair wise modeling to choose the best hypothetical translation by comparing candidate translations with a reference, integrating syntactic and semantic information into NNs. Gupta et al. (2015b) proposed LSTM networks based on dense vectors to conduct TQA, while Ma et al. (2016) designed a new metric based on bi-directional LSTMs, which is similar to the work of Guzmán et al. (2015) but with less complexity by allowing the evaluation of a single hypothesis with a reference, instead of a pairwise situation.

4 Discussion and Perspectives

In this section, we examine several topics that can be considered for further development of MT evaluation fields.

The first aspect is that development should involve both n-gram word surface matching and the deeper linguistic features. Because natural languages are expressive and ambiguous at different levels (Giménez and Márquez, 2007), simple n-gram word surface similarity based metrics limit their scope to the lexical dimension and are not sufficient to ensure that two sentences convey the same meaning or not. For instance, (Callison-Burch et al., 2006a) and (Koehn and Monz, 2006b) report that simple n-gram matching metrics tend to favor automatic statistical MT systems. If the evaluated systems belong to different types that include rule based, human aided, and statistical systems, then the simple n-gram matching metrics, such as BLEU, give a strong disagreement between these ranking results and those of the human assessors. So deeper linguistic features are very important in the MT evaluation procedure.

However, inappropriate utilization, or abundant or abused utilization, of linguistic features will result in limited popularity of measures incorporating linguistic features. In the future, how to utilize the linguistic features in a more accurate, flexible and simplified way, will be one challenge in MT evaluation. Furthermore, the MT evaluation from the aspects of semantic similarity is more reasonable and reaches closer to the human judgments, so it should receive more attention.

The second major aspect is that MT quality estimation (QE) tasks are different to traditional MT evaluation in several ways, such as extracting reference-independent features from input sentences and their translation, obtaining quality scores based on models produced from training data, predicting the quality of an unseen translated text at system run-time, filtering out sentences which are not good enough for post processing, and selecting the best translation among multiple systems. This means that with so many challenges, the topic will continuously attract many researchers.

Thirdly, some advanced or challenging technologies that can be further applied to MT evaluation include the deep learning models (Gupta et al., 2015a; Zhang and Zong, 2015), semantic logic form, and decipherment model.

5 Conclusions and Future Work

In this paper we have presented a survey of the state-of-the-art in translation quality assessment methodologies from the viewpoints of both manual judgements and automated methods. This work differs from conventional MT evaluation review work by its concise structure and inclusion of some recently published work and references. Due to space limitations, in the main content, we focused on conventional human assessment methods and automated evaluation metrics with reliance on reference translations. However, we also list some interesting and related work in the appendices, such as the quality estimation in MT when the reference translation is not presented during the estimation, and the evaluating methodology for TQA methods themselves. However, this arrangement does not affect the overall understanding of this paper as a self contained overview. We believe this work can help both MT and NLP researchers and practitioners in identifying appropriate quality assessment methods for their work. We also expect this work might shed some light on evaluation methodologies in other NLP tasks, due to the similarities they share, such as text summarization (Mani, 2001; Bhandari et al., 2020), natural language understanding (Ruder et al., 2021), natural language generation (Gehrmann et al., 2021), as well as programming language (code) generation (Liguori et al., 2021).

Acknowledgments

We appreciate the comments from Derek F. Wong, editing help from Ying Shi (Angela), and the anonymous reviewers for their valuable reviews and feedback. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The input of Alan Smeaton is part-funded by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight Centre).

References

J. Albrecht and R. Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of the 45th Annual Meeting of the ACL, Prague, Czech Republic*.

Jon Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

D. Arnold. 2003. *Computers and Translation: A translator's guide-Chap8 Why translation is difficult for computers*. Benjamins Translation Library.

Eleftherios Avramidis, Maja Popovic, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of WMT 2011*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005*.

Srinivas Bangalore, Owen Rambow, and Steven Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG 2000*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL 2003*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of WMT 2013*.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp,

Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Christian Buck. 2012. Black box features for the wmt 2012 quality estimation shared task. In *Proceedings of WMT 2012*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007a. (meta-) evaluation of machine translation. In *Proceedings of WMT 2007*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007b. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 64–71. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of WMT 2008*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaridan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the WMT 2010*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT 2012*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the 4th WMT 2009*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaridan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT 2011*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL 2006*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL 2006*, volume 2006, pages 249–256.

John B. Carroll. 1966. An experiment in evaluating the quality of translation. *Mechanical Translation and Computational Linguistics*, 9(3-4):67–75.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58, Montréal, Canada. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Kenneth Church and Eduard Hovy. 1991. Good applications for crummy machine translation. In *Proceedings of the Natural Language Processing Systems Evaluation Workshop*.

Jasob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):3746.

Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. 2012. Verta: Linguistic features in mt evaluation. In *LREC*, pages 3944–3950.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining workshop*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Machine Learning Challenges:LNCS*, 3944:177–190.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. Tesla at wmt2011: Translation evaluation and tunable metric. In *Proceedings of WMT 2011*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT Proceedings*.

Bonnie Dorr, Matt Snover, and etc. Nitin Madnani. 2009. Part 5: Machine translation evaluation. In *Bonnie Dorr edited DARPA GALE program report*.

Jennifer B. Doyon, John S. White, and Kathryn B. Taylor. 1999. Task-based evaluation for machine translation. In *Proceedings of MT Summit 7*.

H. Echizen-ya and K. Araki. 2010. Automatic evaluation method for machine translation using noun-phrase chunking. In *Proceedings of the ACL 2010*.

Matthias Eck and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Project EuroMatrix. 2007. 1.3: Survey of machine translation evaluation. In *EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit*.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the iwslt 2011 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations beyond language modelling. In *Proceedings of EAMT*, pages 103–112.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. *arXiv e-prints*, page arXiv:2102.01672.

Jesús Giméne and Lluís Márquez. 2008. A smorgasbord of features for automatic mt evaluation. In *Proceedings of WMT 2008*, pages 195–198.

Jesús Giménez and Lluís Márquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of WMT 2007*.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

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.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1–10.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceeding of SIGIR*.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015a. Machine translation evaluation using recurrent neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 380–384, Lisbon, Portugal. Association for Computational Linguistics.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015b. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Emperical Methods in Natural Language Processing*, pages 1066–1072. Association for Computational Linguistics, o.A.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint*

Conference of the Asian Federation of Natural Language Processing (ACL'15), pages 805–814, Beijing, China. Association for Computational Linguistics.

Francisco Guzmn, Shafiq Joty, Llus Mrquez, and Preslav Nakov. 2017. Machine translation evaluation with neural networks. *Comput. Speech Lang.*, 45(C):180–200.

Anders Hald. 1998. *A History of Mathematical Statistics from 1750 to 1930*. ISBN-10: 0471179124. Wiley-Interscience; 1 edition.

Aaron L-F Han, Derek F Wong, and Lidia S Chao. 2013a. Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Language Processing and Intelligent Information Systems*, pages 57–68. Springer.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau, Macao.

Lifeng Han. 2016. Machine Translation Evaluation Resources and Methods: A Survey. *arXiv e-prints*, page arXiv:1605.04515.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Lifeng Han, Gareth J. F. Jones, Alan F. Smeaton, and Paolo Bolzoni. 2021. Chinese Character Decomposition for Neural MT with Multi-Word Expressions. *arXiv e-prints*, page arXiv:2104.04497.

Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013b. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Lifeng Han, Derek Fai Wong, and Lidia Sam Chao. 2012. A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING*.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, Shuo Li, and Ling Zhu. 2013c. Phrase tagset mapping for french and english treebanks and its application in machine translation evaluation. In *International Conference of the German Society for Computational Linguistics and Language Technology, LNAI Vol. 8105*, pages 119–131.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, and Yi Lu. 2014. Unsupervised quality estimation model for english to german translation and its application in extensive supervised evaluation. In *The Scientific World Journal. Issue: Recent Advances in Information Technology*, pages 1–12.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013d. A description of tunable machine translation evaluation systems in wmt13 metrics task. In *Proceedings of WMT 2013*, pages 414–421.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Maurice G. Kendall and Jean Dickinson Gibbons. 1990. *Rank Correlation Methods*. Oxford University Press, New York.

Marrgaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. Femti: Creating and using a framework for mt evaluation. In *Proceedings of the Machine Translation Summit IX*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

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 Conference on Association of Computational Linguistics*.

Philipp Koehn and Christof Monz. 2006a. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006b. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of WMT 2006*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jamal Laoudi, Ra R. Tate, and Clare R. Voss. 2006. Task-based mt evaluation: From who/when/where extraction to event understanding. In *in Proceedings of LREC-06*, pages 2048–2053.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie. 2013. Automated metrics for mt evaluation. *Machine Translation*, 11:731.

Guy Lebanon and John Lafferty. 2002. Combining rankings using conditional probability models on permutations. In *Proceeding of the ICML*.

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, 23(2-3).

A. LI. 2005. Results of the 2005 nist machine translation evaluation. In *Proceedings of WMT 2005*.

Liang You Li, Zheng Xian Gong, and Guo Dong Zhou. 2012. Phrase-based evaluation for machine translation. In *Proceedings of COLING*, pages 663–672.

Pietro Liguori, Erfan Al-Hossami, Domenico Cotroneo, Roberto Natella, Bojan Cukic, and Samira Shaikh. 2021. Shellcode_IA32: A Dataset for Automatic Shellcode Generation. *arXiv e-prints*, page arXiv:2104.13100.

Chin-Yew Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL 2003*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL 2004*.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedingsof theACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chi Kiu Lo, Anand Karthik Turmuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of WMT 2012*.

Chi Kiu Lo and Dekai Wu. 2011a. Meant: An inexpensive, high- accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of ACL 2011*.

Chi Kiu Lo and Dekai Wu. 2011b. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in StatisticalTranslation (SSST-5)*.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67.

Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. 2016. Maxsd: A neural machine translation evaluation metric optimized by maximizing similarity distance. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, pages 153–161.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

I. Mani. 2001. Summarization evaluation: An overview. In *NTCIR*.

Elaine Marsh and Dennis Perzanowski. 1998. Muc-7 evaluation of ie technology: Overview of results. In *Proceedingsof Message Understanding Conference (MUC-7)*.

Kathleen R. McKeown. 1979. Paraphrasing using given and new information in a question-answer system. In *Proceedings of ACL 1979*.

Marie Meteer and Varda Shaked. 1988. Microsoft research treelet translation system: Naacl 2006 europarl evaluation. In *Proceedings of COLING*.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Douglas C. Montgomery and George C. Runger. 2003. *Applied statistics and probability for engineers*, third edition. John Wiley and Sons, New York.

Erwan Moreau and Carl Vogel. 2013. Weakly supervised approaches for quality estimation. *Machine Translation*, 27(3–4):257–280.

Erwan Moreau and Carl Vogel. 2014. Limitations of MT quality estimation supervised systems: The tails prediction problem. In *Proceedings of COLING*

2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2205–2216, Dublin, Ireland. Dublin City University and 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 ACL 2002*.

M. Paul. 2009. Overview of the iwslt 2009 evaluation campaign. In *Proceeding of IWSLT*.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *Proceeding of IWSLT*.

Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(5):157–175.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: Ibm1 scores as evaluation metrics. In *Proceedings of WMT 2011*.

M. Popovic and Hermann Ney. 2007. Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of WMT 2007*.

Maja Popović. 2020a. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović. 2020b. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online. Association for Computational Linguistics.

Claus Povlsen, Nancy Underwood, Bradley Music, and Anne Neville. 1998. Evaluating text-type suitability for machine translation a case study on an english-danish system. In *Proceeding LREC*.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Florence Reeder. 2004. Investigation of intelligibility judgments. In *Machine Translation: From Real Users to Research*, pages 227–235, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. *arXiv e-prints*, page arXiv:2104.07412.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado. Association for Computational Linguistics.

Mattthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceeding of AMTA*.

L. Specia and J. Giménez. 2010. Combining confidence estimation and reference-based metrics for segment-level mt evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Naheh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Keh-Yih Su, Wu Ming-Wen, and Chang Jing-Shin. 1992. A new quantitative quality measure for machine translation systems. In *Proceeding of COLING*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceeding of EUROSPEECH*.

Joseph P Turian, Luke Shea, and I Dan Melamed. 2006. Evaluation of machine translation and its evaluation. Technical report, DTIC Document.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Clare R. Voss and Ra R. Tate. 2006. Task-based evaluation of machine translation (mt) engines: Measuring how well people extract who, when, where-type elements in mt output. In *In Proceedings of 11th Annual Conference of the European Association for Machine Translation (EAMT-2006*, pages 203–212.

Warren Weaver. 1955. Translation. *Machine Translation of Languages: Fourteen Essays*.

John S. White, Theresa O' Connell, and Francis O' Mara. 1994. The arpa mt evaluation methodologies: Evolution, lessons, and future approaches. In *Proceeding of AMTA*.

John S. White and Kathryn B. Taylor. 1998. A task-oriented evaluation metric for machine translation. In *Proceeding LREC*.

Billy Wong and Chun yu Kit. 2009. Atec: automatic evaluation of machine translation via word choice and word order. *Machine Translation*, 23(2-3):141–155.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. RED: A reference dependency based MT evaluation metric. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2042–2051.

Jiajun Zhang and Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, (5):16–25.

Appendices

Appendix A: Evaluating TQA

A.1: Statistical Significance

If different MT systems produce translations with different qualities on a dataset, how can we ensure that they indeed own different system quality? To explore this problem, Koehn (2004) presents an investigation statistical significance testing for MT evaluation. The bootstrap re-sampling method is used to compute the statistical significance intervals for evaluation metrics on small test sets. Statistical significance usually refers to two separate notions, one of which is the p-value, the probability that the observed data will occur by chance in a given single null hypothesis. The other one is the "Type I" error rate of a statistical hypothesis test, which is also called "false positive" and measured by the probability of incorrectly rejecting a given null hypothesis in favour of a second alternative hypothesis (Hald, 1998).

A.2: Evaluating Human Judgment

Since human judgments are usually trusted as the gold standards that automatic MT evaluation metrics should try to approach, the reliability and coherence of human judgments is very important. Cohen's kappa agreement coefficient is one of the most commonly used evaluation methods (Cohen, 1960). For the problem of nominal scale agreement between two judges, there are two relevant quantities p_0 and p_c. p_0 is the proportion of units in which the judges agreed and p_c is the proportion of units for which agreement is expected by chance. The coefficient k is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{2}$$

where $p_0 - p_c$ represents the proportion of the cases in which beyond-chance agreement occurs and is the numerator of the coefficient (Landis and Koch, 1977).

A.3: Correlating Manual and Automatic Score

In this section, we introduce three correlation coefficient algorithms that have been widely used at recent WMT workshops to measure the closeness of automatic evaluation and manual judgments. The choice of correlation algorithm depends on whether scores or ranks schemes are utilized.

Pearson Correlation

Pearson's correlation coefficient (Pearson, 1900) is commonly represented by the Greek letter ρ. The correlation between random variables X and

Y denoted as ρ_{XY} is measured as follows (Montgomery and Runger, 2003).

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (3)$$

Because the standard deviations of variable X and Y are higher than 0 ($\sigma_X > 0$ and $\sigma_Y > 0$), if the covariance σ_{XY} between X and Y is positive, negative or zero, the correlation score between X and Y will correspondingly result in positive, negative or zero, respectively. Based on a sample of paired data (X,Y) as $(x_i, y_i), i = 1\,to\,n$, the Pearson correlation coefficient is calculated as:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \quad (4)$$

where μ_x and μ_y specify the means of discrete random variable X and Y respectively.

Spearman rank Correlation

Spearman rank correlation coefficient, a simplified version of Pearson correlation coefficient, is another algorithm to measure the correlations of automatic evaluation and manual judges, e.g. in WMT metrics task (Callison-Burch et al., 2008, 2009, 2010, 2011). When there are no ties, Spearman rank correlation coefficient, which is sometimes specified as (rs) is calculated as:

$$rs_{\varphi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (5)$$

where d_i is the difference-value (D-value) between the two corresponding rank variables $(x_i - y_i)$ in $\vec{X} = \{x_1, x_2, ..., x_n\}$ and $\vec{Y} = \{y_1, y_2, ..., y_n\}$ describing the system φ.

Kendall's τ

Kendall's τ (Kendall, 1938) has been used in recent years for the correlation between automatic order and reference order (Callison-Burch et al., 2010, 2011, 2012). It is defined as:

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \quad (6)$$

The latest version of Kendall's τ is introduced in (Kendall and Gibbons, 1990). Lebanon and Lafferty (2002) give an overview work for Kendall's τ showing its application in calculating how much the system orders differ from the reference order. More concretely, Lapata (2003) proposed the use of Kendall's τ, a measure of rank correlation, to estimate the distance between a system-generated and a human-generated gold-standard order.

A.4: Metrics Comparison

There are researchers who did some work about the comparisons of different types of metrics. For example, Callison-Burch et al. (2006b, 2007b); Lavie (2013) mentioned that, through some qualitative analysis on some standard data set, BLEU cannot reflect MT system performance well in many situations, i.e. higher BLEU score cannot ensure better translation outputs. There are some recently developed metrics that can perform much better than the traditional ones especially on challenging sentence-level evaluation, though they are not popular yet such as nLEPOR and SentBLEU-Moses (Graham et al., 2015; Graham and Liu, 2016). Such comparison will help MT researchers to select th appropriate metrics to use for specialist tasks.

Appendix B: MT QE

In past years, some MT evaluation methods that do not use manually created gold reference translations were proposed. These are referred to as "Quality Estimation (QE)". Some of the related works have already been introduced in previous sections. The most recent quality estimation tasks can be found at WMT12 to WMT20 (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015; Specia et al., 2018; Fonseca et al., 2019; Specia et al., 2020). These defined a novel evaluation metric that provides some advantages over the traditional ranking metrics. The DeltaAvg metric assumes that the reference test set has a number associated with each entry that represents its extrinsic value. Given these values, their metric does not need an explicit reference ranking, the way that Spearman ranking correlation does. The goal of the DeltaAvg metric is to measure how valuable a proposed ranking is according to the extrinsic values associated with the test entries.

$$DeltaAvg_v[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S) \quad (7)$$

For scoring, two task evaluation metrics were used that have traditionally been used for measur-

ing performance in regression tasks: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric. For a given test set S with entries $s_i, 1 \leqslant i \leqslant |S|$, $H(s_i)$ is the proposed score for entry s_i (hypothesis), and $V(s_i)$ is the reference value for entry s_i (gold-standard value).

$$\text{MAE} = \frac{\sum_{i-1}^{N}|H(s_i) - V(s_i)|}{N} \quad (8)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i-1}^{N}(H(s_i) - V(s_i))^2}{N}} \quad (9)$$

where $N = |S|$. Both these metrics are non-parametric, automatic and deterministic (and therefore consistent), and extrinsically interpretable.

Some further readings on MT QE are the comparison between MT evaluation and QE Specia et al. (2010) and the QE framework model QuEst (Specia et al., 2013); the weakly supervised approaches for quality estimation and the limitations analysis of QE Supervised Systems (Moreau and Vogel, 2013, 2014), and unsupervised QE models (Fomicheva et al., 2020); the recent shared tasks on QE (Fonseca et al., 2019; Specia et al., 2020).

In very recent years, the two shared tasks, i.e. MT quality estimation and traditional MT evaluation metrics, have tried to integrate into each other and benefit from both knowledge. For instance, in WMT2019 shared task, there were 10 reference-less evaluation metrics which were used for the QE task, "QE as a Metric", as well (Ma et al., 2019).

Appendix C: Mathematical Formulas

Some mathematical formulas that are related to aforementioned metrics:

Section 2.1.2 - Fluency / Adequacy / Comprehension:

$$\text{Comprehension} = \frac{\#\text{Cottect}}{6} \quad (10)$$

$$\text{Fluency} = \frac{\frac{\text{Judgment point}_{-1}}{\text{S}_{-1}}}{\#\text{Sentences in passage}} \quad (11)$$

$$\text{Adequacy} = \frac{\frac{\text{Judgment point}_{-1}}{\text{S}_{-1}}}{\#\text{Fragments in passage}} \quad (12)$$

Section 3.1.1 - Editing Distance:

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}}. \quad (13)$$

$$\text{PER} = 1 - \frac{\text{correct} - \max(0, \text{output}_{\text{length}} - \text{reference}_{\text{length}})}{\text{reference}_{\text{length}}}. \quad (14)$$

$$\text{TER} = \frac{\#\text{of edit}}{\#\text{of average reference words}} \quad (15)$$

Section 3.1.2 - Precision and Recall:

$$\text{BLEU} = \text{BP} \times \exp\sum_{n=1}^{N}\lambda_n \log \text{Precision}_{\text{n}}, \quad (16)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{1-\frac{r}{c}} & \text{if } c <= r. \end{cases} \quad (17)$$

where c is the total length of candidate translation, and r refers to the sum of effective reference sentence length in the corpus. Bellow is from NIST metric, then F-measure, METEOR and LEPOR:

$$\text{Info} = \log_2(\frac{\#\text{occurrence of } w_1, \cdots, w_{n-1}}{\#\text{occurrence of } w_1, \cdots, w_n}) \quad (18)$$

$$F_\beta = (1+\beta^2)\frac{PR}{R+\beta^2 P} \quad (19)$$

$$\text{Penalty} = LP0.5 \times (\frac{\#\text{chunks}}{\#\text{matched unigrams}})^3 \quad (20)$$

$$\text{MEREOR} = \frac{10PR}{R+9P} \times (1 - \text{Penalty}) \quad (21)$$

$$\text{LEPOR} = LP \times NPosPenal \times Harmonic(\alpha R, \beta P) \quad (22)$$

$$h\text{LEPOR} = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR)$$

$$n\text{LEPOR} = LP \times NPosPenal \times exp(\sum_{n=1}^{N} w_n logHPR)$$

where, in our own metric LEPOR and its variations, *n*LEPOR (*n*-gram *precision* and *recall* LEPOR) and *h*LEPOR (*harmonic* LEPOR), *P* and *R* are for precision and recall, *LP* for length penalty, *NPosPenal* for n-gram position difference penalty, and *HPR* for harmonic mean of precision and recall, respectively (Han et al., 2012, 2013b; Han, 2014; Han et al., 2014).

Linguistic profiles of translation manuscripts and edited translations

Tatiana Serbina
RWTH Aachen University
Kármánstraße 17-19
52062 Aachen, Germany
tatiana.serbina@ifaar.rwth-aachen.de

Mario Bisiada
Universitat Pompeu Fabra
C. Roc Boronat, 138
08018 Barcelona, Spain
mario.bisiada@upf.edu

Stella Neumann
RWTH Aachen University
Kármánstraße 17-19
52062 Aachen, Germany
stella.neumann@ifaar.rwth-aachen.de

Abstract

A range of studies have pointed to the importance of considering the influence of editors in studies of translated language. Those studies have concentrated on particular features, which allowed them to study those features in detail, but also prevented them from providing an overall picture of the linguistic properties of the texts in question. This study addresses this issue by conducting a multivariate analysis of unedited and edited translations of English business articles into German. We aim to investigate whether translation manuscripts have a characteristically different distribution of lexico-grammatical features compared to edited translations, and whether editors normalize those features and thus assimilate the translations to non-translated texts. Findings related to individual features are in line with the previously observed phenomena of sentence splitting and passive voice, and a general tendency towards increasing readability. In general, however, no profound effect of editorial intervention could be observed, even though there was a slight tendency of edited translations to be more similar to comparable originals.

1 Introduction

The aim of this study is to assess the role of editors in the translation workflow. This is achieved using the geometric multivariate analysis (GMA) proposed by Diwersy et al. (2014) and Evert and Neumann (2017) to obtain a holistic account of the linguistic properties that characterize translation manuscripts and edited translations. Our pilot study focuses on the first two steps of GMA, namely performing a Principal Component Analysis (PCA) and visually inspecting its results. Specifically, in this paper we address the following questions:

- Do translation manuscripts have a characteristically different distribution of lexico-grammatical features compared to edited translations?
- Do editors normalize the lexico-grammatical features of translation manuscripts, assimilating them to the comparable non-translated texts?

2 Editorial influence in the translation workflow

A number of recent studies analyzed editors' influence on translated texts raising awareness of the part editors play in the translation workflow (Kruger, 2012, 2017; Bisiada, 2017, 2018a, 2019).

Bisiada (2016) studied the phenomenon of sentence splitting, which is often considered a feature of translation that occurs depending on structural conventions in the target languages (Fabricius-Hansen, 1996, 1999; Solfjeld, 2008). He critiques that there seems to be the assumption of an "automatism that seems to assume that translators have little choice in the matter, as the structural principles of the languages involved determine whether sentences are split" (Bisiada, 2016, 354), so that sentence splitting almost necessarily occurs in translations from languages that are considered to prefer a higher informational density to those with a lower one. It is assumed not to occur in the opposite direction, that is, when translating from "low density" language into "high density" languages, because the latter have the structural resources to present information in a compact way.

In his study, however, Bisiada (2016, 374) observes a notable amount of sentence splitting in translations from English to German, thus providing evidence to suggest that "sentence splitting is

an explicitating strategy in translated language in general rather than a process that is triggered only in specific translation directions". As he finds that a significant amount of sentence splitting is attributable to editors, he argues that explicitation as a translation strategy cannot account for the observed frequency of sentence splitting and points to a possible attempt by editors to increase readability (Bisiada, 2016, 371–374).

This is evidenced by further research into the corpus: a study of nominalizations finds that half of them "consist of extensive changes that lead to a complete reformulation of the sentence in question", so that "translators may thus be affected to a greater extent by the academic nature of the source texts, which conventionally favours a nominal style in German, while the editors in this case incorporate popularising strategies" (Bisiada, 2018a, 46–47).

Those findings are corroborated by a study of grammatical metaphorization on the same corpus, which finds that the main influence editors exert is that of turning nominal constructions in the translation back into verbal ones (Bisiada, 2018b,c) as well as turning passive constructions back to active ones (Bisiada, 2019). Both studies suggest that it is through editorial influence that the published translation receives its notable usage frequency of nominal and passive forms.

Kruger (2017) reports on an ongoing study of 208 English non-translated texts, in both unedited and edited form, from the registers "academic, instructional, popular writing and reportage" (Kruger, 2017, 125). To study the influence of editors on the text, she uses a range of operationalizations as proposed by Kruger (2012); Kruger and van Rooy (2012). Her findings are that editors "prefer explicit, non-redundant, analytical constructions, which also tend to be associated with formal writing", most evidently so "in the popular register, where editors' conventionalising impulses override the register preference for more informal usage" (Kruger, 2017, 146). She further reports "support for the hypothesis that editors demonstrate a tendency towards conventionalization or normalization", though they "reduce conventional lexical patterning in the most-frequent range of trigrams" (Kruger, 2017, 146). The study also supports the view that editors simplify the texts.

Bisiada (2017) has further pursued this idea by studying how translation and editing are different activities as regards explicitation, normalization and simplification. The aim of the study was to address the claim that translation universals are really "mediation universals" (Chesterman, 2004; Ulrych and Murphy, 2008) and that editing and translating are thus comparable linguistic activities. This notion was contested by Kruger, who finds a "consistent difference between the translated and edited subcorpus" (Kruger, 2012, 380) in her data.

Bisiada (2017, 268) finds two significant differences: one is between (manuscript and edited) translations and (edited) non-translated texts in the "universal" of normalization/conservatism, the second is that, in terms of simplification, manuscript translations differ from edited texts (translations and non-translations). Bisiada (2017, 268) argues that "editors' influence has been strongest in this respect" and suggests that this may be because simplification is operationalized mainly by quantitative features, which also attract "speed editing" (Bisaillon, 2007, 306).

In terms of a comparison to Kruger (2017), the editorial tendencies towards simplification is corroborated, but Bisiada (2017) finds no reduction of conventionalized lexical patterns in the form of trigrams in translated German; the translations are more conventional than non-translated texts, both before and after editing. This, however, may be due to language differences, corpus composition and also the fact that Kruger (2017) studied non-translations, i.e. texts written originally in the analyzed language, while Bisiada (2017) examined translations.

Bisiada (2017) concludes that the editing stage seems to have had little effect on the features he measured, but states that this "does not mean that changes to the text are negligible, but rather that editors do not intervene in such a way to make the articles more like the non-translated articles" (Bisiada, 2017, 269). This points to the main limitation of research into linguistic properties based on specific features: even if the study takes into account a wide range of them, the picture provided by the results is often fragmented. Observed results are usually interpreted in terms of the specific feature that the analysis concentrated on, which hinders a holistic analysis. This is why we believe that a multivariate analysis provides a full and equal picture to study the lexico-grammatical

features of texts.

3 Methodology

While the above studies have picked a range of individual features for analysis, the present study adopts the multivariate methodology as proposed by Evert and Neumann (2017), whose aim is to study systematic properties of text which, they argue, are not observable on the basis of individual features: "the use of multivariate techniques appears to be essential for a systematic investigation of translation properties" (Evert and Neumann, 2017, 48). The present study therefore runs such a multivariate analysis technique on the corpus compiled by Bisiada (2018a,b,c) (hereafter: Harvard Business Corpus), which was updated by also including text in boxes appearing next to the main articles. The Harvard Business Corpus consists of articles published in the *Harvard Business Manager*, a German sister publication to the *Harvard Business Review*. The articles are translations of articles published in the American edition. The corpus also contains translation manuscripts, which we define as translated texts that were sent by the translation company to the publisher. At least nine different translators have translated the texts at the translation company (in some cases the translator's details were not specified), and six different editors have worked on the texts at the *Harvard Business Manager*.

The articles present findings of scientific studies in an accessible form, geared to managers and business leaders, and thus resemble what is elsewhere known as a popular-scientific format. Others give advice on how to become a better leader or how to manage a company or staff. The magazine sends out specific instructions to its translators where the editors ask them to avoid the nominal style, jargons, the passive and impersonal language use. They are also instructed to dissolve nested sentences (Bisiada, 2016, 356). As these are instructions given to translators, it seems plausible to assume that editors will work with them to hand and use them as their editorial guidelines.

For the present study, this collection of translation manuscripts and edited translations was complemented by a part of the CroCo corpus (Hansen-Schirra et al., 2012). More specifically, in addition to the German translations (edited and non-edited) of business articles (BUSINESS), our data sample includes the published German translations belonging to the registers of letters to shareholders (SHARE) and popular-scientific texts (POPSCI), as well as the German originals from the same registers. Moreover, to counterbalance the size of the data sample consisting of German originals, two additional registers were added, namely the registers of political essays (ESSAY) and prepared speeches (SPEECH). The texts from SHARE and POPSCI were added due to their similarity to the BUSINESS register: letters to the shareholders refer to the performance of the company and the actions of the management, their aim being both to inform and to convince the shareholders. Similar to the business articles, the German translations from POPSCI are mostly articles published in the popular-scientific magazines. Unfortunately, due to the difficulties of finding comparable translations in the opposite translation direction, the sub-corpus of German originals contains popular-scientific book extracts. The aim is to present the scientific findings to the readers in a comprehensible way (Neumann and Hansen-Schirra, 2012). Table 1 summarizes the data used for the present study. The entire data sample consists of 137 texts.

The meta data contains four distinct categories for corpora, namely two different translation versions from the Harvard Business Corpus (Trans – translation manuscripts, Publ – published translations) as well as originals and translations from the CroCo corpus (GO – German originals, GTrans – German translations), and five categories for registers, namely BUSINESS, SHARE, POPSCI, SPEECH and ESSAY.

All texts from Harvard Business Corpus were POS tagged with the STTS tagset (Schiller et al., 1999) using the TreeTagger (Schmid, 1994). The texts from the CroCo corpus that we drew on were tagged using the TnT tagger (Brants, 2000). Based on the previous work on GMA (Evert and Neumann, 2017), the study is based on a set of lexico-grammatical features that were originally defined for the study of register variation (Neumann, 2013). We argue that together the features result in a linguistic profile of the analyzed texts. The process of feature extraction and quantification of every feature per text in the data sample was performed with the help of a CQP script (Fest et al., 2019; Neumann and Evert, Forthcoming).

In the next step, the raw frequencies are normalized using the appropriate unit of measurement, such as nominalizations/words or finite verbs/

Corpus	Translation Status	Register	Size in words
Harvard Business Corpus	manuscript translations	Business	112,810
Harvard Business Corpus	published translations	Business	106,958
CroCo	originals	Share, Popsci, Speech, Essay	137,747
CroCo	published translations	Share, Popsci	69,937

Table 1: Overview of the data sample

sentences. The features that were too sparse in the data and features with correlations r higher than 0.7 were removed from further analysis. From each pair of correlated features, the feature which deemed to be linguistically more informative was kept for further analysis. An overview of the remaining 36 features is shown in Appendix A.

Analysis of the data is performed in two steps. First, the feature counts are discussed descriptively to get the first impression of the data distribution in translation manuscripts and edited translations. In the second step, the features are used as an input for PCA – an unsupervised statistical technique that reduces the number of dimensions within the data set (Levshina, 2015).

4 Analysis

Before performing a multivariate analysis of the data, the distribution of individual features is compared descriptively between the two translation versions contained in the Harvard Business Corpus – translation manuscripts and edited translations. Since the data contains a large amount of outliers, the comparison is based on median that is less sensitive to extreme values. An initial analysis of raw counts showed that translation manuscripts are characterized by more words but contain less sentences as well as less verbs in general and finite verbs in particular. Due to the fact that a lot of other variables are dependent on these values, further comparison is performed using normalized values (see Section 3). For the purposes of this comparison, most of the feature counts are represented here as percentages.

While the differences between the normalized counts are very small, some minor contrasts can be detected (see Table 2, which contains only differences above 1 per cent). These are related to the values of coordination/finite verb, past tense/finite verb, passive/finite verb, and PP as theme/sentence – all of which are used more frequently in translation manuscripts – as well as to adverbs as theme/sentence and conjunctions as theme/sentence – which are slightly increased in the edited translations. Moreover, one further minor contrast concerns the feature words/sentence (the median of 18.82 for manuscript translations, 17.39 for edited translations, 19.31 for non-translations). In contrast to the features included in Table 2, the feature words/sentence represents the number of words per sentence, rather than the proportion. Therefore, this feature count was not transformed into percentage. When compared to medians of the non-translations, the values for all of these features, with the exception of coordination/finite verb and words/sentence, are higher in both translation versions (see Appendix B for the corresponding boxplots).

In order to perform PCA based on the analyzed features, some further preliminary data processing steps are required. In accordance with GMA procedure introduced in Diwersy et al. (2014) and Evert and Neumann (2017), visual inspection of plots plays an important role both during data preparation and interpretation of results. Due to different ranges and distributions of features visible in box plots, normalized feature counts are standardized as z-scores. In the next step, to reduce the influence of outliers, we applied the signed logarithmic transformation of z-scores. Visual inspection of the PCA with and without the log-transformation revealed that individual outliers were reduced, while the overall shape of the data stayed similar. Therefore, all further analyses are performed using log-transformed values. In these analyses every text is projected into a multidimensional feature space as a feature vector comprising the log-transformed z-scores of 36 indicators. The Euclidean distances between the feature vectors are assumed to represent meaningful differences between texts in terms of the selected lexico-grammatical features (Evert and Neumann, 2017).

PCA returns a ranked list of latent dimensions

Feature	Manuscript translations	Edited translations	Non-translations
pasttense/S	29.46	27.7	9.56
passive/F	11.13	6.66	7.69
coordination/F	40.49	38.63	44.44
prepinitial/S	17.02	14.96	7.17
advinitial/S	15.75	17.29	6.67
textinitial/S	2.45	3.73	2.2

Table 2: Distribution of individual features in per cent

characterizing the data. In the present study, over a half of squared Euclidean distance information, identified through the proportion of variance R^2, is captured in the first four dimensions. Figure 1 shows a scatterplot matrix of these four PCA dimensions: the y-axis in each of the rows corresponds to dimensions 1–3, whereas the x-axis in each of the columns corresponds to dimensions 2–4. For instance, the top left plot shows dimension 1 on the y-axis and dimension 2 on the x-axis. While PCA is unsupervised (i.e. meta information such as corpora or registers is not part of the statistical analysis), this information is visualized in the scatterplots to facilitate interpretation of the results.

As can be seen in Figure 1, particularly the first dimension foregrounds the register differences. However, the separation of the five registers present in the data is not complete. Looking at the first dimension, we can see that texts from the BUSINESS register are grouped together mostly on the negative side of the y-axis. Several texts from the POPSCI translations and ESSAY originals are also located on this side. SHARE was placed on the positive side of the axis together with some originals, mainly belonging to the registers of ESSAY and SPEECH. Moreover, around 0 we find another mixed group consisting of almost all texts from the POPSCI register as well as some originals from ESSAY and SPEECH. This distribution is also visible in the density plot shown in Figure 2.

Density curves visualize distribution of texts belonging to the specified categories – in this case the five registers represented in the data – along one of the PCA dimensions. The marks at the bottom stand for individual texts (Evert and Neumann, 2017, 57). The density plot also suggests that the business articles appear to be most similar to the popular-scientific texts.

Analysis of feature weights for this PCA dimension is inconclusive. Similar to the discussion of factor loadings in Factor Analysis, only features with weights below or above the arbitrary threshold of -/+0.3 are considered as significantly contributing to the distribution of texts (Levshina, 2015, 362). Other feature weights cannot be analyzed with certainty. As can be seen in Figure 3, the only linguistic feature with the weight below -0.3 is verbs/word, all other feature weights being in the range between -0.3 and 0.3. Figure 1 shows that business articles are grouped together on the negative side of the first PCA dimension. Therefore, we can conclude that the higher proportion of verbs in business articles is one of the factors that is responsible for this distribution.

While the separation of registers is even less clear along dimension 2, it is worth looking at the distribution of texts by the category of corpus. As shown in Figure 4, all four corpus categories appear to be spread along the whole dimension. However, comparing areas with the highest density per category, we may see a certain tendency of the published translations to be closer to the originals.

Figure 5 shows that the two corpora corresponding to edited and non-edited translations have almost the same distribution between -1 and 2 with the highest density around 0 on the x-axis, whereas all the texts from the CroCo corpus are spread more or less evenly along this dimension.

Dimension 4 does not seem to reflect any interesting patterns in terms of register, corpus or translation status.

5 Discussion

From the perspective of individual features, only slight tendencies could be observed, especially when considering the normalized counts. Some of these differences could be directly related to the previous studies of edited translations. Thus, the higher number of words per sentence and the lower number of sentences together with the

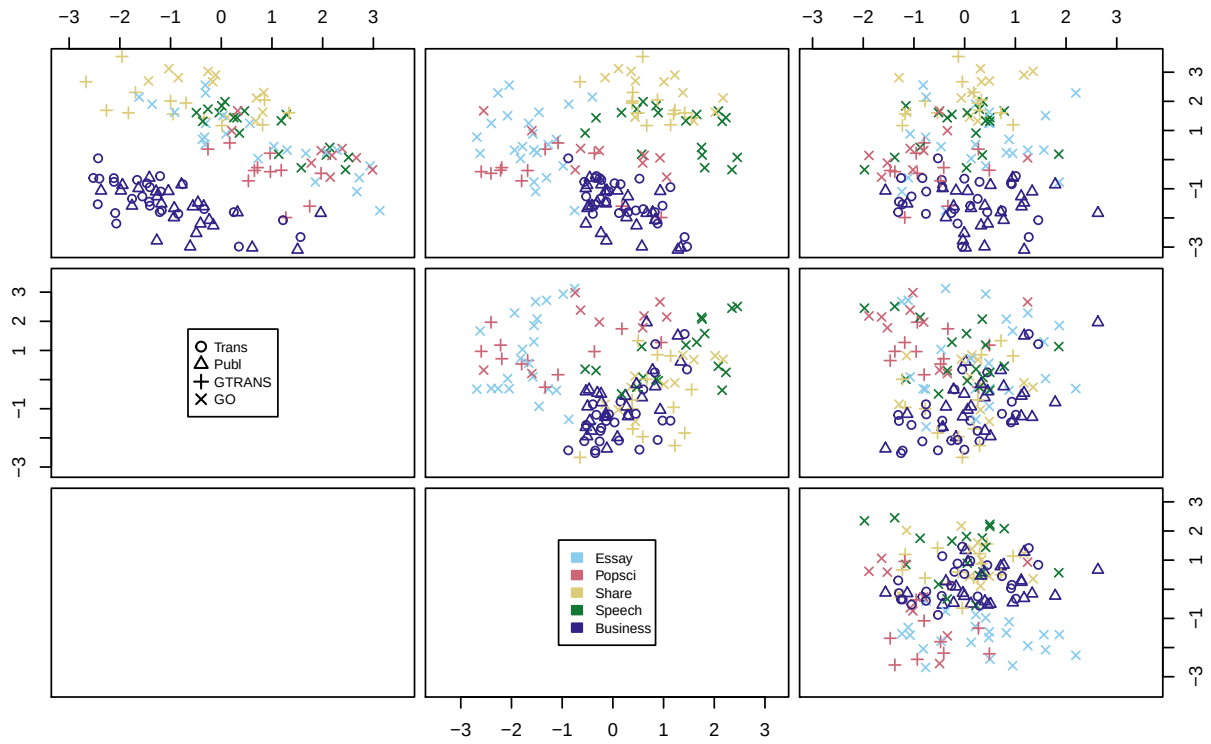

Figure 1: Scatterplot matrix of the first four PCA dimensions

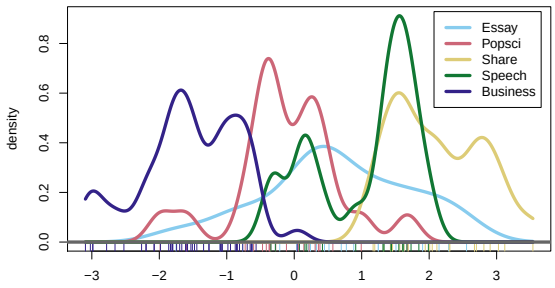

Figure 2: Density plot by register for the first PCA dimension

higher number of coordinating conjunctions attributed to translation manuscripts could be probably explained through sentence splitting (Bisiada, 2016). The difference in terms of the passive voice between the two translation versions within this corpus has been studied by Bisiada (2019). Potential changes by the editors related to the use of past tense and certain elements occurring in the theme position might also be interesting future research questions. The slightly increased numbers for adverbs and conjunctions as theme could indicate a tendency towards introduction of further cohesive devices by the editors – a change that would be in line with the aim of increasing readability of translations. While the comparison of the values to non-translations does not indicate that editors tend to normalize these features, it should be taken into account that the non-translations analyzed in the present study do not contain business articles and are thus not directly comparable to the two translation versions included in Harvard Business Corpus.

From the perspective of a multivariate analysis, we could observe some interesting patterns in the data, even though the identified groups of texts are not clearly separated. Our first research question concerns patterns in the distribution of translation manuscripts and edited translations in terms of their linguistic profiles. Based on the previous research in this area that showed some differences between the two translation versions (see Section 2), we could expect the PCA to separate them into two distinct groups of texts. However, the multivariate analysis did not show a profound effect of editorial intervention. In other words, the combined analysis of the 36 lexico-grammatical features considered in this study suggests that translation manuscripts and edited translations have similar linguistic profiles.

A partial explanation for the differences to the previous research in this area could be a differ-

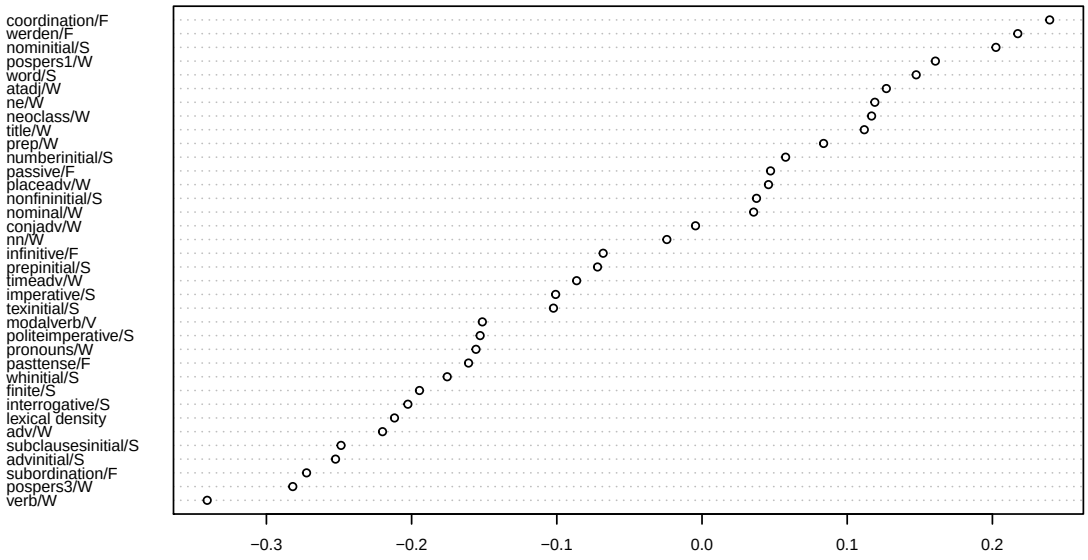

Figure 3: Dot chart of feature weights along the first PCA dimension

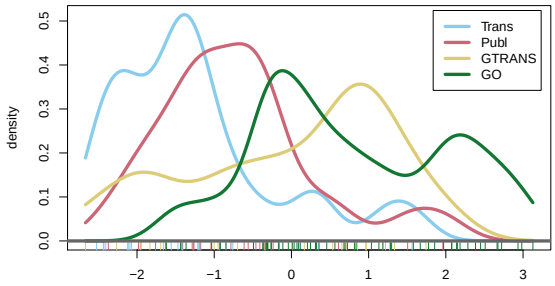

Figure 4: Density plot by corpus for the second PCA dimension

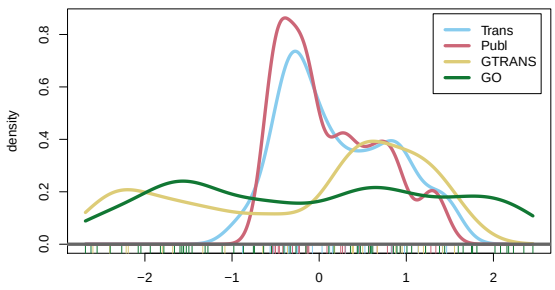

Figure 5: Density plot by corpus for the third PCA dimension

ent type of normalization of feature frequencies as compared to other studies on editorial intervention. As discussed above, a descriptive analysis of normalized feature counts indicated only very minor differences between the two translation versions. For instance, the use of nominalizations, which was reported as one of the differences between edited and non-edited translations (Bisiada, 2018a), is not among the individual features affected by editorial changes when normalized to the number of words per text.

Moreover, and more importantly, our method presents a holistic way of analyzing texts taking into account a large set of linguistic features that together form a linguistic profile. It allows us to generalize on a more global scale than methods focusing on specific features, thus improving our ability to compare text groups in general. That of course does not invalidate existing approaches, as the method applied here cannot detect specific changes that concentrate on a few features and may have a notable effect on the text without changing its overall linguistic profile. The multivariate method we apply in this study of the translation workflow is thus to be seen as complementary to more fine-grained analyses of specific features.

With respect to the second research question

concerned with the translation property of normalization, that is, translated texts being more similar to the comparable originals within the same language (Baker, 1996), we found a slight tendency for some of the edited translations to be closer to the German originals included in our data set, as compared to the translation manuscripts. This means that some of the changes introduced by editors could result in translations being more conventional in the target language, in our case German. The findings should be confirmed using a larger data set. In particular, adding a category of originals comparable to the translation versions included in the Harvard Business Corpus in terms of register, as well as the corresponding English originals could help us explain the unexpected distribution of German translations from the CroCo corpus analyzed for the present paper.

Moreover, the analysis has indicated differences between registers included in our data sample. These contrasts are detected by the most informative first dimension of the PCA. Along this dimension, both translation versions were grouped together as belonging to the same register of business articles. Letters to shareholders, which are comparable to business articles in terms of topic, appear to have very different distributions of analyzed features. In contrast, the popular-scientific register, which is comparable to business articles in terms of aim, seems to have a more similar linguistic profile to the texts taken from the Harvard Business Corpus. One potential explanation could be the fact that our analysis does not contain purely lexical features. It is possible that if individual lexical items were considered as well, then more similarities between business articles and letters to shareholders could be detected. Based on the lexico-grammatical features that are included in the analysis, the results suggest that it is not the topic but rather the aim of texts that is more important for the classification of texts according to register. A follow-up study might consider re-analyzing the business articles as a type of popular-scientific publication.

None of the PCA dimensions has detected differences between originals versus translations within the same language, as was shown, for instance, in Baroni and Bernardini (2006). It is possible that the register effect is so strong that it obscures any effect of translationese.

The present study considers only three sources of variation within the texts, namely translation status (translated vs. non-translated texts), editorial intervention (edited vs. non-edited translations) and register. However, other factors may also play a role. For instance, Figure 5 shows that the CroCo texts are evenly distributed along the third PCA dimension. This might suggest that another source of variation not considered in this study might play a role. It is conceivable that individual variation is responsible for this distribution of texts: taken into account the fact that texts from the CroCo corpus are publications taken from a variety of sources, in contrast to the Harvard Business Corpus, which consists of business articles taken from one magazine, the CroCo texts are likely to contain texts by more individual writers. Unfortunately, both corpora do not contain detailed meta-information, so that it is not possible to include authors/translators/editors as another category that could explain the PCA results.

Following further steps of the GMA procedure (Evert and Neumann, 2017), future research will involve a combination of PCA and a Linear Discriminant Analysis (LDA). This analysis performed on a larger data set involving not only categories considered in the present study but also English originals and German non-translated business articles may lead to finding further meaningful patterns within the data and thus refining the linguistic profiles of translation manuscripts and edited translations.

Acknowledgements

We would like to thank Stefan Evert for developing the R scripts for the GMA procedure and the COMTEX team for modifying the CQP scripts for German. We would also like to thank Florian Frenken for helping us with data pre-processing. Part of the research was funded by the German Research Foundation (DFG) research grant no. NE1822/2-2 and by the Spanish Ministry for Science and Innovation (MICINN), with grant no. PID2019-107971GA-I00.

References

Mona Baker. 1996. Corpus-based translation studies. In Harold Somers, editor, *Terminology, LSP and Translation*, pages 175–186. John 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.

Jocelyne Bisaillon. 2007. Professional editing strategies used by six editors. *Written Communication*, 24(4):295–322.

Mario Bisiada. 2016. "Lösen Sie Schachtelsätze möglichst auf". *Applied Linguistics*, 37(3):354–376.

Mario Bisiada. 2017. Universals of editing and translation. In Silvia Hansen-Schirra, Oliver Czulo, Sascha Hofmann, and Bernd Meyer, editors, *Empirical Modelling of Translation and Interpreting*, pages 241–275. Language Science Press, Berlin.

Mario Bisiada. 2018a. Editing nominalisations in English–German translation. *The Translator*, 24(1):35–49.

Mario Bisiada. 2018b. The editor's invisibility. *Target*, 30(2):288–309.

Mario Bisiada. 2018c. Translation and editing. *Perspectives: Studies in Translation Theory and Practice*, 26(1):24–38.

Mario Bisiada. 2019. Translated language or edited language? A study of passive constructions in translation manuscripts and their published versions. *Across Languages and Cultures*, 20(1):35–56.

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 224–231.

Andrew Chesterman. 2004. Hypotheses about translation universals. In Gyde Hansen, Kirsten Malmkjær, and Daniel Gile, editors, *Claims, Changes and Challenges in Translation Studies*, pages 1–13. John Benjamins, Amsterdam.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, pages 174–204. de Gruyter, Berlin.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In Gert de Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Theoretical and Methodological Traditions*, pages 47–80. Mouton de Gruyter, Berlin.

Cathrine Fabricius-Hansen. 1996. Informational density: a problem for translation and translation theory. *Linguistics*, 34(3):521–566.

Cathrine Fabricius-Hansen. 1999. Information packaging and translation. In Monika Doherty, editor, *Sprachspezifische Aspekte der Informationsverteilung*, pages 175–214. Akademie Verlag, Berlin.

Jennifer Fest, Arndt Heilmann, Oliver Hohlfeld, Stella Neumann, Helge Reelfs, Marco Schmitt, and Alina Vogelgesang. 2019. Determining response-generating contexts on microblogging platforms. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, pages 171–182.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*. de Gruyter, Berlin.

Haidee Kruger. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target*, 24(2):355–388.

Haidee Kruger. 2017. The effects of editorial intervention: Implications for studies of the features of translated language. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, pages 113–156. de Gruyter, Berlin.

Haidee Kruger and Bertus van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures*, 13(1):33–65.

Natalia Levshina. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Benjamins, Amsterdam.

Stella Neumann. 2013. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. de Gruyter, Berlin.

Stella Neumann and Stefan Evert. Forthcoming. A register variation perspective on varieties of english. In Elena Seoane and Douglas Biber, editors, *Corpus based approaches to register variation*. de Gruyter, Berlin.

Stella Neumann and Silvia Hansen-Schirra. 2012. Corpus methodology and design. In Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner, editors, *Cross-linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*, pages 21–33. de Gruyter, Berlin.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging Deutscher Textcorpora mit STTS*. Universität Stuttgart, Universität Stuttgart.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Kare Solfjeld. 2008. Sentence splitting—and strategies to preserve discourse structure in German–Norwegian translations. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *Subordination versus Coordination in Sentence and Text: A Cross-Linguistic Perspective*, pages 115–133. John Benjamins, Amsterdam.

Margherita Ulrych and Amanda Murphy. 2008. Descriptive translation studies and the use of corpora: Investigating mediation universals. In Carol Taylor Torsello, Katherine Ackerley, and Erik Castello, editors, *Corpora for University Language Teachers*, pages 141–166. Peter Lang, Frankfurt/M.

Appendix A: List of features

Feature name	Description
word/S	Number of words/number of sentences
lexical density	Number of lexical words/number of words
nn/W	Number of common nouns /number of words
ne/W	Number of proper nouns/number of words
nominal/W	Number of nominalizations/number of words
neoclass/W	Number of neoclassical compounds/number of words
pronouns/W	Number of all pronouns/number of words
pospers1/W	Number of 1st person pronouns/number of words
pospers3/W	Number of 3rd person pronouns/number of words
adv/W	Number of adverbs/number of words
atadj/W	Number of attributative adjectives/number of words
prep/W	Number of prepositions/number of words
finite/S	Number of finite verbs/number of sentences
pasttense/F	Number of past tense verbs/number of finite verbs
werden/F	Number of instances of the modal verb *werden* (future)/number of finite verbs
modalverb/V	Number of modal verbs/number of verbs
verb/W	Number of all verbs/number of all words
infinitive/F	Number of infinitives with *zu*/number of finite verbs
passive/F	Number of instances of passive voice/number of finite verbs
coordination/F	Number of coordinating conjunctions/number of finite verbs
subordination/F	Number of subordinating conjunctions/number of finite verbs
interrogative/S	Number of instances of interrogative mood/number of sentences
imperative/S	Number of instances of imperative mood/number of sentences
politeimperative/S	Number of polite imperatives/number of sentences
title/W	Number of titles/number of words
placeadv/W	Number of adverbs of place/number of words
timeadv/W	Number of adverbs of time/number of words
conjadv/W	Number of conjunctive adverbs/number of words
nominitial/S	Number of nominal elements in theme position/number of sentences
numberinitial/S	Number of numbers in theme position/number of sentences
prepinitial/S	Number of prepositions in theme position/number of sentences
advinitial/S	Number of adverbs in theme position/number of sentences
textinitial/S	Number of conjunctions in theme position/number of sentences
whinitial/S	Number of *wh*-elements in theme position/number of sentences
nonfininitial/S	Number of infinitives with *zu* in theme position/number of sentences
subclausesinitial/S	Number of subordinate clauses in theme position/number of sentences

Table 3: List of features

Appendix B: Boxplots

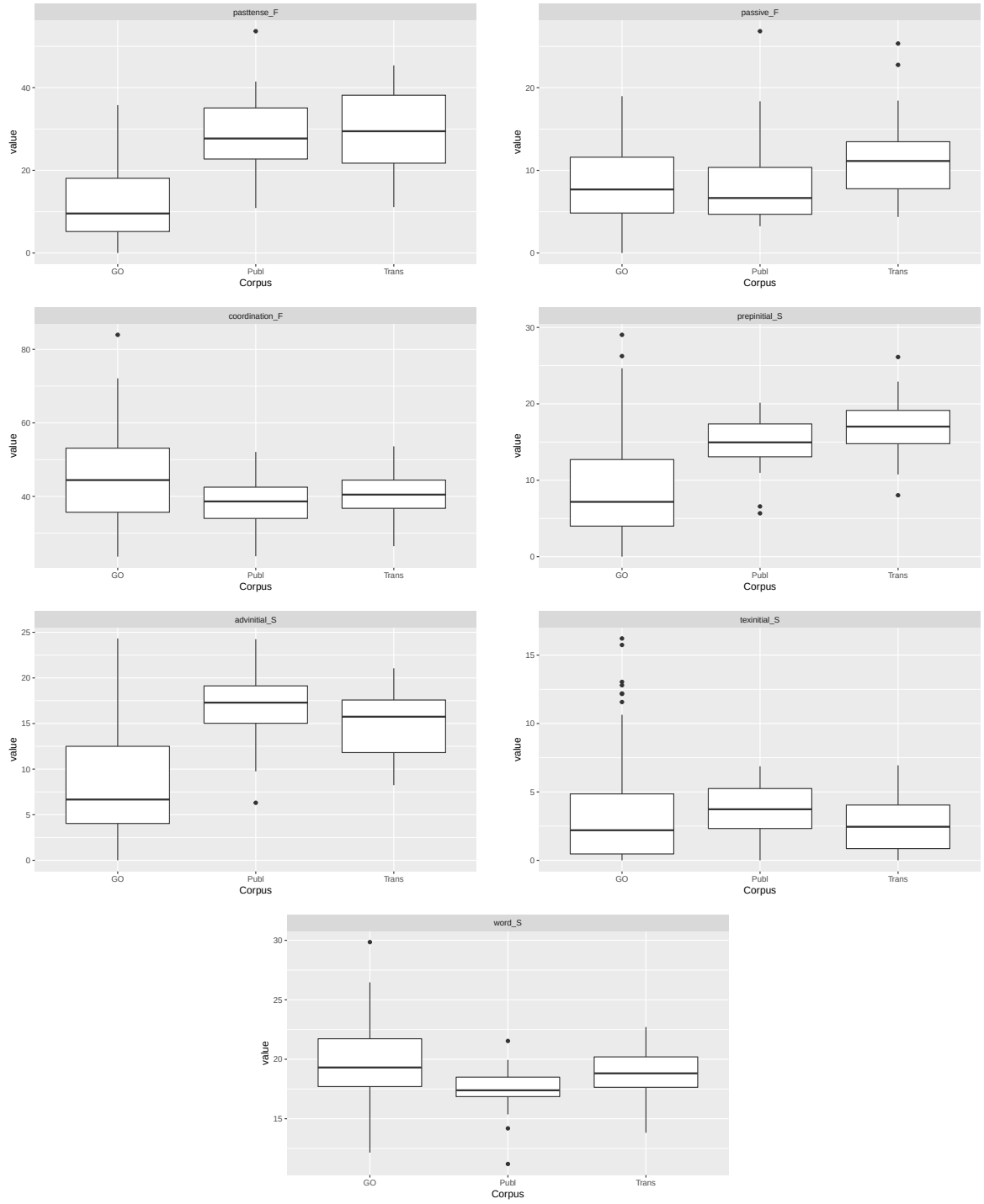

Figure 6: Distribution of selected features across three sub-corpora

Multiword expressions as discourse markers in Hebrew and Lithuanian

Giedre Valunaite Oleskeviciene
Institute of Humanities,
Mykolas Romeris university
Ateities 20 LT-08303,
Vilnius, Lietuva
gvalunaite@mruni.eu

Chaya Liebeskind
Department of Computer Science,
Jerusalem College of Technology
21 Havaad Haleumi st. 9116001,
Jerusalem, Israel
liebchaya@gmail.com

Abstract

Multiword expressions are of key importance in language generation and processing. Certain multiword expressions also could operate as discourse markers. In this research, we combined the alignment model of the phrase-based statistical machine translation and manual treatment of the data in order to examine English multiword discourse markers and their equivalents in Lithuanian and Hebrew, by researching their changes in translation. After establishing a full list of multiword discourse markers in our generated parallel corpus, we focused on the two most frequent ones functioning as stance attitudinal discourse markers: *I think* and *you know* aiming to research if they demonstrate their functional stability as stance attitudinal discourse markers in translation and what changes they undergo in Lithuanian and Hebrew translation. Our research proves that the examined multiword discourse markers preserve their function as stance attitudinal discourse markers and tend to remain multiword discourse markers in the Hebrew translation but turn into one-word discourse markers in Lithuanian due to the translation tendency relying on inflections.

1 Introduction

Research on multiword expressions has identified that language is not produced just word by word but it usually involves generating certain chunks using a lot of formulaic constructions (Barlow, 2011). Native speakers have a multitude of memorized sequences to perform various functions within language, for example, organizing discourse (Nattinger and DeCarrico, 1992), or processing language by the speaker and the hearer (Siyanova-Chanturia et al., 2011). Formulaic language includes idioms and proverbs, various clichés and collocations, lexical bundles, and phrasal verbs. Biber et al. (2004) observed that lexical bundles constitute a high percentage of the produced language and the authors identified that one function of lexical bundles is to organize discourse by providing an example of such bundles, for example, *I think*, which relates to the research on discourse markers (DMs). Phrases such as *you know* and *I think* have also been classified as DMs that perform certain discourse organising functions. However, Maschler and Schiffrin (2015) observe that there is no a priori theoretical classification of DMs and the analysis of function in the data is necessary. Research on DMs as tools of discourse management prove that they carry several functions, including signposting, signalling, and rephrasing. Furthermore, there are ongoing attempts to investigate the importance of discourse layers in language production, communication, second language learning, and translation. Additionally, Dobrovoljc (2017) has recently attempted to research multiword expressions as DMs in a corpus of spoken Slovene, identifying structurally fixed discourse marking multiword expressions.

The underlying assumption is that DMs *I think* and *you know* are indicators of stance in discourse used to express and understand points of view and beliefs. The purpose of the current research is to examine multiword expressions used as DMs in TED talk English transcripts focusing on stance attitudinal DMs *I think* and *you know* and compare them with their counterparts in Lithuanian and Hebrew by following Maschler and Schiffrin (2015) observation on the necessity of closer investigation on their function as stance DMs. To achieve the aim of the research, the set objectives were to create a parallel research corpus to identify multiword expressions used as stance at-

titudinal DMs and to analyse their translations in Lithuanian and Hebrew to determine if they function as stance DMs and are also multiword expressions or one word translations, or if they acquire any other linguistic forms. An additional benefit of the study was extending the available resources and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English, Lithuanian, and Hebrew); the created corpus is shared and interlinked via CLARIN open language resources. What is more, the current research could be extended to other languages. The future research envisions applying machine learning and using the model for discourse marker identification in other languages to research how stance signalling is treated.

2 Theoretical background

The literature overview briefly takes into account the research languages, studies related to multiword expressions and their use as DMs, the importance of DMs for discourse management, and certain insights into DM translation.

2.1 Cultural heritage and research languages

First, it is necessary to briefly discuss the cultural heritage of the languages of the research, which, in a way, guided the choice of languages for our study. According to Bieliauskienė (2012), Jewish and Lithuanian cultures coexisted on the same territory from the first half of the 14th century. The author stressed that from 19th century onwards, in the Republic of Lithuania, Vilnius was called Lithuania's Jerusalem, attracting knowledgeable people in the field of education and inspiring a flourishing high culture, for example, in theatre, art, and literature. In fact, both languages, Lithuanian and Hebrew, formed the cultural heritage of the region. In this study, we research the Lithuanian and Hebrew corpus in parallel with pivotal English.

Lithuanian is an old surviving Baltic language, retaining forms related to Sanskrit and Latin and preserving the most phonological and morphological aspects of the Proto-Indo-European language. Thus, it has gained importance in Indo-European language studies and has been researched by many scientists so far, including Ferdinand de Saussure, who considered Lithuanian "the Galapagos of linguistic evolution" (Joseph, 2009). Lithuanian is rich in declensions and cases inside the declensions and the oldest layer of the Lithuanian language vocabulary is related to the Indo-European language, which is dated to be approximately over 5000 years old.

Hebrew is a very old Semitic language and it is a successful example of a revived dead language. It survived in the medieval period as the language of religious scriptures, being revived, in the 19th century, into a spoken and literary language (Joslyn-Siemiatkoski, 2007). Hebrew is an important language for researchers specializing in Middle East civilizations and Christian theology studies.

2.2 Multiword expressions as DMs

The research areas of natural language processing (NLP), linguistics, and translation are closely related to discourse research, focusing on discourse relations between clauses or sentences. NLP research focuses more and in depth on multiple language-related areas, such as semantic phenomena, dialogue exchange structure, and discourse textual structure (Webber and Joshi, 2012). NLP recognizes that language is not just placing words in the right order but getting the meaning and deeper textual relations as well as organizing ideas into a logical textual flow. According to researchers (Barlow, 2011; Sinclair, 1991), language is not just generated word by word; it is also formulaic. Speakers possess multiple learnt formulaic sequences, which, according to Siyanova-Chanturia et al. (2011) are important in organizing discourse and help the language producer and recipient to manage language processing. However, formulaic language is not easy to manage and categorize for NLP research, as it may seem at first sight, since the sequences that could be considered formulaic vary in length, meaning, fixedness, etc., and the finalized definition of formulaic language has not fully crystallized. It could be considered as an umbrella term embracing idioms, proverbs, clichés, phrasal verbs, collocations, and lexical bundles (Wray, 2012). According to Wei and Li (2013), formulaic language covers approximately 60% of written texts in their researched corpus of English academic language. According to Biber et al. (1994, 1999), lexical bundles are groups of words that show a statistical tendency to co-occur and could be considered as extended collocations, for example, *I think*. Biber et al. (2004) identify that lexical bundles have functional purposes, such as organizing discourse, expressing stance, and

referential meaning. Based on the evidence of the formulaic nature of language for communication, research has turned to investigating multiword expressions used as DMs (Dobrovoljc, 2017), identifying structurally fixed discourse marking multiword expressions.

Another important issue in NLP is discourse management, which is related to discourse relations, connecting ideas between sentences and bigger parts of the text. Discourse relations may remain implicit or be expressed explicitly through discourse markers, which help textual coherence and discourse management, and are used for making coherent speech appropriately segmented to enable textual understanding. DMs perform important functions, such as signposting, signalling, and rephrasing, by facilitating discourse organization. They are mainly drawn from syntactic classes of conjunctions, adverbials, and prepositional phrases (Fraser, 2009), as well as expressions such as *you know*, *you see*, and *I mean* (Schiffrin, 2001; Hasselgren, 2002; Maschler and Schiffrin, 2015). Hasselgren (2002) advocated that better DM signalled fluency contributes to interaction and even makes the speaker sound more 'native-like'. Recently, discourse relations and DM research has gained certain impetus with corpora annotation for exploring discourse structure in texts, for example, the Penn Discourse Tree Bank (PDTB) (Webber et al., 2016). Furthermore, there was a rise in annotated multilingual corpora for researching different means of expressing discourse relations and managing discourse (Stede et al., 2016; Zufferey and Degand, 2017; Oleskeviciene et al., 2018).

Language, especially spoken, is characterised by DM use; however, some of them (e.g., *you know*, *I think*, *well*) are sometimes referred to in a critical manner, as indicating a lack of fluency (O'Donnell and Todd, 2013). Still, DMs are abundantly used and, according to Crystal (1988), they enhance communication if used appropriately and should not be considered unnecessary or undesirable. As Biber (2006) observed, DMs, such as *you know*, or *well*, are very rare in written language. However, they are quite common in spoken discourse and should not be treated as just fancy words since they serve the function of organizing discourse by signalling, rephrasing, marking, or relating ideas. Svartvik (1980) observed that, if a foreign language learner makes a mistake (e.g., he goed), it can be easily identified and redeemed by the native speaker; however, if a learner misses words such as *you know*, or *well*, the native speaker cannot identify any error and the speech might sound impolite or even dogmatic. The same idea is also supported by Hasselgren (2002), who observed that DMs enhance interaction. Furthermore, it has also been researched using learner corpora to demonstrate the importance of discourse level knowledge, especially at more advanced levels of language learning (Granger, 2015; Cobb and Boulton, 2015).

2.3 Translation issues of DMs

DMs are used in both written texts and spoken discourse to connect ideas and guide the reader or the listener through expression by ensuring that the ideas are grasped correctly. DMs have been researched by applying various theoretical approaches, such as Rhetorical Structure Theory (Mann and Thompson, 1988), Segmented Discourse Representation Theory (Asher et al., 2003), and PDTB (Prasad et al., 2008), first focusing on the monolingual approach, which resulted in multilingual studies focusing on translation (Degand and Pander Maat, 2003; Pit, 2007; Dixon, 2009; Zufferey and Cartoni, 2012). As Zufferey and Cartoni (2012) observed, multilingual studies are more complicated as languages differ in the use of DMs and their expression. The authors also added that often DMs are poly-semic, which means that a single expression of a DM may perform in expressing various discourse relations. They provided an example of the English *since*, which could express temporal or causal discourse relations depending on the surrounding contexts.

Recently, much research has gained interest in using parallel translated corpora. For example, Dupont and Zufferey (2017) focused on the investigation of translation corpora to study if the effect of register, translation direction, or translator's expertise could influence the shifts of meaning and omissions of English and French markers of concession. Hoek et al. (2017) investigated a parallel corpus on English parliamentary debates translated into Dutch, German, French, and Spanish, searching what types of DMs might have a higher tendency to be more frequently omitted in translation. Baker (2018), in her extensive studies on translation, observed that DMs could be used to signal different relations and these relations could

be expressed by a variety of means. The author provided the example that, in English, the expression of causality could be realized through content verbs, such as *cause* or *lead*, or more simply, through a DM signalling the causality relation. Further, different languages demonstrate different tendencies – some languages prefer using simpler structures connected by a variety of DMs, while other languages favour complex structures, sparsely using explicit DMs. The author analysed the example of an evident difference between English and Arabic, identifying that, while English prefers signalling discourse relation through DMs, Arabic prefers grouping the information into bigger grammatical chunks and using fewer DMs. The finding is supported by Al-Saif and Markert (2010), who observed that, in Arabic, many discourse relations are expressed via prepositions with nominalizations. Therefore, translation poses a challenge in adapting various preferences of the source and target languages. Translators face various choices of inserting DMs to make the flow of the ideas smoother in the target text, however, they risk making the translation sound foreign or transposing the grammatical syntactic structure, ending up using different means of expressing DMs or simply omitting them. It appears that it is not always possible to use the word for word technique and natural changes in translation are sometimes inevitable. According to Baker (2018), grammatical changes in translation involve certain techniques, such as substitution, transposition, omission, and supplementation.

Substitution is the change of the grammatical category of the source unit in translation.

For example, active voice is more common in Lithuanian; therefore, English passive voice units could be changed into active units:

1. He was told the news. – jam pranešė naujienas

   Similarly, in the following example, the verb in the source language is changed into a noun in Hebrew translation.

2. We should have broken ten minutes before. – היינו צריכים לצאת להפסקה לפני עשר דקות

   Transposition represents a change of position in the order of elements of the source textual unit or changing the part of speech in translation, which implies the change in the order of the elements in the translated text.

   In Lithuanian translation, we observe a change in the order of the elements in the sentence.

3. After he had left – Jam išėjus.

   In the case of Hebrew translation, the change of the order of the elements could be observed in the following example.

4. Classical music – מוזיקה קלאסית

   Omission occurs when some elements of the original text could be considered excessive or redundant in translation.

   In the Lithuanian translation example, the whole phrase I thought is omitted.

5. I thought you said you were alright. – Bet tu sakei, kad viskas gerai.

   In the following example in Hebrew, the translation of *are* is omitted.

6. We still are – אנחנו עדיין

   Supplementation involves changes when new elements, which are non-existent in the source text, appear in the translated text in order to ensure structural adequacy of the latter. Such modifications are usually considered structurally or contextually motivated.

   For example, due to the elliptical nature of the English language, the Lithuanian translation should use supplementation to make the translation understandable.

7. Soap star – muilo operos žvaigždė (although the word opera is omitted in English due to ellipsis, it should be added in Lithuanian translation to make it contextually coherent).

   The same technique should be applied in the Hebrew translation.

8. Soap star – כוכב אופרת סבון

As shown above, translation is not a mere process of transposing words from one language into another but requires certain motivated changes. Thus, translation involves grammatical transformations, as a result of the process of looking for approximate correspondences in the translated texts.

2.4 Research data resources

It should be stressed that parallel data resources are not extensive, and researchers still need to work on creating parallel corpora for their research, especially if they would like to cover the variety of languages and areas. One of the most prized parallel multilingual resources is Europarl (Koehn, 2005). It comprises the translations of the European Parliament proceedings (at most 50 million words) in most European languages; however, it covers just one specific domain of parliamentary proceedings. TED talks subtitles to their videos seem to be a growing resource of parallel linguistic material, covering a multitude of languages. In addition, being an open and a developing resource, TED talks attract attention of researchers and their subtitles cover a wide variety of knowledge fields (Cettolo et al., 2012), which makes the data of the talks widely applicable. However, researchers should keep in mind that the talks are translated by volunteers although with administratively managed quality checks, and the translation is mostly unidirectional from source English subtitles to other target languages. Furthermore, Dupont and Zufferey (2017) identified that such talks contain features of both spoken and written language, as they are semi-prepared speeches by nature. Additionally, Lefer and Grabar (2015) observed that subtitle translation bears certain specificity in itself. Even by taking into account the features of TED talks discussed by researchers, TED talks are extensively useful as they are an open resource and could provide large amounts of parallel data for research. Besides, parallel corpora are employed as a pool of data for statistical machine translation systems and TED talks is one of the most frequent data resources referred to explore multilingual Neural MT (NMT) (Aharoni et al., 2019; Chu et al., 2017; Hoang et al., 2018; Khayrallah et al., 2018; Tan et al., 2018; Xiong et al., 2019; Zhang et al., 2019). NMT, as currently the newest technique of MT, stems from the model of the functioning of the human brain neural networks, which place information into different layers for processing it before generating the outcome. With the technological advancements, NMT gained impetus, as it used to be, resource and computation wise, too costly to outdo phrase-based MT, which operates on the basis of translating entire sequences of words. Now, the neural approach of NMT started challenging the long-lasting prevalence of phrase-based MT techniques.

3 Research methodology

The detailed description of the research procedures is provided in the research methodology section. In the current research, phrase-based MT was applied relying on two main reasons: NMT techniques do not allow extensive processing of phrases and NMT procedures are not as explicit as phrase-based MT processes. The current study does not involve the full set of phrase-based MT systematic procedures, as it is used just for a phrase table construction, which is a single step of the phrase-based MT paradigm. The research aim comprised examining multiword expressions used as DMs in TED talk English transcripts and comparing them with their counterparts in Lithuanian and Hebrew. Thus, there was a need to achieve the double objectives of creating the parallel corpus for the research data and carrying out the research on multiword expressions used as DMs in the studied languages. Unlike working on one language and using statistical methods we used parallel corpus knowledge alignment algorithm. Initially, the list of multiword and one word expressions that could potentially be used as DMs was generated relying on theoretical insights by Schiffrin (1987) and the classification provided by Fraser (2009). Fraser's extensive classification was taken as a basis, and Huang (2011) theoretical analysis of DM characteristics for spoken discourse, for example, *you know*, *you see*, *I mean*, *I think*, was also included.

3.1 Parallel Corpus creation

First, a parallel corpus meeting the research aim needed to be created. We decided to use TED Talk transcripts, as they are publicly available and provide appropriate material for parallel data. In order to create a substantial parallel corpus containing data in English, Lithuanian, and Hebrew, the talks were extracted automatically using a special code, which ensured that English sentences with the candidate DMs from the theoretically based list were extracted and matched with their Lithuanian and Hebrew counterparts. The process of creating the parallel corpus allows parallelizing the data of any researched languages. While building the corpus, the parallel texts in English, Lithuanian, and Hebrew were extracted from TED talk transcripts. Then, the sentences were aligned to make

a parallel corpus for further research. The corpus contains 87.230 aligned sentences (published in LINDAT/CLARIN-LT repository `http://hdl.handle.net/20.500.11821/34`).

3.2 Multiword DM extraction

Another stage of the research focuses on multiword expressions that are used as DMs to ensure textual cohesion and, according to Fraser (2009), to relate separate discourse messages. For example, phrases such as *you know*, *I mean*, *of course*, are characteristic of spoken language (Maschler and Schiffrin, 2015; Furkó and Abuczki, 2014; Huang, 2011). Thus, 3.314 aligned sentences containing the earlier mentioned multiword expressions were extracted and manually annotated, spotting the cases in which the expressions were used as DM. One-word DM identification did not represent much challenge; however, turning to multiword expressions, they certainly caused challenges. For example, to identify if the expression *you know* is used as a DM, the context in which it occurs should be examined by identifying if the expression serves as a DM. As such, two situations arise: (1) the multiword expression *you know* is used to introduce a new discourse message, or (2) they are content words fully integrated into the sentence.

1. You know, this is really an infinite thing.

2. You know exactly what you want to do from one moment to the other.

After that, the variations of the translations of DMs into Lithuanian and Hebrew were extracted automatically for a comparative study, determining the variations in translation. We ran an NLP word-alignment algorithm to extract a phrase table of all the possible translations of the researched DMs, using our parallel corpus (in our case, source = English, target = Lithuanian/Hebrew). The extraction of the translation variations was dependent on the phrase-based statistical machine translation model introduced by Koehn et al. (2003). The model could be visually represented in the research languages by the figures below.

Figure 1: Lithuanian – English phrase alignment

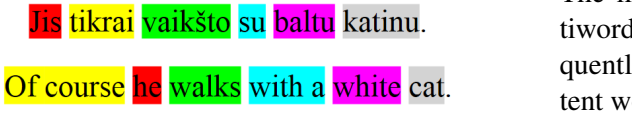

Figure 2: English – Hebrew phrase alignment

Figure 1 visualizes Lithuanian–English corresponding phrases marked in respective colours. Figure 2 shows English–Hebrew respective phrase alignment, with a note for the reader that Hebrew text should be read from right to left.

The model applies the segmentation of the input into sequences of words, which are called phrases, and then each phrase is translated into English phrases that could later be reordered in the output. Such a model ensures the correspondence between the units of phrases. After being extracted, all the possible translations were manually filtered to reject the wrong translation variants and prepare the data for the machine analysis stage. This helped us extract sentences with translations of the researched DMs from the target language corpus and analyse their use.

While analysing the data, we noticed that there was a small amount of data left which did not fit the variations of possible translations. The first supposition was that it might represent the cases of omissions; however, we decided to analyse it closely to verify. We checked manually the extracted non-attached data and established that most of the analysed cases involved omission with some minor grammatical transformation cases, incorrect translations, and some phrases not included in the possible translations by the machine.

4 Research findings

4.1 Multiword DM distribution

The most frequent multiword expressions used in the study corpus have been extracted and are presented in the table below.

It could be seen in Table 1 that the two most frequent multiword expressions in the corpus are *I think* and *you know*.

As mentioned earlier, multiword expressions needed to be manually annotated, spotting the cases when the expressions were used as DMs. The manual annotation revealed that some multiword expressions were used as DMs more frequently while others were more often used as content words fully integrated into sentences.

Multiword expression	Frequency
I think	580
You know	573
That is	370
Of course	312
You see	287
In fact	256
I mean	199
For example	161

Table 1: Multiword expressions in the corpus

Multiword expression	Discourse marker	Content word
I think	473	107
You know	380	193
That is	29	341
Of course	233	79
You see	47	240
In fact	217	39
I mean	168	31
For example	117	44

Table 2: Multiword expressions used as DMs

It is visible in Table 2 that multiword expressions *That is* and *You see* although identified as DMs by the theoretical literature, in this study, they demonstrate a weak tendency to be used as DMs and are mainly used as content words in the current corpus. While multiword expressions *I think* and *you know* demonstrate a high tendency of being used as DMs and the stability of remaining DMs in Lithuanian and Hebrew translation.

4.2 DM 'I think' translations

Further, following our research aim, we present a detailed analysis of the translations of the two most frequent multiword expressions used DMs – *I think* and *you know*. The alignment approach allowed extracting direct output of the translations together with the figures of the translation frequency. First, we explore the translations of the most frequent multiword DM, *I think*.

The most frequent multiword expression in the researched corpus, *I think*, has a number of translation variants in both researched languages, Hebrew and Lithuanian. The most frequent one in Lithuanian is a one-word expression – an inflected verb, *manau*, which, due to Lithuanian being a highly inflected language (Zinkevičius et al., 2005), fully represents the verb-pronoun cases. Other one-verb variants and multiword expressions do not demonstrate high numbers. A separate case is represented by omission, which comprises 48 situations, showing that such a technique is also chosen by the translators.

Referring to Hebrew, the most frequent translation is **אני חושב**, which refers to a male derivative, while the female derivate, **אני חושבת**, comprises only 51 cases. The assumption could be that the choice of gender in first person pronouns depends on the gender of the speaker. However, Hebrew translation variant choices differ from the Lithuanian ones, as they mostly remain multiword expressions in translation. Another interesting observation in Hebrew is that a number of 70 cases include the additionally integrated connective *and* into the derivative **ואני חושב**. It reveals that sometimes translators prefer inserting additional information into the translation, which could be related not to the direct semantic meaning of addition of *and* but more to the pragmatic inferences drawn by the translators form the surrounding contexts, which relates to the observations of Blakemore and Carston (1999), and Moeschler (1989). Hebrew demonstrates less omission cases than Lithuanian for the DM *I think*. The number of omissions in Hebrew is 23, while the Lithuanian omission number is approximately double in the parallelized corpus sentences.

4.3 DM 'you know' translations

Another commonly used multiword DM, *you know*, demonstrates far more variable translations. A closer investigation into the translations of DM *you know* reveals that the most common ones in Lithuanian are also one-word verbs *žinote/ žinai/ žinot*, which represent verb-pronoun cases. Another quite frequent translator choice is the single particle *na*. Although not numerous, very interesting cases of multiword expressions with particles could be found, such as *na jūs žinote* or *na suprantate*, or a single particle *juk*. Even a single particle is used as a DM, which is characteristic of the Lithuanian language. There are also cases of multiword expressions involving a connective and inflected verb phrases, for example, *kaip žinote*, *bet žinote*. The translator's choice to additionally use particles or connectives is obviously related not to the translation of semantic meaning but more to the pragmatic meaning inferred by them from

the surrounding context. It connotes with the deep observation made by Nau and Ostrowski (2010b) that Lithuanian particles contain the component of subjectivity and inter-subjectivity, and their meaning is mostly coloured by the surrounding context.

In Hebrew, the translation variants for the DM you know are not as variable. The most frequent ones, again, are the variants referring to the male gender, including both plural (191) אתם יודעים and singular (26) אתה יודע, which by far exceeds the number of female derivatives in plural (2) אתן יודעות and singular (17) את יודעת. The prevalence of male derivatives could be explained by the nature of the Hebrew language, which has the feature that male derivatives are used while addressing purely male and mixed audiences (Tobin, 2001). In Hebrew, this DM is much prone to omission, as the number of omissions amounts to 113 cases, which are a bit less than the number of the translated cases. Again, multiword expressions remain multiword expressions with just one case of one-word choice in translation. The translation choices for the multiword expression serving as a DM *you know* are more versatile than those of *I think* and certain cases of grammatical transformation could be observed in the case of the former.

In Lithuanian, eight cases of grammatical changes were found and, even amongst those, one-word DMs prevail. The multiword DM *you know* is translated also into a connective, *taigi* (so), and adverbs *gerai* (okay) and *iš tiesų* (really). However, such translator choices are absolutely rare, considering the size of the dataset.

The grammatical transformation cases are more numerous, comprising of 21 occurrences, and much more versatile in Hebrew. The most interesting cases include: טוב נו (okay), which is a usual colloquial saying in Hebrew, נחשו מה (guess what), and two connectives used successively, כאילו (as if). There are also some cases when a connective is just added as in the following example ואז כמובן, (then of course), which could be done by the translator simply to stress the discourse management role of the DM used or possibly attaches a rhetorical function to the integrated connective. Even among the limited cases of grammatical transformation, multiword expressions as DMs prevail in Hebrew. What is similar to Lithuanian is that there are also adverbs used in the Hebrew translation: הרי (indeed), נו (well), ברור (clearly). Reflecting why different DMs demonstrate different translation choices could be based on the nature of the target language into which the texts are translated; for example, Lithuanian is rich in particles and, as the analysis has demonstrated, translators choose to additionally integrate particles into DMs to add supplementary discourse expressions.

In Hebrew, the male gender prevails in translation, and translators automatically give preference to male derivatives as in English; the gender is not expressed in English and the choice of the gender of the derivative is completely the translator's choice. Another observation regarding Hebrew is that multiword DMs remain multiword because of the translator choice to relay more on word for word translation, while in Lithuanian there is a tendency to omit the pronoun by using just an inflected verb, and this way, multiword DMs turn into one-word DMs.

5 Conclusions and Future research

The study results showed that English multiword expressions *I think* and *you know*, identified as DMs according to Maschler and Schiffrin (2015) function-based approach, remain stance attitudinal DMs in Lithuanian and Hebrew translation but they demonstrate variability in Lithuanian and Hebrew translations: they are either translated into multiword expressions or one inflected word, or they are completely omitted. In Hebrew translation there is a tendency to use multiword discourse marker translations to express stance, and there is a clear tendency for translators to give preference to male over female derivatives, which is due to the nature of the Hebrew language (Tobin, 2001). However, in Lithuanian, there is a clear tendency observed for one-word DMs in translation. One-word translations mainly include verbs, which, due to Lithuanian being a highly inflected language (Zinkevičius et al., 2005), fully represent the verb-pronoun cases. It should be noted that Lithuanian translations of pronoun-verb multiword expressions and one-word verb cases could be considered almost word-for-word translations. Concerning translation modelling the research reveals stance signalling in discourse preserved as an important element in translation.

More interesting cases include translator choices of particle or connective integration into multiword expressions. The integration of particles for Lithuanian and connectives for both languages might carry the pragmatic meaning that

could have been inferred from the surrounding contexts by the translators (Nau and Ostrowski, 2010a; Blakemore and Carston, 1999; Moeschler, 1989), or translator choices might be also guided by the inner discourse managing system of the target language. The translator's choice to insert particles and connectives needs closer investigation and might be studied in future research. Furthermore, keeping in mind that each language is a unique system with unique features, research could be carried out without English as a pivotal language, which means furthering the current research and using linguistically linked open data (LLOD) and thus accessing related linguistic data directly and comparing the languages. This has already been done for related languages; for example, Snyder et al (2010) analysed Ugaritic (an ancient Semitic language spoken in the second millennium BCE) through resources originally developed for Hebrew. However, linked data provide a sound basis and potential for interoperable resources relating across various languages and enable research across languages and areas.

Acknowledgments

The research described in this paper was conducted in the context of the COST Action CA18209 "Nexus Linguarum" ("European network for Web-centred linguistic data science").

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*, pages 2046–2053.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Michael Barlow. 2011. Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, 16(1):3–44. Publisher: John Benjamins.

Douglas Biber. 2006. https://doi.org/10.1016/j.jeap.2006.05.001 Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2):97–116.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405. Publisher: Oxford University Press.

Douglas Biber, Susan Conrad, and Randi Reppen. 1994. Corpus-based approaches to issues in applied linguistics. *Applied linguistics*, 15(2):169–189. Publisher: Oxford University Press.

Douglas Biber, Stig Johansson, Geoffrey Leech, S. Conrad, Eclwarcl Finegan, and Randolph Quirk. 1999. Longman. *Grammar of spoken and written english*.

Roza Bieliauskienė. 2012. Vilnius–jidiš kalbos Jeruzalė. *Krantai*, (4):56–61.

Diane Blakemore and Robyn Carston. 1999. The interpretation of and-conjunctions. *Iten, C. &*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation*, pages 261–268.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

Tom Cobb and Alex Boulton. 2015. *Classroom applications of corpus analysis*.

David Crystal. 1988. Another look at, well, you know…. *English Today*, 4(1):47–49. Publisher: Cambridge University Press.

Liesbeth Degand and Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. *LOT Occasional Series*, 1:175–199. Publisher: LOT, Netherlands Graduate School of Linguistics.

Robert MW Dixon. 2009. The semantics of clause linking in typological perspective. *The semantics of clause linking: A cross-linguistic typology*, pages 1–55. Publisher: Oxford University Press Oxford.

Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene. *International journal of corpus linguistics*, 22(4):551–582. Publisher: John Benjamins.

Maïté Dupont and Sandrine Zufferey. 2017. Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives. *International journal of corpus*

linguistics, 22(2):270–297. Publisher: John Benjamins.

Bruce Fraser. 2009. An account of discourse markers. *International review of Pragmatics*, 1(2):293–320. Publisher: Brill.

Péter Furkó and Ágnes Abuczki. 2014. English discourse markers in mediatised political interviews.

Sylviane Granger. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1):7–24. Publisher: John Benjamins.

Angela Hasselgren. 2002. Learner corpora and language testing: Smallwords as markers of learner fluency. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 143–174. Publisher: John Benjamins Amsterdam, The Netherlands.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131. Publisher: Elsevier.

Lan Fen Huang. 2011. *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers*. PhD Thesis, University of Birmingham.

John E. Joseph. 2009. Why Lithuanian accentuation mattered to Saussure. *Language & History*, 52(2):182–198. Publisher: Taylor & Francis.

Daniel Joslyn-Siemiatkoski. 2007. The Cambridge history of Judaism: The late Roman-rabbinic period. *Theological Studies*, 68(4):924. Publisher: Sage Publications Ltd.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Science Inst.

Marie-Aude Lefer and Natalia Grabar. 2015. Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures*, 16(2):187–208. Publisher: Akadémiai Kiadó.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. Publisher: Berlin.

Yael Maschler and Deborah Schiffrin. 2015. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 2:189–221. Publisher: Wiley Online Library.

Jacques Moeschler. 1989. Pragmatic connectives, argumentative coherence and relevance. *Argumentation*, 3(3):321–339. Publisher: Springer.

James R. Nattinger and Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford University Press.

Nicole Nau and Norbert Ostrowski. 2010a. Background and perspectives for the study of particles and connectives in Baltic languages. *Particles and connectives in Baltic*, pages 1–37. Publisher: Vilnius: Vilniaus universitetas, Academia Salensis.

Nicole Nau and Norbert Ostrowski. 2010b. Particles and connectives in Baltic.

William R. O'Donnell and Loreto Todd. 2013. *Variety in contemporary English*. Routledge.

Giedre Valunaite Oleskeviciene, Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2018. Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI*, volume 2155, pages 53–58.

Mirna Pit. 2007. Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast*, 7(1):53–82. Publisher: John Benjamins.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Deborah Schiffrin. 2001. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:54–75. Publisher: Wiley Online Library.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Anna Siyanova-Chanturia, Kathy Conklin, and Walter JB Van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776. Publisher: American Psychological Association.

Manfred Stede, Stergos Afantenos, Andreas Peldzsus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1051–1058.

Jan Svartvik. 1980. Well in conversation. *Studies in English Linguistics for Randolph Quirk*, 5:167–177. Publisher: London: Longman.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual Neural Machine Translation with Knowledge Distillation. In *International Conference on Learning Representations*.

Yishai Tobin. 2001. Gender switch in modern Hebrew. *Gender across languages: The linguistic representation of women and men*, 1:177–198.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Naixing Wei and Jingjie Li. 2013. A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18(4):506–535. Publisher: John Benjamins.

Alison Wray. 2012. What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual review of applied linguistics*, 32(1):231–254. Publisher: Cambridge University Press.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Yuqi Zhang, Kui Meng, and Gongshen Liu. 2019. Paragraph-Level Hierarchical Neural Machine Translation. In *International Conference on Neural Information Processing*, pages 328–339. Springer.

Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 365–370.

Sandrine Zufferey and Bruno Cartoni. 2012. English and French causal connectives in contrast. *Languages in contrast*, 12(2):232–250. Publisher: John Benjamins.

Sandrine Zufferey and Liesbeth Degand. 2017. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 13(2):399–422. Publisher: De Gruyter Mouton.

Translation Competence in Machines: A Study of Adjectives in English-Swedish Translation.

Lars Ahrenberg
Department of Computer and Information Science, Linköping University
lars.ahrenberg@liu.se

Abstract

Recent improvements in neural machine translation calls for increased efforts on qualitative evaluations so as to get a better understanding of differences in translation competence between human and machine. This paper reports the results of a study of 1170 adjectives in translation from English to Swedish, using the Parallel Universal Dependencies Treebanks for these languages. The comparison covers two dimensions: the types of solutions employed and the incidence of debatable or incorrect translations. It is found that the machine translation uses all of the solution types that the human translation does, but in different proportions and less competently.

1 Introduction

The performance of today's machine translation systems is sometimes characterized as 'human-level' or achieving 'human parity' (Hassan et al., 2018; Bojar et al., 2018). While claims of this kind have been criticized for not being based on proper evaluations, e.g. by (Graham et al., 2019; Läubli et al., 2020), it is nevertheless a fact that the quality of machine-translated text have improved considerably in recent years, due to new neural models such as the Transformer (Vaswani et al., 2017).

These developments do not only motivate the need for better quantitative evaluations but also call for qualitative evaluations that can pinpoint the differences between different translators, whether human or digital. In this paper my interest is with comparing state-of-the art online translation with human translation of the same source data.

1.1 An example

Comparisons of machine translations with human translations tend to focus either on errors or on general quality criteria such as accuracy and fluency. With the advances of neural machine translation such criteria just seem too blunt to be useful. Neural MT is both accurate and fluent most of the time so any search for differences requires something more fine-grained.

Consider the following English sentence and three possible Swedish translations:

1. He is a bad lier
2a. Han är en dålig lögnare
'He is a bad lier'
2b. Han ljuger dåligt
'He lies badly'
2c. Han är dålig på att ljuga
'He is bad at lying'

All three translations are accurate, fluent and understandable. Still, the first is an example of interference or what (Katourgi, 2020) calls *översättningssvenska*, which we can translate as 'translational Swedish' or 'Swedish translationese'. For this example it means that the translation has the same structure, word by word, as the source sentence, while other, more natural or idiomatic alternatives exist.

Translation (2b) has turned a copulative sentence with a noun phrase predicative into a verb phrase, where the verb translates the noun and an adverb translates the adjective. Translation (2c) is again copulative but involves a head switch; the adjective is translated by an adjective, but this adjective is now the only predicative, having a verb in the infinitive as dependent, this verb translating the English noun.

Katourgi does not say that translations of the type 2a are bad, nor that those of 2b and 2c are better. However, as the title of his book reveals, he claims that they can be too 'noticeable' especially if there are too many of them. The point is thus that a translator should know all available alternatives and not least those that are natural and more

common in an indigenous Swedish context.

The aim of this study is to assess the quality of the translations produced by a state-of-the-art online NMT system at a particular point in time for one language pair, English-Swedish. Compared to Chinese or German, Swedish is a small language, but it still has high-quality MT systems available. The study has two quality aspects in focus: the types of solution the system can produce and to what extent it can apply those solutions accurately.

As for the first part, it is related to taxonomies of what has variously been termed *translation procedures* (Vinay and Darbelnet, 1958/77) or *translation relations* (van Leuven-Zwart, 1989), though in this paper we will call them *solution types*. The second part is an analysis of what we can call debatable solutions or issues (Lommel et al., 2015). This analysis uses a linguistically-based taxonomy of issues, and has been done by the author only. It is however supported with evidence from the human translation.

(Ahrenberg, 2017) concluded for a study on the same language pair and direction that "... the MT is in many ways, such as length, information flow, and structure more similar to the source than the HT. More importantly, it exhibits a much more restricted repertoir of procedures, and its output is estimated to require about three edits per sentence". Here, an edit is caused by an issue that was judged to require alteration. A specific aim of this paper is to see whether these conclusions are still valid.

The data used for this study comes from the English and Swedish Parallel Universal Dependencies treebanks (Zeman et al., 2017). We wished also to investigate the possibilities of using this data set for translation studies, although they were not collected for this purpose.

2 Translation competence

The notion of translation competence has been approached in different ways. One type of analysis seeks to identify all possible properties that are required of a translator. A major proponent of this approach is the PACTE group at the University of Barcelona, who have elaborated a model based on subcomponents in several works, e.g. in (PACTE, 2011). Others, like (Malmkjær, 2009), have argued that a characterization of translation competence should focus on those factors that distinguish it from any other profession, including any other type of bilingualism. From this perspective the only subcomponent of the PACTE model that Malmkjær finds relevant is the transfer competence:

> "... the ability to complete the transfer process from the ST (source text) to the TT (target text), i.e. to understand the ST and re-express it in the TL (target language), taking into account the translation's function and the characteristics of the receptor".

Nevertheless, when PACTE studies how translation competence is acquired they put translation problems in focus, in particular what they call Rich Points, i.e., passages that may be challenging to translate. In this work I take translation competence to mean the ability to find an appropriate solution for all words in the source text, where a solution may of course be to leave it untranslated. Thus, the set of 'problems' includes not just difficult ones, but all words or constructions meeting a given criterion. This prevents the selection of rich points from being skewed and allows for quantitative analysis. All in all this should give a better picture of the abilities of a translator. The chosen construction is adjectives in relation to a head.

We can interpret produced translations in terms of skills that we ascribe to the translators, whether human or digital. Necessary skills are specified in many text books on translation. Here, I mention a few of them from a text book by (Ingo, 2007)[1]

- a robust sense of style in the target language
- active and creative language skills in the target language
- familiarity with target language genre conventions
- ability to express oneself naturally in the target language
- ability to change style in accordance with the style of the source text
- possession of an imaginative mind so as not be bound by the patterns of the source text and hindered from finding the natural and idiomatic expressions of the target language

We can observe that this list demands of a translator to be able to strike a delicate balance between stylistic and genre-related conventions on the one hand, and creativeness and imaginative abilities on the other.

[1] Translations by the author.

2.1 Translation problems

NMT models make mistakes of different kinds. Even (Hassan et al., 2018) who claims human parity performed an error analysis of the English output and found most of the errors in the categories Incorrect words, Ungrammatical, Missing words, and Named Entities. Thus, both accuracy and fluency are affected. Looking at the same sentences (Läubli et al., 2020) observe that fluency mistakes (word order, ungrammaticality, ...) are still common in the machine translations, somewhat contrary to the expectations that NMT systems have specifically improved as regards fluency. They also observe that cross-sentential constraints affect machine translations more often than human translations. This means that a sentence translation can appear fluent in isolation, but be judged as inappropriate in the document context.

(Ahrenberg, 2017) found that the most frequent errors in his data related to the accuracy of word translations. In that study, close to 50% of the errors noted were of this kind. The next error type in frequency concerned morphological form, slightly less than 25% of all.

2.2 Comparing machine translation and human translation

There have been quite a number of studies trying to distinguish machine translations from human translations by automatic means e.g., (Aharoni et al., 2014). These studies usually favour features that can be detected automatically as well, such as the occurrence of common function words and part-of-speech ngrams. (Nguyen-Son et al., 2017), found other features, such as word distributions, complex phrase constructions, and the occurrence of phrasal verbs to be helpful, in particular when combined with coherence features across sentences or within whole paragraphs. The data used in this study consist of isolated sentences, so we cannot employ coherence features. However, we can compare our approach to what can be gained from other features.

There have also been studies aimed at automating, at least partly, the recognition of solution types, or divergencies as they are often called, for example (Deng and Xue, 2017; Zhai et al., 2019). This was an initial aim also of this study, but was abandoned for reasons that will be explained below.

3 Adjectives in English-Swedish translation

Adjectives have very much the same behaviour in English and Swedish. They can be modifiers/attributes, predicatives, heads of arguments such as subjects and objects, conjuncts and be part of lexicalized phrases. All of these functions are actually found in the English source data. The distribution of these functions in the source data is shown in Table 1 together with simple examples to show what is meant.

Function	Frequency	Example
modifier	956	a *red* shirt
predicative	122	it is *red*
conjunct	42	white and *red*
argument head	34	help the *poor*
lexphrase	16	*at best*
Total	1170	

Table 1: Adjectival functions in the source data.

Given that the vast majority of adjectives can be translated straightforwardly by an identical syntactic construction, it can be expected that this pattern will be over-used by inexperienced translators and by machine translations that tend to prefer frequent patterns over rarer ones. Thus, when the adjective and head noun are independent semantic units that form a complex that can be interpreted compositionally, the unmarked translation is a word-for-word translation, in particular if both items are part of the core vocabulary of the language. This applies to the first four examples of Table 1. The lexphrase also has a standard translation, though one which is not compositional: *i bästa fall*, 'in (the) best case'.

For some English adjectives a common alternative translation is to form a compound. This happens with *wooden – trä-*, *main*, where *huvud-* is a common choice, and *special* with the translations *speciell*, *särskild* or *special-*. Examples are: *wooden table – träbord*, *main purpose – huvudsyfte*, *special unit – specialenhet*. This solution type is actually quite common in the studied data set.

Another possibility is that the adjective and the noun form a single designation of some referent, which acts as a term or name for the referent. This requires that the translator knows this and also is able to find out the term or name used in the target language. Common results then are compounds,

red herring – avledningsmanöver, 'distraction action', transfers, such as *British Council* or *American Express*, or a lexicalized phrase, where the translation of the adjective may also be an adjective, but one that does not occur outside of that phrase, as in *common sense – sunt förnuft*, 'sound sense'.

Swedish has a greater propensity than English for using adjectives as heads of nominal phrases. Thus, a nominal head in an English source text is sometimes not translated. *white people* may be translated by *vita människor* but just *vita* 'whites' would do just as well, if not better. The word *one* is often not translated when it is used instead of repeating a mentioned noun, or when the referent is understood from the context: *the only one – den enda*, *they will build a new road and tear* **the old one** *up. – de ska bygga en ny väg och riva upp* **den gamla**.

Yet another possibility is that the pair of adjective and noun are part of a larger construction that acts as a unit in the translation. It may simply be that a preceding preposition gives the whole an adverbial function and the possibility to translate the whole thing with an adverb. Examples:

> in *early* morning
> *tidigt* på morgonen
> 'early in the_morning'
>
> She was killed *in cold blood*
> Hon mördades *kallblodigt*
> 'She was_murdered coldblooded-ly'

The relevant embedding construction may also be larger:

> at your *earliest* convenience
> *så fort du kan*
> 'as fast you can'

If the embedding construction is found superfluous for the target audience, it may not be translated at all, and this will then affect the adjective-noun pair in the same way. With this as background we now proceed to the study.

4 Data

The source sentences for the study are taken from the English part of the Parallel Universal Dependencies treebanks (PUD)[2]. These treebanks were created for a shared task on multilingual parsing from raw text (Zeman et al., 2017). The sentences are taken from news and Wikipedia articles, but only a few from each article. Thus, there may be lexical overlaps but no coherent paragraphs. This means that we cannot study discourse phenomena such as cohesion.

PUD-segments were translated into Swedish outside of the shared tasks. The Swedish translations follow the same directions as for other PUD treebanks, namely that "Translators were instructed to prefer translations closer to original grammatical structure, provided it is still a fluent sentence in the target language" (*ibid. p.4*). This requirement is one that we could also ask of a machine translation system.

Only those sentences where English is the source language have been used. They amount to 750 segments. We define an adjective as any token assigned the UD part of speech ADJ in the English PUD treebank. There are 1170 of them.

The machine translations were produced by Google Translate on 25-26th of February, 2021. They were then tagged and parsed with the UDPipe tools (Straka, 2018) using a model for UD_Swedish-Talbanken.

Basic statistics for the data can be found in Table 2. The figures follow a standard pattern for English-to-Swedish translations. It can be noted that the Type-Token-Ration for the machine translation is much closer to the human translation than to the English source text.

Dataset	Types	Tokens	TTR
English PUD	4714	15840	0.297
Swedish PUD	5125	14432	0.355
MT-translated PUD	4949	14129	0.350

Table 2: Statistics of the datasets.

The human translations have earlier been provided with manual word alignments by the author. We hoped that the structural properties of the image of an adjectival relation could be determined automatically from the word alignment. This approach, however, turned out to be problematic as the annotations for part-of-speech and dependencies are not harmonized across the two languages. The translation of many words, such as 'many' and 'same' that were tagged ADJ in the source treebank, were translated in the expected, standard fashion, but had a different tag (PRON

[2]https://github.com/UniversalDependencies/UD_English-PUD/

and DET for these words), causing the automatic analysis to suggest a part-of-speech shift. Similarly, a reference such as the 'Metropolitan Club' has been translated verbatim and is thus word-to-word. However, where the English annotates 'Metropolitan' as an adjective modifying a noun, the Swedish sees two proper nouns, where the second is a dependent of the first via the UD relation *flat*. The automatic analysis thus suggests a shift of parts-of-speech and a reversal of the dependency. Cases of this kind abound, and for this reason the sorting and the analysis have required more manual effort than anticipated. Thus, all data points have had a manual review and the same holds for the machine translations.

5 Analysis

We compare a machine translation with human translations of 750 English sentences which are part of the Parallel Universal Dependencies (PUD) dataset. We analyse translations along two dimensions, solution types based on structural properties and issues.

5.1 Solution types

The solution types are divided into two major classes, **Isomorphisms** and **Restructurings**

A translation is an isomorphism if the following properties hold: (1) the adjective is translated by a single token, a; (2) the head token is translated by a separate single token, h; (3) h is the head of a in the translation; (4) a and h have the same part-of-speech as their source tokens and the dependency relation and the order between them is also the same. It may be the case that the distance between a and h is different than the distance between the corresponding source words. These differences are not directly caused by the adjective and its head and so are not considered relevant.

Restructuring is an umbrella term for all other situations. We sub-classify restructurings according to the structural effect. Table 3 gives examples of each category from the corpus.

A *shift* occurs when the first three clauses above hold, but there is a change in part-of-speech and/or relation. Using the dependency relations of the UD framework, a change in part-of-speech will almost always involve a shift of dependency relation as well, so we will note a relation shift only when there is no change in part-of-speech. An example is when the dependency of an adjective is changed from 'xcomp' (head of a subject-less verb phrase) to 'ccomp' (head of a finite clause with subject).

An *omission* occurs when the adjective in the source sentence lacks a corresponding target token. This means as well that there is no corresponding dependency either. In case the head has not been translated we use the label *head-omission*.

A *convergence* occurs when the adjective and its head are mapped onto the same target token, or the same set of target tokens. The opposite situation, a *divergence*, happens when either the adjective or its head is aligned with two or more target tokens, so that the single edge of the source tree is mapped on some subgraph with two or more edges in the target tree.

A *head-shift* occurs when both the adjective and its head are aligned with single tokens, but the dependency relation is reversed, i.e., h will be a dependent of a. This category is different from the category of *head changes*, which means that both the source adjective and its head have been translated, but they are no longer related as a dependent to a head. Finally, an *order-reversal* means that the order between a and h is reversed in comparison with the order between their source words.

From Table 4 we see that the human translation is more prone to restructure than the machine translation. In fact, this difference is consistent across all of the five adjectival functions shown in Table 1. However, the difference is not so great as to be statistically significant at a 0.05 critical level using a Chi-Square test with one degree of freedom.

Also, for some 43% of all instances (506 out of 1170) HT and MT have produced identical translations, see Table 5. The large majority of these cases are isomorphisms and the tokens concerned are common lexical items with more or less standard translations, such as *first – första, many – många, , new – nya, other – andra, possible – möjliga, whole – hela*. Another set of adjectives for which translations are shared are words with a common historical root, or words that Swedish has borrowed from English, such as *artificial – artificiell, civil – civil, international – internationell, military – militär, popular – populära*. In 85% of the instances (1003 out of 1170) the two translations agree on the broad type of solution, and in most of these (941 out of 1170) they also agree on the sub-type.

Category	English	Swedish
	Isomorphisms	
modifier	the *peaceful* transition	den *fredliga* övergången
predicative	this will be a little *different*	kommer detta bli lite *annorlunda*
	Restructurings	
convergence	The *South Korean* company initially thought...	Det *sydkoreanska* företaget trodde ...
divergence	over 70 % are *alive*	mer än 70 % var *vid liv*
omission	*provincial* police surveillance operations	polisens övervakningsoperationer
headomission	of new ideas with *old ones*	som nya idéer bildade med *gamla*
head shift	*preferential*$_1$ *access*$_2$ to government	*företräde*$_1$ i regeringens *tillgänglighet*$_2$
head change	*Much*$_{nsubj:5}$ ... has been about *identity*	*Mycket*$_{nsubj:3}$ ... har handlat om *identitet*
shift of POS	the protein ... that's *responsible*$_{ADJ}$	det protein ... som *ansvarar*$_{VERB}$ för
shift of deprel	I'd be *amazed*$_{root}$ if	Jag skulle bli förbluffad$_{xcomp}$
order-reversal	in the realm of the *unimaginable*	i det *ofattbaras* rike

Table 3: Different types of solutions.

System	Isom	Restr	Total
MT	943	227	1170
HT	878	292	1170

Table 4: Distribution of isomorphic and restructured solutions for MT and HT.

Criterion	Isom	Restr	Total
Token identical	455	51	506
Type identical	827	176	1003
Sub-type identical	825	116	941

Table 5: Number of identical translations between MT and HT.

The largest differences between the two systems concern the use of restructurings, as shown in Table 6. When we look at these more fine-grained sub-types of restructurings, there are several cases where the two translations choose the same type of solution, convergence being the most common. Examples are found with geographical adjectives such as *South Korean – sydkoreansk, northern Sami – nordsamiska*

While the two systems agree to a large extent in the use of convergences, the case is quite different with divergences. The human translation employs this solution type four times as often as the machine translation. The same is true, though to a lesser degree, of part-of-speech shifts and headshifts. A possible explanation is that the human translator has a better sense of what fluency or naturalness means for the target language.

The system, on the other hand, has a greater use of omissions, although for quite a small percentage of the full dataset. It also produces more of head changes where the direct connection between dependent and head in the source is broken up in the translation.

Type	MT	HT
convergence	120	110
divergence	17	70
headchange	15	6
headshift	5	12
headomission	3	6
omission	22	8
posshift	46	75
reversal	0	2
relshift	2	3
Total:	230	292

Table 6: Distribution of different types of restructurings in MT and HT.

5.2 Issues

For issue classification we use the taxonomy shown in Table 7. It is basically structured according to which linguistic level is affected. The label *Meaning* means that one can debate the accuracy of the choice. It includes cases that (Hassan et al., 2018) label Incorrect words, but also what they call Unknown words, a category which is only rarely found in their translations. However, in our machine translations they are quite common, to be further discussed below. *Word choice* means that we may discuss whether the chosen word in the translation is the best choice. It is less serious than the previous category. The label *Morphol-*

ogy means that there is a lack of congruence between the adjective and its head, or of any one of them in relation to another related token such as a determiner. *Grammar* means that the translation has produced an ungrammatical substring that includes the adjective or its head. *Style* is similar to Word choice but in relation to grammar. Thus, the grammar is ok, but there are perhaps better solutions, i.e., a different type of solution could be preferred. Finally, *Orthography* concerns spellings and the use of capital letters, etc.

Issue	Frequency	Same as HT
Meaning	54	2
Word choice	118	25
Morphology	26	4
Grammar	8	1
Style	30	5
Orthographic	3	0
Total:	239	40

Table 7: System solutions that may be debated according to linguistic levels.

First it should be said that the table shows that issue classification is a subjective process. At least one person, i.e., the translator responsible for Swedish PUD, can be assumed to accept the system translations as they coincide with those of her own. However, we can note with some relief that the issue type where the differences are most pronounced concerns Meaning. For this reason we look at this category in more detail.

5.3 Problems with accuracy

Looking further at the issues pertaining to Meaning, they can largely be divided into three classes: (i) innovations, where the system seems to make up words, probably based on its models of subwords; they may sometimes be understood nevertheless; (ii) mistranslations, where the translation may mislead the reader but is perfectly fluent; and (iii) odd mistranslations that affect both accuracy and fluency and probably will cause the reader to stop for a while and try to infer what is meant.

The innovative solutions produces words that either don't exist in Swedish, or have alternatives that are vastly more common. The following are a few examples:

- 'villainous' is translated as *skurkig*, 'crooky' instead of *skurkaktig*, 'like a crook'.
- 'skerry-protected waterway' is translated as *skärvägsskyddad*, where the human translator found herself forced to rewrite as *av skär skyddad*, 'by skerries protected'.
- 'isthmus' is translated as *landmus*, 'land mouse' instead of the correct *näs*.
- 'zodiacal' is translated as *zodiakal*, a word which exists but is uncommon.

To this list we may add a few cases where the English words are copied into the Swedish translation: 'glitchy, twitchy Odi' is left as such where the human translator provides normal Swedish words. We can observe that the source adjectives in these cases are quite rare; in fact none of them can be found, even at the C2 level, in the English Vocabulary Profile. This means that even as a proficient speaker of English as a second language you are not expected to know them.

The second subset is made up of plain mistranslations, sometimes yielding the opposite of what was in the source as when 'uncooperative' is translated as *samarbetsvillig*, 'cooperative'. References to centuries are a problem; the system sometimes gets it right as with '16th century' becoming *1500-talet*, but mostly gets it wrong; for example with the '6th', '8th' and '14th' centuries.

Inconsistencies are found also with other adjectives of nationality, so that 'Macedonian' is translated either as *makedonisk*, as in the HT, or *makedonsk*.

In the third type of situation the system's choice is just odd, making you wonder what is actually meant. Some examples of this kind are:

- 'the dress code was too stuffy': the HT says *stel*, 'stiff', which is correct, whereas the system says *täppt*, which would be appropriate if you were talking about someone's nose.
- 'skilled jobs' is rendered as *skickliga arbeten*, with an adjective that is appropriate for a 'skilled worker'. The HT has the correct *kvalificerade*.
- 'lower forgone earnings' was translated as *nedre förlorade inkomster*, where *nedre* is appropriate for positions and geography but cannot be applied to earnings.

5.4 Problems with compounds

The system is very happy at producing convergencies and normally does so quite accurately. But sometimes it is overdoing it, producing clumsy compounds such as *Obama-*

specialassistent, 'Obama special assistant', *lagstiftningsförlamning*, 'legislative paralysis', *Post-Classic-perioden*, 'the Post Classic period'. It may also pick an unfortunate translations for one of the parts of a compound, as in *södersamiska* instead of *sydsamiska* for 'south Sami'.

It also occassionally separates the parts of a compound which results in a breach of grammar: *storstads kommun*, 'a city's municipality' instead of *storstadskommun* for 'metropolitan municipality' or *ras tolerans*, 'a race's tolerance' instead of *rastolerans* for 'racial tolerance'.

6 Conclusions

A general conclusion is that the system seems to have improved, for example compared to (Ahrenberg, 2017). It has gained in the type of solutions it has available, and in creativity, but this has come with a price. As regards translation competence with respect to the translation of adjectives, it can be summarized as follows:

- The system is more prone than the human translator to choose an isomorphic solution. The tendency is consistent across grammatical functions;
- The system uses the same types of restructurings as the human translation, but to different degrees;
- In particular, the human translation employs divergencies, part-of-speech shifts, and head shifts to a much larger extent than the system;
- On the other hand, the system shows more of head changes and omissions than the human translations;
- As shown in Table 4, the system produces some 200 debatable translations including about 50 (4.3% of all) that can be considered errors of accuracy.
- (Not surprisingly) the system has the greatest problems with uncommon words. For these words the system often produces innovative solutions, probably on the basis of its subword models. However, this means that the system essentially lacks the competence to distinguish words from non-words.

It is interesting to note that the large restructurings that were illustrated in the introduction are rare. There is one example on the model of sentence (1), where someone is described as 'a keen guitarist'. Both human and machine chose a word-by-word translation in spite of the fact that there is no Swedish word that exactly corresponds to 'keen'. The human translator chose *flitig*, 'diligent, hard-working', and the system chose *skicklig*, 'skilled', none of which is optimal. A more idiomatic way of expressing the meaning of 'keen' in Swedish would be to use a verb such as *gilla* or *tycka om*, both meaning 'like'. The human translator was instructed to stay close to the source, so that may be an explanation for not chosing a major rewriting; the system, however, has no awareness of the directive.

As for the use of PUD treebanks to study differences between human and machine translations there are both pro's and con's. On the positive side, the sentences contain both common and uncommon words and thus provides a nice sample of problems for translation across frequency ranges. The same is actually true of grammar so there are a good number of 'Rich Points' that can be selected. On the downside from the point of view translation studies is the fact that the resource consists of isolated sentences, so that discourse effects cannot be studied. Another drawback is that the annotations of the English and Swedish treebanks are not harmonised. This can partly be explained by differences in annotation practices, and partly by parsing errors that have not been corrected. Even though I had made a complete alignment at the word level for all sentences, attempts to automate the categorisation of solution types failed because of the inconsistencies. Similarly, UDpipe was helpful for tagging the machine translations, but also makes many parsing errors.

In future work, the study can be extended to dependencies of nouns and verbs using a similar approach. And the study of adjectives can be repeated at a future date.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection och machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), June 23-25 2014*, pages 289—295, Baltimore, ML, USA. Association for Computational Linguistics.

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 21–28, Varna, Bulgaria.

Ondřej Bojar, Christian Federmann, Mark Fishel,

Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Dun Deng and Nianwen Xue. 2017. Translation divergences in chinese–english machine translation: An empirical investigation. *Computational Linguistics*, 43(3:521—565.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Rune Ingo. 2007. *Konsten att översätta: översättningens praktik och didaktik*. Studentlitteratur.

Alexander Katourgi. 2020. *Svenskan går bananer: En bok om översättningar som syns*. Lys förlag.

Kitty van Leuven-Zwart. 1989. The relation of translation to original. *Target*, 1(2):100–101.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2015. Multidimensional quality metrics. `http://www.qt21.eu/mqm-definition/definition-2015-12-30.html`.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *JAIR*, 67:653–672.

Kirsten Malmkjær. 2009. What is translation competence? *Revue française de linguistique appliquée*, XIV:121–134.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T. Tieu, Huy H. Nguyen, Junichi Yamagishi, and Isao Echi Zen. 2017. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1504–1511.

PACTE. 2011. Results of the validation of the pacte translation competence model: Translation problems and translation competence. In *Methods and Strategies of Process Research: Integrative Approaches to Translation Studies*, pages 317–343. John Benjamins.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

J.-P. Vinay and J. Darbelnet. 1958/77. *Stylistique comparée du français et de l'anglais*. Paris, Didier.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. 2019. Towards recognizing phrase translation processes: Experiments on english-french.

Polarity in Translation: Differences between Novice and Experts across Registers

Ekaterina Lapshinova-Koltunski[1] **Fritz Kliche**[2]
Anna Moskvina[2] **Johannes Schäfer**[2]
[1]Saarland University [2]University of Hildesheim
e.lapshinova@mx.uni-saarland.de kliche@uni-hildesheim.de
moskvina@uni-hildesheim.de johannes.schaefer@uni-hildesheim.de

Abstract

Translation can obscure the subjectivity of the sources and flatten down positive and negative aspects. Thus, we perform an explorative analysis of translation in terms of sentiment properties focusing on the differences between student and professional translations of various registers. However, we do not compare translations with their sources, but analyse polarity items in two translation variants from the same text sources. We propose a multi-step analysis to investigate the distribution of polarity items and report on small experiments on a corpus of English to German translations to identify the lack of experience in translation by students. Our results show that pragmatic differences expressed in the usage of polarity words is highly dependent on the register a text belongs to. Following this, we identify registers, such as popular-scientific articles, where students translate sentiment using more and heavier polarity words.

1 Introduction

Most computational studies of translationese[1] concentrate on the analyses of lexico-grammatical, morpho-syntactic and textual language patterns ignoring semantic and pragmatic properties (Baroni and Bernardini, 2006; Volansky et al., 2015). However, multilingual computational sentiment studies show that textual sentiment, e.g. positive and negative aspects, as well as subjectivity, are altered and even vanished in translation (Mihalcea et al., 2007; Balahur and Turchi, 2014; Salameh et al., 2015; Mohammad et al., 2016). These features are linked to pragmatic competence of translators that can vary depending on their level of expertise. Moreover, pragmatic aspects and the related translation competence may also vary across textual registers as novice and professional translators have different degrees of register sensitivity as shown by Lapshinova-Koltunski (2020) and Redelinghuys (2016).

In the present paper, we analyse sentiment-related properties of English-German translations that were produced by translators of different levels of expertise. We concentrate on the distribution of positive and negative polarity items across different registers[2] translated either by students or by professionals. Although the sentiment of the source texts would bring us interesting insights, we are constrained to exclude them, as the required comparable analytical resources[3] are missing at the moment. Therefore, we concentrate on the analysis of variation in translation in terms of polarity properties. Our data contains student and professional translations of the same sources – texts belonging to various registers. We aim to identify differences in the polarity of the two translation varieties and analyse if these differences are subject to register settings. We expect that student and professional translators alter the sentiment of the originals differently, which should be reflected in the different use of the sentiment lexicon in their translations. On the one hand, as students are repetitive in their lexical choices (as shown by Kunilovskaya et al. (2018) and Redelinghuys (2016) a.o.), we might observe their overuse of certain words which follows in higher or lower sentiment of their translations. On the other hand, their lack of register sensitivity (see Bizzoni and Lapshinova-Koltunski, 2021; Redelinghuys, 2016, for details) may cause a more

[1]Linguistic characteristics of translations showing their differences from non-translated texts (Gellerstam, 1986; Baker, 1993).

[2]We understand register as contextual text variation which is reflected in distinctive distributions of linguistic patterns (Biber, 1995).

[3]This kind of analysis requires comparable polarity lists for English and German.

levelled use of sentiment lexicon in different registers.

We perform an explorative analysis of translation in terms of polarity, focusing on specific differences between professional and student translations of various registers.

2 Main Concepts and Related Work

We understand sentiment analysis as determining the polarity of a piece of text as positive or negative and measure it with the help of polarity items – negative or positive words. This approach is a type of lexicon-based sentiment analysis (Taboada et al., 2011).

As sentiment is not always similarly marked in the source and in the target, translations do not always preserve the original sentiment (Salameh et al., 2015; Mohammad et al., 2016), which was also shown for machine translation (Troiano et al., 2020). Although we measure polarity of the target texts only, we deal with translation, a product of multilingual communication. Therefore, our work is also related to multilingual sentiment analyses that have mainly addressed mapping sentiment resources from one language onto another (e.g. Mihalcea et al., 2007; Balahur and Turchi, 2014). Contrastive studies show pragmatic differences between English and German (Kranich, 2016; House, 2006) that have impact on sentiment realisation in both languages, as it was shown by Taboada et al. (2014) in the analysis of evaluative language and by Fronhofer (2020) in the analysis of emotions. The latter study points to specific language preferences in the morpho-syntactic realisation of emotions (their parts-of-speech, tenses, etc.).

Knowing about these cross-lingual contrasts, we expect translators to adapt a text's sentiment to the target language preferences. Without sufficient experience in doing so, students may fail in appropriate choices for polarity transformations or their lexico-grammatical settings. Munday (2012) shows in a study on translating attitude that students have difficulty because of the missing knowledge on lexico-grammatical features of both the source and the target language. Another study of student translations reveal their missing pragmatic competence (Pisanski Peterlin and Zlatnar Moe, 2016). Interestingly, students showed more difficulties in transferring structures that had no direct translation equivalent with similar lexico-grammatical patterning, as novice translators frequently translate word-by-word. Therefore, we should also expect variation in our data in terms of lexico-grammar, i.e. morpho-syntactic types of polarity items.

3 Methodology

In this section, we introduce the features we extract from the sentiment analysis (Section 3.1), outline the used data set (Section 3.2) and tools (Section 3.3) with our analysis methods in Section 3.4.

3.1 Features

Building upon existing studies in sentiment and translation, we formulate a number of features to analyse polarity in student and professional translations. Our aim is to find lexical differences between student and professional translators. Therefore, we don't use a classifier which would yield sentiment scores for whole texts. Instead, as the first step of our pipeline we extract sentiment words using the list SentiWS (Remus et al., 2010) containing weighted negative and positive items. We formulate the following features:

Overall polarity. 1. the total number of positive polarity words per text (`Pos`), 2. the total number of negative polarity words per text (`Neg`), 3. the sum of weights of positive polarity items (`SumWeightedPos`), 4. the sum of weights of negative polarity items (`SumWeightedNeg`).

Morpho-syntactic subtypes of polarity items. 5-7. Distribution of positive polarity nouns, adjectives and verbs (`PosN, PosV, PosA`), 8-10. Distribution of negative polarity nouns, adjectives and verbs (`NegN, NegV, NegA`), 11-13. Proportion of positive polarity nouns, adjectives and verbs calculated against the total number of nouns, verbs and adjectives, respectively (`PosNprop, PosVprop, PosAprop`), 14-16. Proportion of negative polarity nouns, adjectives and verbs calculated against the total number of nouns, verbs and adjectives, respectively (`NegNprop, NegVprop, NegAprop`).

3.2 Data

We use a dataset of German texts translated by both professional and student translators from English (PT – professional translations and ST – student translations), representing translation variants of the same original texts. These texts cover the following registers: political essays (ESSAY), fiction (FICTION), manuals (INSTR),

popular-scientific articles (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). Professional translations were exported from the CroCo corpus (Hansen-Schirra et al., 2012), whereas the student translations come from the corpus VARTRA (Lapshinova-Koltunski, 2013). The main difference between the two variants in our data is the degree of expertise – professionals have a good degree of experience in translating, whereas students are trainees with little experience in translating. The whole data set contains 102 texts (51 for each translation variant) with 272,195 tokens in total (more details are given in Table 1).

	ST	PT
ESSAY	15,794	15,595
FICTION	12,549	11,226
INSTR	19,866	20,718
POPSCI	22,692	19,739
SHARE	24,739	24,450
SPEECH	24,303	23,373
TOU	19,687	17,464
TOTAL	139,630	132,565

Table 1: Dataset size in tokens.

3.3 Sentiment Analysis in Geist

The data is pre-processed and analysed using Geist[4] (Kliche, 2020), a web tool for converting text data in different formats[5] into formats required by applications in the Digital Humanities context, e.g. topic modeling or stylometric analyses. For the present study, the SentiWS list and the pipeline to extract the features detailed in Section 3.1 were integrated into Geist. Using the TreeTagger (Schmid, 1994), the texts are tokenised and labeled with part-of-speech tags. When one or two tokens left to a sentiment word is a negation, the polarity swaps from negative to positive or vice versa. Geist analyses each of the 102 translations separately and creates a CSV file containing the features for each document. The student translations contained in sum 139,630 Tokens (122,715 words), 8,088 of which were positive and 2,138 were negative. The texts of professional translators consist of 132,565 tokens (116,086 words), with 7,613 positive and 2,103 negative words.

[4] https://geist.uni-hildesheim.de

[5] Including PDF, RTF, Open Office or Microsoft Office formats.

3.4 Explorative and descriptive analyses

As our aim is to exploratively analyse translations and find specific differences between professionals and students, we decide for several techniques that include Correspondence Analysis (CA, Greenacre, 2007), Hierarchical Agglomerative Clustering (HC, Rokach and Maimon, 2005) and boxplots.

Correspondence Analysis. CA allows us to explore relations between features and subcorpora in our data. With the help of this explorative technique, we identify which subcorpora have similarities or differences and how these differences correlate with the selected features. For our purposes, we intend to find groupings of subcorpora based on either the experience of translators or the register a text belongs to. The feature distributions across the subcorpora are used to measure Weighted Euclidean distances, termed the χ^2 distances. The distances are represented in a two-dimensional graph. The larger the differences between the subcorpora, and also between the subcorpora and features (dots and triangles in Figure 1), the further apart they are on the graph. The dimensions are computed in such a way that any subset of *k* dimensions accounts for as much variation as possible in one dimension, the first two principal axes account for as much variation as possible in two dimensions, and so on. The length of the feature arrows indicates associations between subcorpora and features: the longer the line, the stronger is the association.

Clustering. In the next step, we perform HC on texts using the 'strongest' features resulting from CA. With this technique, we investigate whether texts cluster according to registers or according to the level of expertise in translation. To be consistent with the previous analysis, we use the Euclidean distance and performed Ward's linkage to calculate the distance between new clusters on a condensed distance matrix. In each iteration, two clusters that have the smallest distance are merged together, until every text is linked into a dendrogram. The order of the initial clusters (texts we used for the analysis) represented by features that we want to analyse has little significance, the distance between clusters increases with each merging iteration and the height of each merge gives the distance between two clusters.

Boxplots. In the final step, we use boxplots to more closely observe the discovered specific dif-

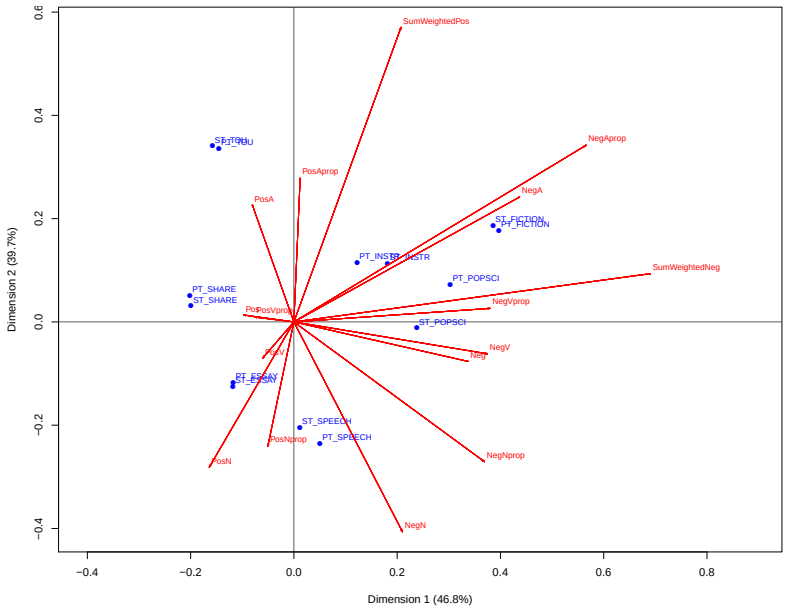

Figure 1: CA for all subcorpora with all features, dimensions 1 and 2.

ferences between professional and student translations. Boxplots are helpful to visually represent summary statistics (central tendency values and spread of data) and to compare descriptive statistics across groups.

4 Analyses

Correspondence analysis. We perform CA on the expertise and register-based subcorpora using the whole set of features defined in Section 3.1. Figure 1 presents the resulting two dimensional graph. The differences between the plotted subcorpora and features can be interpreted on both axes (dimensions 1 and 2 that explain 86,5% of the data variance). Here most student and professional subcorpora of the same registers group together. Dimension 1 (x-axis) separates translations of letters-to-shareholders (leftmost), tourism leaflets and political essays from political speeches, instructions, popular science and fiction (rightmost). Almost all negative polarity features seem to contribute to this division, as the feature arrows show positive values in the direction of the x-axis, with `SumWeightedNeg` being the most contributing feature. Interestingly, its counterpart `SumWeightedPos`, is not opposing (i.e. pointing into the opposite direction), but rather contributes most to the other breakdown in our data – the division of subcorpora observed along the y-axis (dimension 2). Here again, most of the observed groupings are register-based, except for popular science. This is the only difference between professional and student translations uncovered with CA in our data. This means that there is more variation in terms of register than experience in our data, with some text registers being more similar between each other than the others. As the features `SumWeightedNeg` and `SumWeightedPos` were found to contribute the most in determining the differences in texts, they were used for further analysis with clustering and box plots.

Clustering. We use the two features, `SumWeightedNeg` and `SumWeightedPos`, contributing most to the variation along the two dimensions discovered in the previous analysis step rather then using all of the features. This allows us to further target the differences in texts, based on the particular use of positive and negative words within different registers and translation variants. The resulting dendrograms are given in Figures 4 and 5 in Appendix, with x-axis containing texts and y-axis representing the distance.

The dendrogram based on `SumWeightedPos` visualises two major clusters, where the smaller cluster consists mostly of texts from the registers TOU, SHARE and FICTION, with student and professional translations being equally linked together. Most of the texts from other registers can be found in the second major cluster. Deeper towards the leaves of the tree, translation variants of the same text within a register are linked earlier (the distance

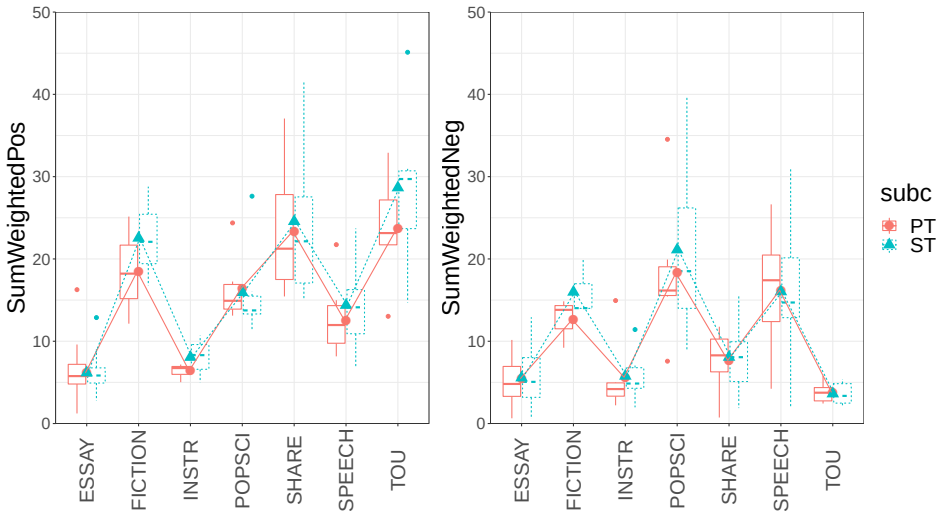

Figure 2: Polarity item weights at text level across registers in professional and student translation.

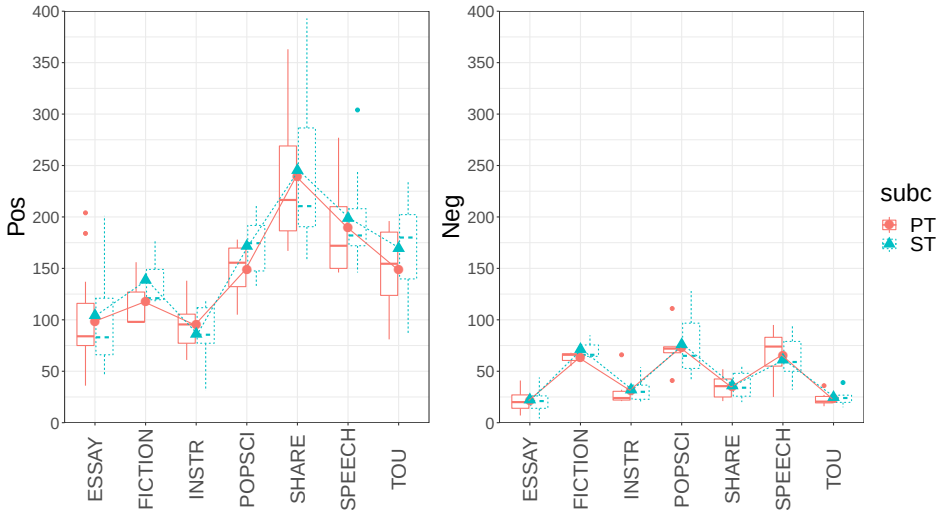

Figure 3: Polarity item distribution at text level across registers in professional and student translation.

between such clusters is smaller), which is later followed by linking of texts from the same register: most of the essay texts end up being linked together before being linked to texts from the register SPEECH and INSTR.

The dendrogram based on `SumWeightedNeg` also has two distinctive clusters, though the texts are more equally distributed across these two (in comparison to the previous dendrogram). Likewise with `SumWeightedPos`, the closest distance is found within the register, though not necessarily within the variants of translation: we can find examples like PT-ESSAY_004 and PT-ESSAY_009 being linked together earlier then the corresponding translation variant from students. Moreover, though the earliest clusters tend to belong to the same register, the registers are interwoven together as distance grows. The results of clustering confirm the observation from CA: Translation variants are highly similar in terms of sentiment features and differences are observed for groups of registers only.

Boxplots. We use boxplots (Figure 2) to directly compare student and professional translations across registers in terms of the two selected features. We observe more variation between professional and student translations when analysed across registers. As seen from the plot for `SumWeightedPos`, student translations of most registers are more positive than the professional ones, except in ESSAY and POPSCI. However, the differences do not seem to be significant in most cases, except for fictional texts and instructional manuals. The plot for `SumWeightedNeg` reveals

EO	*Using this self-administration setup and related techniques, researchers mapped the regions of the brain that mediate addictive behaviors and discovered the central role of the brain's reward circuit.*
ST	*Mithilfe dieser Selbstverabreichungsmethode und ähnlichen Methoden haben Forscher die Regionen im Gehirn lokalisiert, die das* ***Abhängigkeit****sverhalten steuern. Zudem hat sich herausgestellt, dass das Belohnungssystem im Gehirn eine zentrale Rolle bei der Bildung einer* ***Abhängigkeit*** *spielt.*
PT	*Mittlerweile haben Hirnforscher die am Drogenmissbrauch beteiligten Gehirnregionen kartiert. Sie kennen heute die zentrale Funktion des Belohnungssystems dabei.*

Table 2: Example illustrating the difference between student and professional translations (ST and PT), as well as the original English source (EO).

that fictional and popular science texts are more negative when translated by students. The variation of negative weights within the POPSCI texts translated by students is also remarkable pointing to heterogeneous negativity of these translations.

We also compare the overall distribution of positive and negative words in student and professional translations to discover a slightly different view (see Figure 3). Instructions translated by students contain less positive words (although being more positive). Students use more positive words in the POPSCI translations than professionals, although the overall positive polarity of both translation variants of this register remain similar. All this points to the differences in the lexicon choices by students and professionals.

A glance at the data confirms this as well: the negative polarity noun *Abhängigkeit* occurs 24 times in the student translations of POPSCI, whereas professionals use this word 5 times only. Table 2 contains an example from our corpus illustrating the observed differences in translation and showing that students (ST) are more repetitive in their lexical choices also because their translations are longer and more explicit.

5 Conclusion and Discussion

We performed explorative analysis of polarity in translations that differ with regard to the level of translators' expertise. The variation discovered in our data turned to be more register-related, than expertise-related. However, differences between student and professional translations could be observed within registers and register groupings. This points to dependency of pragmatic differences in translation on the functional text variation – the register a text belongs to.

Students use more and heavier polarity words in certain registers only. Moreover, they seem to show similar register sensitivity as professionals do, as their translations also vary in terms of polarity features, which is against our expectations.

In future, we plan to perform a more detailed analysis of distinct features. We also intend to investigate differences between the polarity vocabularies used by both groups of translators, as preliminary insights show that students tend to repeat the same words. Moreover, a cross-lingual comparison involving the sources' analysis would be an asset, which, however, requires comparable polarity lists for English and German.

References

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.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1):56 – 75.

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.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Surprising professionals and conventional students: structural register diversification and convergence in translation. In *Proceedings of the NoDaLiDa-2021*.

Nina-Maria Fronhofer. 2020. *Emotion Concepts in Context - A Contrastive Analysis of English and German Discourse*. Ph.D. thesis, Universität Augsburg.

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.

Michael J. Greenacre. 2007. *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Juliane House. 2006. Communicative styles in english and german. *European Journal of English Studies*, 10:249–267.

Fritz Kliche. 2020. *Die Erschließung heterogener Textquellen für die Digital Humanities*. Dissertation, Universität Hildesheim.

Svenja Kranich. 2016. *Contrastive Pragmatics and Translation*. John Benjamins Publishing, Amsterdam.

Maria Kunilovskaya, Natalia Morgoun, and Alexey Pariy. 2018. Learner vs. professional translations into Russian: Lexical profiles. *Translation and Interpreting*, 10.

Ekaterina Lapshinova-Koltunski. 2013. VARTRA: A Comparable Corpus for Analysis of Translation Variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria. ACL.

Ekaterina Lapshinova-Koltunski. 2020. Tracing normalisation and shining through in novice and professional translations with data mining techniques. In Sylviane Granger and Marie-Aude Lefer, editors, *Translating and Comparing Languages: Corpus-Based Insights*, pages 45–59. Presses universitaires de Louvain, Louvain-la-Neuve.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic. ACL.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Int. Res.*, 55(1):95–130.

Jeremy Munday. 2012. *Evaluation in Translation: Critical Points of Translator Decision-making*. Evaluation in Translation: Critical Points of Translator Decision-making. Routledge.

Agnes Pisanski Peterlin and Marija Zlatnar Moe. 2016. Translating hedging devices in news discourse. *Journal of Pragmatics*, 102:1 – 12.

Karien Redelinghuys. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: a corpus-based study. *Stellenbosch Papers in Linguistics*, 45(0):189–220.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a publicly available German-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta.

Lior Rokach and Oded Maimon. 2005. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 767–777, Denver, Colorado. ACL.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307. ACL.

Maite Taboada, Marta Carretero, and Jennifer Hinnell. 2014. Loving and hating the movies in english, german and spanish. *Contrastive Linguistics*, 14:127–161.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online).

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

Appendix

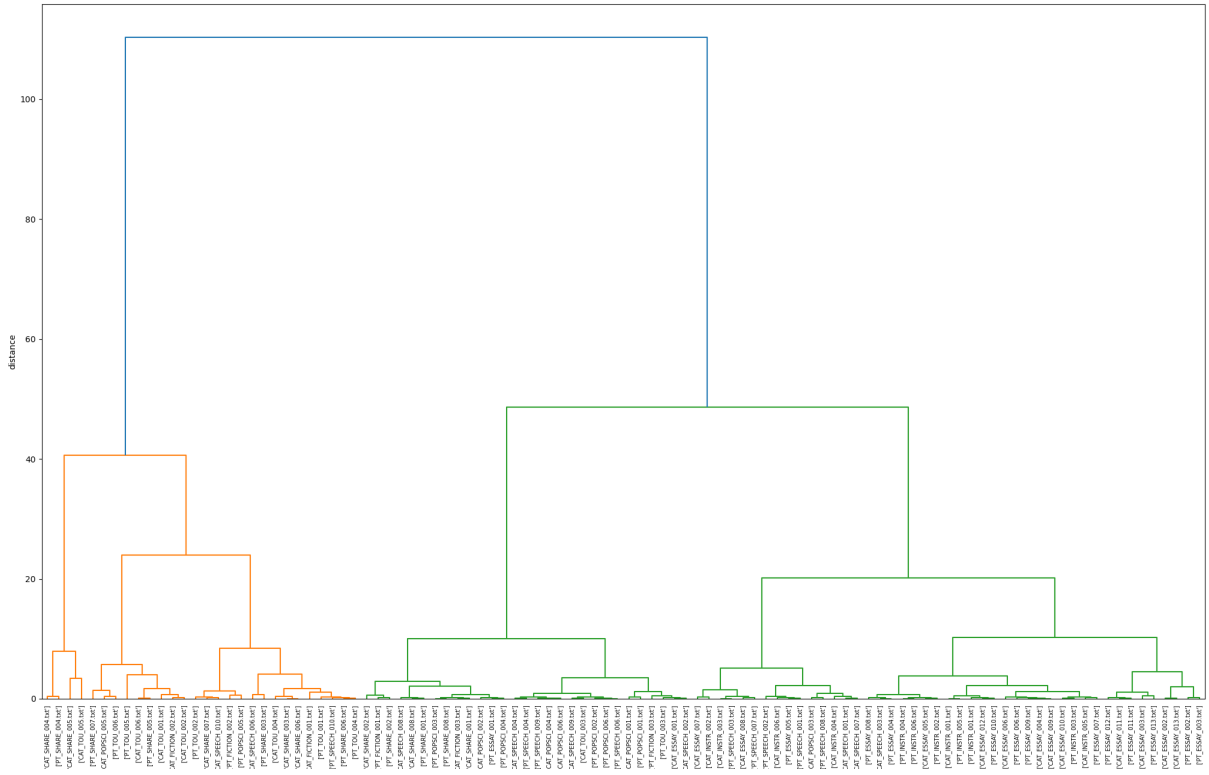

Figure 4: HC for SumWeightedPos.

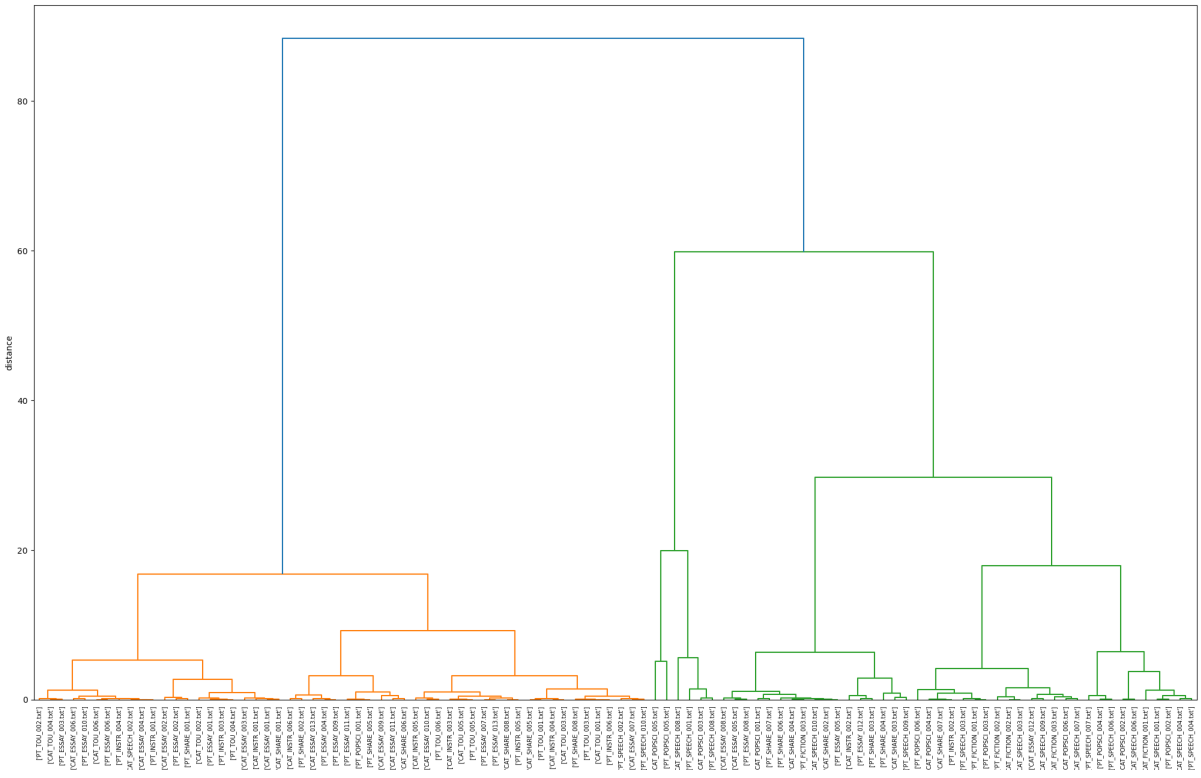

Figure 5: HC for SumWeightedNeg.

Introducing a Word Alignment Dissimilarity Indicator: Alignment Links as Conceptualizations of a Focused Bilingual Lexicon

Devin Gilbert
CRITT dg@devrobgilb.com

Michael Carl
CRITT mcarl6@kent.edu

Abstract

Starting from the assumption that different word alignments of translations represent differing conceptualizations of cross-lingual equivalence, we assess the variation of six different alignment methods for English-to-Spanish translated and post-edited texts. We develop a word alignment dissimilarity indicator (WADI) and compare it to traditional segment-based alignment error rate (AER). We average the WADI scores over the possible 15 different pairings of the six alignment methods for each source token and correlate the averaged WADI scores with translation process and product measures, including production duration, number of insertions, and word translation entropy. Results reveal modest correlations between WADI and production duration and insertions, as well as a moderate correlation between WADI and word translation entropy. This shows that differences in alignment decisions reflect on variation in translation decisions and demonstrates that aggregate WADI score could be used as a word-level feature to estimate post-editing difficulty.

1 Introduction

Alignment error rate (AER) is a segment-based metric that compares one alignment (usually automatically generated) against another gold standard word alignment, assigning errors when the hypothesis alignment's links differ from those of the gold standard (Och and Ney, 2003). It is a normalized score with values between 0–1 for entire segments where a score of 0 indicates identical word alignments and a score of 1 indicates completely different sets of alignment links. When reported, AER scores are usually multiplied by 100 for readability. Usually an average AER score over many segments is reported, and automatic alignment systems have ranged between average AER scores of 3.7–50.6 (Liu et al., 2010) and 14.5–33.2 specifically for the English-to-Spanish language pair (Lambert, 2008). We have conducted alignment experiments with six different alignments of the same English-to-Spanish translations (a total of 1045 segments, 25936 tokens, translated by 31 participants), two manual alignments (M1, M2) and four automatic alignments (A1–A4).[1]

M1 is the original manual alignment (Mesa-Lao, 2014) which was later amended by another group of researchers. M2 is a realignment done by a group of researchers with very specific alignment criteria and, above all, the stipulation that only one aligner would sign off on the alignment of all translations for a given text in order to ensure consistency. A1 was aligned with GIZA++ (Och and Ney, 2003), trained on almost 2M en-es Europarl segments. A2 was aligned with SIMALIGN Match, A3 with SIMALIGN Argmax, and A4 with SIMALIGN Itermax (Sabet et al., 2020).

We obtain average AER scores between 8.8 and 26.3 (see Table 1). Perhaps not surprisingly, the lowest alignment scores (< 10.0, i.e., the most similar alignments) are between two automated alignment systems (A2-A4 and A3-A4) while the alignment scores between the two human alignments, M1 and M2, average out to 14.6. Also note that A4 is the automatic system that comes closest to human alignment M2 as well as human alignment M1. These scores give us a measuring stick with which to optimize word alignments, but we argue that word alignment links could be a much

[1] All data is publicly available on the CRITT website (Center for Research and Innovation in Translation and Translation Technology). For these alignments' study IDs in the CRITT Translation Process Research Database (TPR-DB), see Appendix A

Pairing	AER score	Pairing	AER score
M1-M2	14.6		
M1-A1	26.2	M2-A1	22.0
M1-A2	25.7	M2-A2	19.5
M1-A3	26.3	M2-A3	19.3
M1-A4	23.8	M2-A4	18.0
A1-A2	25.0	A2-A3	10.4
A1-A3	24.9	A2-A4	9.5
A1-A4	23.7	A3-A4	8.8

Table 1: Cumulative AER scores for six different alignments

richer source of information if we examine them on a more granular level than is afforded by AER. Instead of seeing dissimilarities between different alignment methods as errors, we suggest thinking of different word alignments as instantiations of a different contextualized and focused bilingual lexicon which may dynamically emerge in a translator's mind during the translation process.

We take it that alignment links are probabilistic in nature and that chances of two different alignment methods (human or machine) generating the exact same alignment links for any given segment are extremely slim. If we term a segment's set of alignment links an "alignment configuration," then a translation with m source words and n target words allows for 2^{m*n} unique alignment configurations. Think of a segment's alignment space as a grid where each source word is a row and each target word is a column. If a square of the grid is filled in, this represents an alignment link. The different possible patterns on this grid are an alignment configuration. A sentence with 10 source and 10 target words has 2^{100} (1.267e30) different possible alignment configurations[2]; finding the exact same alignment configuration on a segment level is not very likely.

Additionally, AER is usually reported for entire texts; an averaged AER score may be computed based on thousands of word alignments. While much effort has gone into developing systems to decrease global averaged AER (GIZA++, Och and Ney (2003); SIMALIGN, Sabet et al. (2020); FASTALIGN, Dyer et al. (2013); UALIGN, Hermjakob (2009); etc.), we posit that the agreement—as well as the disagreement—about alignment relations on the level of individual words carries crucial information about translation difficulties.

While some words may be 'easy' to align—i.e., with little or no discrepancies between different alignment methods—translational equivalents for other words may be disputable or 'controversial,' resulting in differences between different methods. In statistical MT, alignment links carry information about an underlying, contextualized bilingual dictionary. Along this line of thinking, differing alignments of the same translations represent differing conceptualizations of translational relations between words or phrases. Moreover, differing conceptualizations of translation equivalence point to potential discrepancies and difficulties, and therefore variation in alignment links could potentially be used as an indicator of ambiguity and translation difficulty.

In this paper we investigate differences among alignment links between individual source words and their translations as produced by different alignment methods. We posit that dissimilarities between different alignment methods are indicators for translational choice and difficulty, and we correlate variation in alignment links with other measures of translation difficulty such as production time. The next section discusses our method of calculating alignment error rate at the word level.

2 From segment to word alignment scores

Word alignments are commonly represented as sets of tuples, where each tuple represents one source-target alignment link. The first value in each tuple is the ordinal number of a token in the source segment; the second value in each tuple is the ordinal number of a token in the target segment to which the source word is linked. Figures 1 and 2 show two different alignment configurations of the same translation.

[2]This includes 'incomplete' phrase alignment with missing alignment links. Assume, for instance, the phrase translation {have bread ↔ Tengo pan} (see Figures 1 and 2). This should result in a set of alignment links {(1,0),(1,1),(2,0),(2,1)}. However, without further post-processing, MOSES' phrase-based system *grow-diag-final(-and)* may produce 'incomplete' phrase alignments in which one of the four alignment links may be missing, resulting in five possible alignment configurations for this phrase translation.

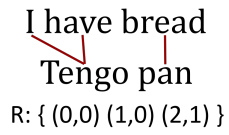

Figure 1: Reference

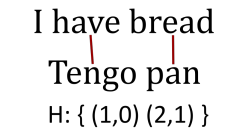

Figure 2: Hypothesis

Source sentence (S):	I	have	bread
$\text{WADI}_{\{0,1,2\}}$(R,H):	[1,	0,	0]

Table 2: WADI for 3-word source sentence

The first (reference) alignment configuration has two alignment links which connect the first and second source word ('I' and 'have') to the first word in the translation ('Tengo'), while the third source word ('bread') has a single alignment link that ties it to the second word in the target ('pan'). The set of hypothesis alignments consists of the same links with the exception to the first source word, which is unaligned. While the sets of reference and hypothesis alignments agree with respect to the translation equivalence of 'bread' and 'pan,' they differ on whether or not 'I' is conceptualized as being part of 'Tengo'—Spanish, as a pro-drop language, would also allow 'yo tengo'.[3]

$$\forall_{i \in S} \text{WADI}_i(\text{R,H}) = |\text{R}_i \cup \text{H}_i - \text{R}_i \cap \text{H}_i| \quad (1)$$

In order to assess the (dis)agreement between two alignment methods, we compute a word alignment dissimilarity indicator (WADI_i) that indicates the number of diverging alignment links for each source word position i. The WADI score (see Equation 1) takes as arguments the set of reference tuples (R) and hypothesis tuples (H) (see Figures 1 and 2) and produces a list that contains a WADI_i for every source word i which indicates the number of mismatches between the reference and the hypothesis. Table 2 shows the list of WADI results for the example in Figures 1 and 2 in which the first position corresponds to the first source word 'I' and a $\text{WADI}_i = 1$. For the two other positions ('have' and 'bread'), $\text{WADI}_i = 0$.

3 Examples of Alignment Dissimilarity

Here are some examples of high WADI scores that we have calculated between the M1 and M2 alignment methods.

Example 1:

Source
His withdrawal comes in the wake of fighting flaring up again in Darfur and *is set to* embarrass China ...
Target
Su retiro se produce a raíz de la lucha que surge de nuevo en Darfur y *tuvo lugar con el objetivo de* avergonzar a China..."

ST	W	M1	M2
is	4	tuvo lugar con el objetivo de	tuvo lugar
set	3	tuvo lugar con el objetivo de	con el objetivo
to	4	tuvo lugar con el objetivo de	de

Table 3: Alignment Dissimlarities in Example 1

One half of a segment from our English-to-Spanish data collection is shown in Example 1.[4] The respective alignments of M1 and M2 of the sub-segment "is set to" ↔ "tuvo lugar con el objetivo de" are shown in Table 3, together with their WADI scores (W in the Table). As the example shows, M1 aligns the ST and TT in a more compositional manner than M2. M1 linked 'is' as part of a three-word alignment group "is set to" and aligned it with a large target alignment group, "tuvo lugar con el objetivo de", which is repeated for each ST word in Table 3. M2, however, aligned more compositionally: {is ↔ tuvo lugar}; {set ↔ con el objetivo} and {to ↔ de}. WADI scores will be higher if one alignment method produces larger alignment groups than the other, as shown in Table 3.

Example 2 shows how alignments can have similarly long alignment groups yet high WADI scores because of the different conceptualizations of what these long alignment groups are equivalent to in translation.[5]

[3]Note that, according to our assumption above, this translation allows for $2^{3*2} = 64$ different alignment configurations, in which every ST word could or could not be paired with any TT word.

[4]Extracted from Participant 29's translation of segment 3 of multiLing Corpus Text 3. The text deals with Steven Spielberg not participating in the Beijing Olympics to protest China's backing of Sudan.

[5]Extracted from Participant 10's translation of segment 3 of multiLing Corpus Text 4. The text covers the topic of climate change and developing countries.

Example 2:

Source
Some of the most vulnerable countries of the world have contributed the least to climate change, *but are bearing the brunt of it*.
Target
Algunos de los países más vulnerables del mundo son precisamente los que menos han contribuido al cambio climático, a pesar de que *precisamente son algunos de los que más lo sufren*

The source word 'are' has a high WADI score of 5 because M1 aligns it by itself with the target word 'son'. This might seem to be a perfectly valid way to conceive of equivalence between the source and target, but when considering the other tokens in the surrounding phrase, we see that the M2 alignment also has a valid way of conceiving of the links of equivalence for this translation (see Table 4).

ST	W	M1	M2
but	1	a pesar de	pesar de que
are	5	son	precisamente son algunos de los que
bearing	6	algunos de los que más lo sufren	sufren
the	6	algunos de los que más lo sufren	más
brunt	6	algunos de los que más lo sufren	más
of	6	algunos de los que más lo sufren	lo
it	6	algunos de los que más lo sufren	lo

Table 4: Alignment Dissimlarities in Example 2

The source text in the last clause of this segment features a null subject due to a subtle bit of anaphora ('...but are bearing...'); the subject is inferred from the first part of the segment. The translator, in striving to create a target text that sounds natural in Spanish, has explicitated the subject by writing, 'son algunos de los que [are some of those that].' While M1 has this circumlocution aligned together with the verbal phrase 'bearing the brunt of it' yet separately from the verb 'are', M2 has treated this long stretch of text in Spanish as the argument of this clause and aligned the entire phrase together with 'are' while splitting the last verbal phrase into 'bearing', 'the brunt', and 'of it'. Additionally, M2 does not leave any target tokens unaligned, whereas M1 leaves 'que' and 'precisamente' unaligned. This example demonstrates how high-WADI tokens tend to "flock together" around longer alignment groups: all of the source tokens from this last verbal phrase, 'bearing the brunt of it,' have WADI scores of 6.

AER					
0	0–10	10–20	20–30	30–40	>40
% 13.4	28.5	27.4	18.4	8.4	3.9
WADI					
0	1	2	3	4	5+
% 76.1	11.8	7.7	2.5	1.1	0.8

Table 5: AER and WADI (M1 & M2) Distribution Pattern

4 WADI for different Alignment Methods

Table 5 compares the distribution patterns of WADI's word-level scores and AER's segment-level scores, as calculated between the M1 and M2 alignment methods. For WADI (M1-M2), the range of values is between 0 and 11. AER is a continuous variable whereas WADI is essentially categorical (ordinal) since it is only possible to have scores that are whole numbers. While calculated in a similar way, WADI and AER are quite different in how they are shaped. For example, WADI scores of zero are highly common in our data, while it is more rare to get AER scores of zero. Both have distributions that are skewed to right, but AER's distribution is much more even than WADI's. Let us compare what each metric shows us about our alignment methods. Figure 3 shows the relation between AER scores using M1 and M2 as references and A1 to A4 as hypotheses. Average AER scores show the automatic methods to be less similar to the manual methods than the manual methods are to each other (see Figure 3).

We can also see that when M1 is used as refer-

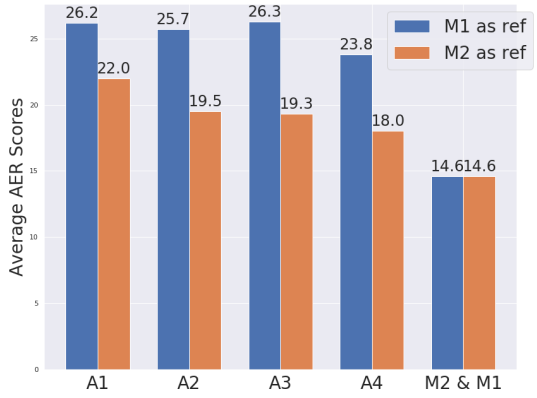

Figure 3: Average AER by Alignment Method

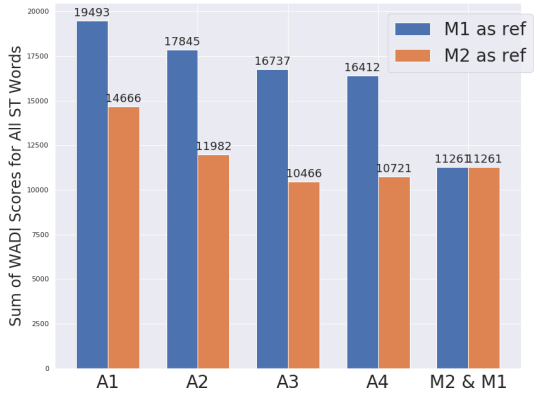

Figure 4: Sum WADI by Alignment Method

ence, all automatic alignment methods are nearly identical. However when M2 is used as reference, all automatic methods' average AER scores drop by 4.2–7.0, and A4 is 4.0 (rather than 2.4) lower than A1 and approaches the AER score of M1 (see Figure 3). These results would suggest that A4 would be the best automatic alignment method with respect to the gold standard.

When we consider the sum of WADI scores for each method instead of AER, we can see that the drop in dissimilarities persists when using M2 as reference instead of M1, but we can also see that both A3 and A4 (sum WADI scores of 10466 and 10721 respectively) are more similar to M2 than both manual methods are to each other (sum WADI score of 11261; see Figure 4). As opposed to the AER results, the WADI results suggest that A3 would be the best automatic alignment method with respect to the gold standard.[6]

[6]There are, of course, other considerations that go into what might be the "best" automatic alignment method for a particular use-case. For example, A3 leaves a lot more tokens unaligned than the other automatic alignment methods, which could make it less suitable for preparing data for translation process research or for using WADI scores as a quality estimation feature.

However, rather than thinking of the different sum WADI scores of each alignment method as an indication of the "best" alignment, we can also think of these WADI scores as measures of how much two alignment methods differ with regard to their conceptualization of a focused bilingual lexicon made up of all the words in the source and target text. If we take each source word to be the source-text component of one "entry" in a bilingual dictionary, then we can take the WADI scores as measures of agreement with respect to the target-text component(s) of that entry. Taking the example from Figures 1 and 2 in Section 2, the WADI score of 1 for 'I' shows the dissimilarity of the two alignments as to whether 'I' has a relationship of partial equivalence with the Spanish word 'tengo', and thus a different conceptualization of the bilingual lexicon.

Following this line of thinking, we can take the WADI data displayed in Figure 4 and conclude that methods A2–A4 all agree with M2 in their conceptualizations of our texts' bilingual lexicon about the same amount that M2 and M1 agree with each other. It is remarkable that automatic alignment methods agree with a human gold standard to the same degree that another human alignment agrees with this gold standard.

5 Examining WADI

Let us examine how WADI scores are distributed relatively by word class. Figure 5 shows a relative distribution for WADI scores of 0, 1, 2, and 3 or greater for the following word classes: adjective (Adj), adverb (Adv), function words (Func), nouns (N), numbers (Num), prepositions and conjunctions (PC), punctuation and other symbols (Sign), verbs (V), and wh-words (Wh). Figure 5 shows WADI scores for the two human alignments M1 and M2. It shows that adverbs and verbs are the least agreed-on (only 67% of adverbs and 71% of verbs have WADI scores of zero), suggesting that verbs and adverbs may be conceptualized differently in the bilingual translation lexicon more often than other word classes. On the other hand the alignment of punctuation and wh-words are the most agreed-on (91% and 88%, respectively, have WADI scores of zero), indicating that these items and their corresponding translations are less prone to dissimilar conceptualiza-

tions. Also, a surprisingly low share of function words have WADI scores of zero (73%).

We already observed some examples of verbs with high WADI scores with Examples 1 and 2 in Section 3. These examples showed how verbs in larger alignment groups had high WADI scores. The fact that verbs exhibit lower alignment agreement is consistent with research showing that verbs also tend to have significantly higher translation entropy values than other word classes (Ogawa et al., 2021), which indicates that translators tend to vary more when translating verbs. This could suggest that there is an association between variation in translation solutions and variation in how translations get aligned, which we test in Section 6.

Function words include determiners (e.g., 'the', 'a', 'this'), pronouns (e.g., 'they', 'he', 'their'), and the word 'to.' It makes sense that function words exhibit less agreement in alignment because the presence of these words across translations of the English-Spanish language pair is often asymmetrical, which would lead to function words tending to be aligned in larger alignment groups rather than by themselves, and this will tend to cause disagreements among aligners as to which neighboring words these function words get grouped with. This seems to be the case since the mean size (length in words) of target alignment group (TAGnbr) for M1 is 1.39, whereas mean TAGnbr for M2 is 1.06. Conversely, it makes sense for punctuation and wh-words to have high levels of agreement in alignment because there do tend to be clear-cut equivalents across languages for these two word classes, at least for the English-Spanish language pair.

Figure 6 plots relative shares of tokens belonging to different target alignment group sizes (TAGnbr), by word class. It shows TAGnbr figures from the M2 alignment method. Compared to most word classes, function words (Func) and punctuation/symbols (Sign) have a very high share of one-word target alignment groups (about 90% and 98%, respectively). This means that function words and symbols have less multi-word alignments. On the other hand, nouns have a large share of two-word alignment groups (over 30%; see Figure 6). It is also interesting to note that adverbs tend to be unaligned more often than other word classes (they have the highest share of target alignment groups of zero).

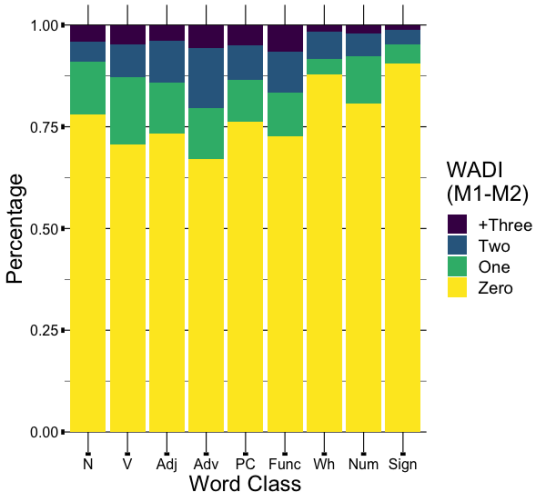

Figure 5: WADI Scores (M1-M2) by Word Class

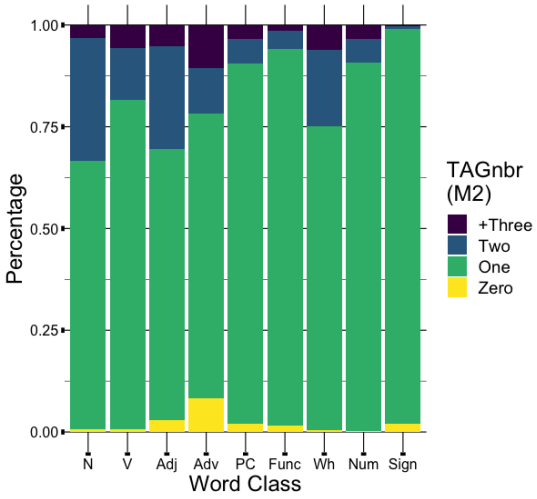

Figure 6: Target Alignment Group Size (M2) by Word Class

Comparing WADI scores (Figure 5) and the size of the target group by source word class (Figure 6), some interesting observations can be made. Function words exhibit alignment dissimilarity that is disproportionate to their high share of single-word alignment groups. This can be explained by the difference in mean TAGnbr between M1 and M2 that we discussed above. Another example: even though English nouns are more often linked to two Spanish target words, the alignment agreement seems to be a relatively uncontroversial; nouns have a higher share of zero WADI scores than verbs, adjectives, adverbs, function words, and prepositions/conjunctions (about 77%; see Figure 5). This could simply be due to the fact that many nouns occur in multi-word phrases yet are fairly straightforward to align because their trans-

lations have an easier-to-identify relationship of equivalence.

The correlation between WADI scores (M1-M2) and TAGnbr from the M2 alignments is significant yet extremely weak (Spearman $\rho(25934) = .08$, $p < .001$). However, the correlation between the same WADI scores and TAGnbr from the M1 alignments is remarkably stronger (Spearman $\rho(25934) = .55$, $p < .001$). This would seem to indicate that a great deal of the alignment differences that WADI (M1-M2) indicates are due to the discrepancies between TAGnbr for the M1 and M2 alignments.

6 Aggregating WADI across alignment methods

We calculate WADI scores for all 15 possible pairings of our six alignment methods and calculate the mean of these 15 different WADI scores for each source word. We investigate how this averaged value correlates with word-level translation process and product metrics such as production duration, insertions, and word translation entropy (HTra).

There is a positive, significant correlation between average WADI scores and log-transformed production duration per word, $r(25934) = .18$, $p < .001$ (see Figure 7), which is similar to the correlation between AER and production duration per segment (Spearman $\rho(1043) = -.11$, $p < .001$). There is also a positive, significant correlation between average WADI scores and number of insertions, $r(25934) = .28$, $p < .001$ (see Figure 8). Here we see a relationship between average WADI scores and behavioral indicators of translation effort which suggests that average WADI scores could be used as indicators for word-level quality estimation.

We also found there to be a positive, significant and moderate correlation between average WADI scores and HTra $r(25934) = .40$, $p < .001$ (see Figure 9). This demonstrates the relationship between the variation in alignment decisions (even among the four automatic alignment methods) and variation in translation. This evidence from production duration, insertions, and HTra leads us to conclude that aggregate WADI scores can be used as an indicator of translation (and post-editing) difficulty. That is, average WADI scores over several different alignment methods might be used to estimate post-editing difficulty on the word level.

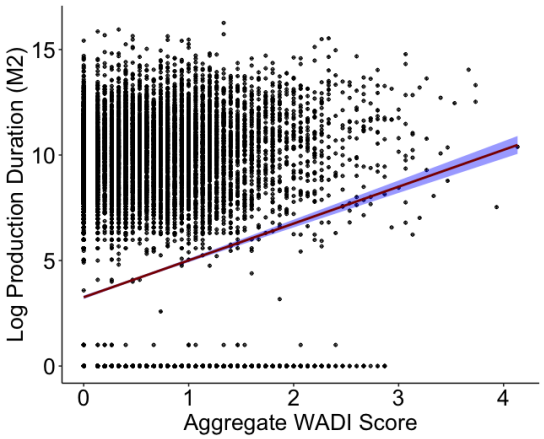

Figure 7: Scatterplot: Average WADI Scores and Log Production Duration

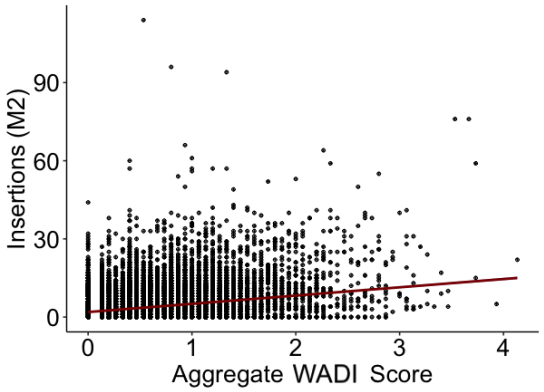

Figure 8: Scatterplot: Average WADI Scores and Number of Insertions

7 Conclusion

There are many ways to conceptualize equivalence in translation. We hypothesize that aligning translations is itself an act of declaring a bilingual focused dictionary, and different alignment relations represent differing possible conceptualizations of translation equivalents. We have developed a metric that, given two word alignments of the same translation, operationalizes dissimilar conceptualizations at the word level: word alignment dissimilarity indicator (WADI).

We observe that some word classes, such as verbs and adverbs, are more prone to dissimilar alignment conceptualizations while other word classes, such as wh-words, numbers, and punctuation/symbols are relatively uncontroversial in alignment. We also observe that size of alignment groups is related to word alignment dissimilarity,

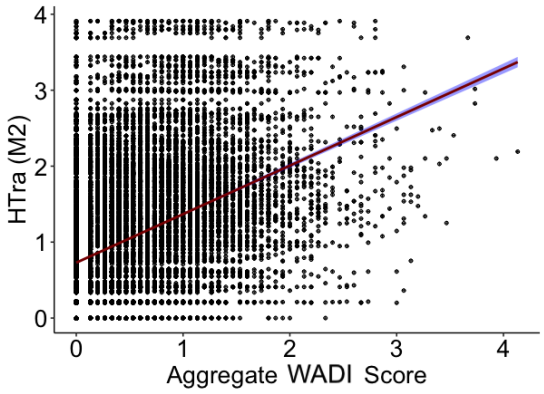

Figure 9: Scatterplot: Average WADI Scores and HTra

which shows that the fundamental conceptualization of what the source or target component of a unit of translation is could explain much of the observed variation in WADI.

Dissimilarities in word-to-word alignment between humans—but also between automatic alignment systems—of the same translations correlates with increased variation in the translation options produced by humans (i.e., aggregate WADI scores correlate with HTra), and we also observe a tendency for increased translation/post-editing effort—as indicated by production duration and number of insertions—to increase with WADI scores.

The observation that word alignments of different human annotators diverge substantially, and sometimes more than some automatic alignments differ from human alignments, suggests that there is no one gold standard for alignment relations. Rather, it stipulates that different conceptualizations of the same translation are possible and valid. Variation in translational conceptualization, however, has been shown to indicate translation difficulty and post-editing effort. The WADI score might captures some of these difficulties.

References

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Ulf Hermjakob. 2009. Improved Word Alignment with Statistics and Linguistic Heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 229–237, Singapore. Association for Computational Linguistics.

Patrik Lambert. 2008. *Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation*. Doctoral Dissertation, Universitat Politècnica de Catalunya, Barcelona.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Bartolomé Mesa-Lao. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors, *Post-editing of Machine Translation: Processes and Applications*, pages 219–245. Cambridge Scholars Publishing.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51. Publisher: MIT Press.

Haruka Ogawa, Devin Gilbert, and Samar A. Almazroei. 2021. redBird: Rendering Entropy Data and ST-Based Information Into a Rich Discourse on Translation: Investigating relationships between MT output and human translation. In Michael Carl, editor, *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications. Springer.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. In *EMNLP (Findings) 2020: arXiv:2004.08728 [cs]*, Online. ArXivLabs.

A Alignment Methods

Table 6 gives the CRITT TPR-DB study IDs for the six alignment methods used in this study.

M1	BML12	A2	BML12_SM
M2	BML12_re	A3	BML12_SA
A1	BML12_giza	A4	BML12_SI

Table 6: CRITT TPR-DB Study IDs for all Alignment Methods

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).

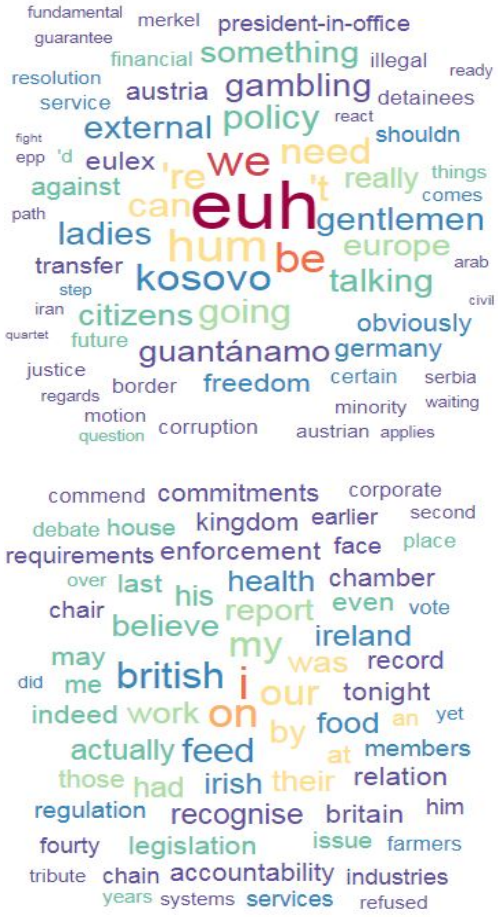

Figure 1: SI based on ORGsp (top), ORGsp based on SI (bottom) (source language: German). Item color denotes relative frequency (relF) (red=high relF, blue=low relF), item size denotes KLD score (large=high KLD, small=low KLD)

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.

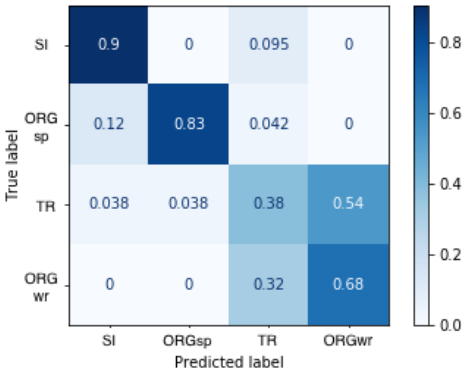

Figure 2: Confusion matrix using the 400 most informative words and PoS n-grams.

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.

Automatic Classification of Human Translation and Machine Translation: A Study from the Perspective of Lexical Diversity

Yingxue Fu
School of Computer Science
University of St Andrews
KY16 9SX, UK
yf30@st-andrews.ac.uk

Mark-Jan Nederhof
School of Computer Science
University of St Andrews
KY16 9SX, UK

Abstract

By using a trigram model and fine-tuning a pretrained BERT model for sequence classification, we show that machine translation and human translation can be classified with an accuracy above chance level, which suggests that machine translation and human translation are different in a systematic way. The classification accuracy of machine translation is much higher than of human translation. We show that this may be explained by the difference in lexical diversity between machine translation and human translation. If machine translation has independent patterns from human translation, automatic metrics which measure the deviation of machine translation from human translation may conflate difference with quality. Our experiment with two different types of automatic metrics shows correlation with the result of the classification task. Therefore, we suggest the difference in lexical diversity between machine translation and human translation be given more attention in machine translation evaluation.

1 Introduction

The initial interest in and support for machine translation (MT) stem from visions of high-speed and high-quality translation of arbitrary texts (Slocum, 1985), but machine translation proves to be more difficult than initially imagined. In recent years, progress has been made in MT research and development, and it is claimed that MT achieves human parity in some tasks (Wu et al., 2016; Hassan et al., 2018; Popel et al., 2020). However, these statements are challenged by other researchers and remain open to debate (Läubli et al., 2018; Toral et al., 2018; Toral, 2020).

The typical automatic approach to evaluating MT is to compare a machine translated text with a reference translation. The assumption is that the closer a machine translation is to a professional human translation, the better it is (Papineni et al., 2002). Automatic metrics for MT are developed based on this assumption. Human translation (HT) is treated as gold standard and the deviation from it is transformed into a measure of translation quality of MT.

Many studies have shown that translated texts are different from originally written texts (Baroni and Bernardini, 2006; Ilisei et al., 2010). The typical method used for the identification of translationese is automatic classification of translated texts and originally written texts (Baroni and Bernardini, 2006). There are some studies that compare translation varieties such as professional and student translations and post-edited MT (Kunilovskaya and Lapshinova-Koltunski, 2019; Toral, 2019; Popović, 2020). While surface linguistic features and simple machine learning techniques are capable of classifying translated texts and originally written texts with high accuracy, it is difficult to use the same method to classify translation varieties, with the accuracy being barely over the chance level (Kunilovskaya and Lapshinova-Koltunski, 2019; Rubino et al., 2016).

When comparing translation varieties, MT is used as a translation variety independent of HT or other translation varieties in some studies (Toral, 2019). Different from the conventional practice of MT evaluation that treats HT as the gold standard, some studies adopt a descriptive approach to comparing MT and HT (Bizzoni et al., 2020; Ahrenberg, 2017; Vanmassenhove et al., 2019). Among these studies, Bizzoni et al. (2020) find that MT shows independent patterns of translationese and it resembles HT only partly. This implies that MT may be different from HT in a systematic way, and

it remains a question as to whether the deviation of MT from HT is a reliable measure of the quality of MT, and whether the current automatic metrics conflate differences between HT and MT with the quality of MT.

According to research by Toral (2019), translation varieties differ in multiple ways. Based on research by Vanmassenhove et al. (2019), we focus on lexical diversity in our experiments.

We try to answer three questions in this study:

- Can MT and HT be classified automatically with an accuracy above the chance level?
- In what way does lexical diversity influence the classification result?
- Are the results of automatic metrics influenced by the difference in lexical diversity between HT and MT?

2 Related Work

As our study essentially involves comparing translation varieties, we present an overview of previous studies that compare originally written texts and translations, other translation varieties, and HT and MT.

2.1 Comparing Originally Written Texts and Translations

Translated texts show distinctive features which make them different from originally written texts. These features are typically studied under the framework of translationese. Gellerstam (1986) is the first to use this term to refer to the "fingerprints" that the source text leaves on the translated text. This notion is developed by Baker, who proposes the idea of universals of translation. As suggested by Baker et al. (1993), universals of translation are linguistic features that typically occur in translated texts as opposed to originally written texts, and these features are independent of the specific language pairs. Automatic means to distinguish translated texts and originally written texts have been developed and generally achieve high accuracy (Baroni and Bernardini, 2006; Ilisei et al., 2010; Lembersky et al., 2012; Rabinovich and Wintner, 2015). Meanwhile, computational approaches (Teich, 2003; Volansky et al., 2015) contribute evidence for some translation universals.

2.2 Comparing Translation Varieties

Compared with the considerable amount of research on identifying translationese, the differences between translation varieties are less studied.

Rubino et al. (2016) perform the classification between originally written texts and translations as well as between professional and student translations. They use surface features and distortion features which are inspired by quality estimation tasks, and surprisal and complexity features which are derived from information theory. Their experiment shows that originally written texts and professional translations are different mainly in terms of sequences of words, part-of-speech and syntactic tags, and originally written texts are closer to professional translations than to student translations. While the originally written texts and translations can be classified with high accuracy, automatic classification of different translation varieties is a more challenging task. Professional translations and student translations can only be classified with an accuracy barely above 50%.

This finding is consistent with the result of a study by Kunilovskaya and Lapshinova-Koltunski (2019). While morpho-syntactic features can be used to distinguish translations from non-translations with high accuracy, the performance of the same algorithm on classifying professional and student translations only slightly exceeds the chance level.

The differences of translations authored by human translators with different expertise and native languages are studied by Popović (2020). Similar to other studies on distinguishing originally written texts from translated texts or comparing translation varieties, surface text features at word and part-of-speech levels are used. It concludes by suggesting that detailed information about the reference translation including translator information be provided in the scenario of MT evaluation.

Toral (2019) compares post-edited MT with HT in terms of lexical variety, lexical density, sentence length ratio and part-of-speech sequences. The research shows that post-edited MT has lower lexical diversity and lower lexical density than HT, which is linked to the translation universal of simplification, and post-edited MT is more normalized and has greater interference from the source text (in terms of sentence length and part-of-speech sequences) than HT.

2.3 Comparing MT and HT

While the number of studies on comparing translation varieties is much smaller than on the identification of translationese, there are even fewer studies that explore the differences between MT and HT.

Ahrenberg (2017) compares MT and HT by means of automatically extracted features and statistics obtained through manual examination. By comparing the shifts (i.e. deviation from literal translation) and word order changes, he finds that HT contains twice as many word order changes. Meanwhile, an analysis of the number and types of edits required to give the machine translated text publishable quality is made. He argues that MT is likely to retain interference from the source text even after post-editing, and the machine translated text is more similar to the source text than the human translated text in many ways, including sentence length, information flow and structure.

Research by Vanmassenhove et al. (2019) shows another aspect where MT differs from HT. Three MT systems based on different architectures are trained. The lexical diversity of the translations of the MT systems is measured with three metrics including type/token ratio, Yule's K, and measure of textual lexical diversity (MTLD). It is found that the output of neural machine translation (NMT) systems has a loss of lexical diversity compared with the human translated text. The reason for this phenomenon is that the advantage of NMT systems over statistical machine translation (SMT) systems in terms of learning over the entire sequence is obtained at the expense of discarding less frequently occurring words or morphological forms. This finding is consistent with the research by Toral (2019), who observes that the lexical variety of post-edited MT is lower than of HT and the lexical variety of MT is lower than of post-edited MT, which is attributed to the tendency of MT to choose words used more frequently in the training data (Farrell, 2018).

Bizzoni et al. (2020) study the differences between HT and MT in relation to the original texts. Part-of-speech perplexity and a syntactic distance metric are used to measure the differences between translations in written and spoken forms and produced by different types of MT systems. It is found that MT shows structural translationese, but the translationese of MT follows independent patterns that need further understanding.

3 Experiment

We adopt two approaches for classifying MT and HT: developing a trigram language model with Witten-Bell smoothing and fine-tuning a pre-trained BERT model for sequence classification from the Transformers library (Wolf et al., 2020).

3.1 Data

The dataset is from the News commentary parallel corpus v13 (Tiedemann, 2012) provided in the WMT2018 shared task[1]. We use Google Translate[2] to obtain the corresponding machine translation.

The language pairs used in the experiment, the number of sentences for each language pair and the average sentence length for HT and MT are presented in Table 1.

	Number of sentences	**MT avg sentence length**	**HT avg sentence length**
CS-EN	30384	26.33	25.83
DE-EN	30345	26.61	26.15
RU-EN	30387	28.00	27.51

Table 1: Statistics of the dataset: translations from Czech, German and Russian to English.

3.2 Classifying HT and MT

Trigram Model

We train two trigram models on the HT and MT training sets. Let p_{MT} denote the trigram model trained on MT sentences, and p_{HT} the model trained on HT sentences. A sentence s is classified as MT if $p_{MT}(s) > p_{HT}(s)$ and as HT otherwise. If s is from the HT test set and classified as HT, we count it as a success, and the same goes for the case when s is from the MT test set and classified as MT. The classification accuracy is obtained by dividing the number of correct classifications by the total number of sentences in the respective test set. Since the two classes are balanced, accuracy is an appropriate metric. The result is shown in Table 2.

[1] http://www.statmt.org/wmt18/translation-task.html
[2] https://translate.google.co.uk

CS-EN		
Total	MT	HT
0.69	0.79	0.58
DE-EN		
Total	MT	HT
0.66	0.75	0.57
RU-EN		
Total	MT	HT
0.67	0.76	0.58

Table 2: Classification accuracy of the trigram model.

CS-EN		
Total	MT	HT
0.78	0.90	0.66
DE-EN		
Total	MT	HT
0.78	0.87	0.69
RU-EN		
Total	MT	HT
0.78	0.90	0.65

Table 3: Classification accuracy of the BERT model.

From Table 2 it is clear that HT and MT can be classified automatically with an accuracy above the chance level. However, it is noticeable that MT can be classified with higher accuracy than HT.

Based on research by Vanmassenhove et al. (2019) and Toral (2019), this imbalance in classification accuracy may be partly explained by the higher lexical diversity of HT, so that p_{HT} is a probability distribution over sentences composed of a larger set of words than in the case of p_{MT}, thereby typically assigning a lower probability to any particular sentence, regardless of whether it is from MT or from HT.

From Table 1, it can be seen that the difference in average sentence length between MT and HT is only around 0.5. Therefore, we assume that the influence of sentence length is not significant in this study.

BERT Model

We apply the BERT model on the same dataset, which is divided into training, test and validation sets by the ratio of 70%, 10% and 20%. The sentences are padded to the maximum length of sentences in the dataset. We find that the pretrained BERT model for sequence classification achieves higher accuracy and lower loss in the first epoch. The result is shown in Table 3.

From Table 3, it can be seen that fine-tuning the pretrained BERT model for sequence classification can achieve higher accuracy for this task than the trigram model. Moreover, we can see the same pattern of imbalance in classification accuracy between MT and HT. Similar to the case of the trigram model, we hypothesize that it is because greater lexical diversity makes HT more difficult to classify correctly than MT.

3.3 Changing Lexical Diversity

To investigate further whether differences in lexical diversity could be the reason for the observed imbalance in the classification accuracy of MT and HT, we manipulate the lexical diversity of the two. As the lexical diversity of HT is generally higher than of MT (Vanmassenhove et al., 2019; Toral, 2019), we reduce the lexical diversity of HT until it becomes close to or lower than MT, and for comparison, we also reduce the lexical diversity of MT.

Method of Changing Lexical Diversity

Our general strategy of reducing lexical diversity is to replace rare words with words that are close to them in a vector space. First, we find rare words based on the frequency of lemmas in the corpus. Since there are many numerals and proper names and it is difficult to find meaningful candidates to replace them in the vector space, we set token.like_num and token.is_oov in spaCy processing[3] to false. Among the remaining lemmas, those lemmas whose frequency is lower than a threshold will be considered to be rare words. We found that setting the frequency threshold to two is effective in reducing the lexical diversity.

Second, we choose words whose vectors are close to the rare words from the pretrained GloVe embeddings (Pennington et al., 2014), which are computationally less expensive than contextualized word embeddings like BERT. We found that the words which are closest to the rare words are not necessarily the optimal candidates in terms of part-of-speech or meaning, and so we choose the top three most similar words for each rare word. We convert the GloVe vectors into word2vec for-

[3]https://spacy.io

mat with the gensim glove2word2vec API[4] and set restrict_vocab to 30000 in the most_similar function[5] so that the search for the most similar words is limited to the top 30000 words in the pretrained embeddings. The vocabulary size 30000 was determined empirically.

After this step, we apply a check on the fine-grained tags of the rare words and the fine-grained tags of the respective three candidates, the tags being obtained with spaCy [6] and containing more information than the coarse-grained part-of-speech tags from the Universal POS tag set[7]. The candidates with the same tags as the rare words will be chosen. Where there is more than one matched candidate, only the first is chosen, and when there are no matched candidates after the check, the rare words will not be replaced. In this way, we obtain texts with modified lexical diversity. For ease of reference, modified HT texts will be referred to as HT_modf, modified MT texts will be referred to as MT_modf, original HT texts as HT_orig and original MT texts as MT_orig.

To compute the lexical diversity of the texts, based on research by McCarthy and Jarvis (2010) and Vanmassenhove et al. (2019), we choose the measure of textual lexical diversity (MTLD) (McCarthy, 2005), which is reasonably robust to text length difference. We refer those interested in the specific computation and statistical significance of MTLD to McCarthy and Jarvis (2010). The lexical diversity of the texts is presented in Table 4.

MTLD	**Original**	**Modified**
CS_MT	62.02	43.00
CS_HT	63.80	43.04
DE_MT	62.53	42.44
DE_HT	64.59	42.76
RU_MT	61.06	42.66
RU_HT	64.51	43.05

Table 4: MTLD of the original texts and of the modified texts.

From Table 4, it can be seen that the MTLD values of HT texts are generally higher than of MT texts, which is consistent with the result of previous studies (Vanmassenhove et al., 2019, 2021; Toral, 2019). With our method, the difference in MTLD value between MT and HT texts is reduced.

Experimental Result of Trigram Model

We conduct another set of binary classification experiments on the original and modified MT and HT texts paired in different ways. For example, "MT_modf & HT_modf" in the following tables means that the binary classification is performed on the modified MT text and the modified HT text. The result of the trigram model is shown in Table 5. For comparison, the results from Table 2 are repeated in the lines MT_orig & HT_orig.

CS-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.69	0.79	0.58
MT_modf & HT_modf	0.69	0.77	0.61
MT_orig & HT_modf	0.69	0.56	0.83
DE-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.66	0.75	0.57
MT_modf & HT_modf	0.67	0.74	0.60
MT_orig & HT_modf	0.67	0.52	0.82
RU-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.67	0.76	0.58
MT_modf & HT_modf	0.67	0.75	0.59
MT_orig & HT_modf	0.67	0.52	0.82

Table 5: Binary classification of MT and HT by the trigram model under different combinations of MT and HT texts.

From Table 5 in combination with Table 4, we can see that when the difference in lexical diversity between MT and HT becomes smaller, the imbalance in classification accuracy is reduced, and the classification accuracy of MT goes down while the classification accuracy of HT goes up.

Since the lexical diversity of HT is generally higher than MT, we conduct an experiment where the lexical diversity of HT is significantly lower than MT, and the result is shown in the lines MT_orig & HT_modf. Under this condition, the classification accuracy of MT is much lower than HT. In this way, we reverse the previously observed trend that the classification accuracy of MT is higher than HT. Note that the overall classification accuracy does not change much in this experiment.

[4]https://radimrehurek.com/gensim/scripts/glove2word2vec.html

[5]https://radimrehurek.com/gensim/models/word2vec.html

[6]https://spacy.io/api/token#attributes

[7]https://universaldependencies.org/docs/u/pos/

Experimental Result of BERT Model

For fine-tuning the pretrained BERT model for sequence classification, similar experiments were done, with different combinations of MT and HT texts. Accuracies are presented in Table 6.

CS-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.90	0.66
MT_modf & HT_modf	0.78	0.89	0.68
MT_orig & HT_modf	0.82	0.91	0.73
DE-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.87	0.69
MT_modf & HT_modf	0.78	0.86	0.71
MT_orig & HT_modf	0.81	0.89	0.73
RU-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.90	0.65
MT_modf & HT_modf	0.77	0.89	0.65
MT_orig & HT_modf	0.81	0.95	0.68

Table 6: Binary classification of MT and HT by the BERT model under different combinations of MT and HT texts.

Similar to the trigram model, the classification accuracy of HT goes up in the case of CS-EN and DE-EN and the classification accuracy of MT goes down a little, when the lexical diversity of MT and of HT are closer, as shown in the lines MT_modf & HT_modf, and when the lexical diversity of HT is much lower than MT, the classification accuracy of HT goes up, as shown in the lines MT_orig & HT_modf. However, changing the difference in lexical diversity does not tend to decrease the classification accuracy of MT for the BERT model. Recall that with the trigram model, the classification accuracy of HT increases while the classification accuracy of MT decreases. In contrast, with the BERT model, even when the lexical diversity of MT is much higher than HT, the overall classification accuracy and the separate classification accuracies of MT and HT all go up. The difference of the two models in terms of the classification accuracy of MT may be explained by the fact that the pretrained BERT model for sequence classification calculates cross-entropy loss for the classification task[8] while the trigram model results from relative frequency estimation.

3.4 Automatic Metrics

We hypothesize that the performance of the two models in the binary classification task may be reflected in the result of MT metrics that are based on n-gram matching or that use contexualized embeddings.

Since BLEU is a commonly used metric based on n-gram matching, we test the performance of BLEU on the dataset to see if the difference in lexical diversity between MT and HT would influence the result. We calculate the corpus-level BLEU score for MT, as implemented in NLTK[9], using HT as reference. The result is presented in Table 7.

BLEU	MT_orig & HT_orig	MT_modf & HT_modf	MT_orig & HT_modf
CS-EN	0.42	0.46	0.39
DE-EN	0.41	0.45	0.38
RU-EN	0.37	0.40	0.34

Table 7: BLEU score.

As can be seen from Table 7, when the lexical diversity of MT is closest to HT, as shown by the column MT_modf & HT_modf, the MT BLEU score is the highest. When the lexical diversity of the reference is much lower than MT, as is the case in the column MT_orig & HT_modf, the MT BLEU score is the lowest. Much as in the discussion of the results of the trigram model, the difference in lexical diversity between MT and HT is a factor that needs to be taken into account when an n-gram matching based metric like BLEU is used for MT evaluation.

The majority of automatic MT metrics developed in recent years such as BERTScore (Zhang et al., 2019) and Yisi (Lo, 2019) adopt contextualized embeddings. Based on accessibility and performance, we choose MoverScore (Zhao et al., 2019) as an example of a metric that uses BERT representations. Since MoverScore is not a corpus-level metric, we calculate the average

[8] https://github.com/huggingface/transformers/blob/9aeacb58bab321bc21c24bbdf7a24efdccb1d426/src/transformers/modeling_bert.py

[9] https://www.nltk.org/

sentence-level score. The result is presented in Table 8.

Mover-Score	MT_orig & HT_orig	MT_modf & HT_modf	MT_orig & HT_modf
CS-EN	0.57	0.56	0.55
DE-EN	0.57	0.56	0.55
RU-EN	0.52	0.50	0.50

Table 8: MoverScore result for MT.

The MoverScore result in Table 8 shows a different pattern from the BLEU scores. The scores are basically inversely proportional to the overall accuracy of the binary classification task shown in Table 6. As the difference in MoverScore results under different combinations of MT and HT texts is small, more work is needed.

4 Conclusion and Future Work

With the above experiments, we have shown that MT and HT can be classified with an accuracy above the chance level. The trigram model does not involve a machine learning algorithm but is capable of capturing the differences between MT and HT. By fine-tuning the pretrained BERT model for sequence classification, we obtain a higher accuracy for this task.

Similar to the identification of translationese, we may claim that MT and HT belong to different translation varieties. The result serves as supporting evidence for the study by Bizzoni et al. (2020), which maintains that MT only resembles HT in part and often follows independent patterns. This finding calls into question the longstanding assumption in MT evaluation that the more similar an MT output is to a professional human translation, the better it is. If MT and HT are two translation varieties and have different patterns, it leaves room for doubt as to the legitimacy of evaluating MT by its similarity to HT.

Moreover, there is a noticeable imbalance in the classification accuracy of HT and MT. For the trigram model, while more than 70% of the MT test sentences can be classified correctly, fewer than 60% of the HT test sentences are classified correctly. This imbalance also exists in the experiment with the BERT model. Generally speaking, it is easier to correctly classify MT sentences than HT sentences.

Based on previous studies and analysis from the probabilistic perspective, we consider lexical diversity as one of the major reasons for this imbalance in classification accuracy. We change the lexical diversity of the MT and HT texts and conduct another set of experiments with the same models. With the trigram model, if the difference in lexical diversity between MT and HT decreases, the imbalance in classification accuracy between the two is reduced, and we can reverse this imbalance in classification accuracy when the lexical diversity of MT is higher than HT. The result of the experiment with the BERT model shows a different pattern. An increase in classification accuracy of HT is accompanied by an increase in the classification accuracy of MT. This may be explained by the different ways of performing binary classification by the two models.

The performance of automatic MT metrics based on n-gram matching, represented by BLEU in this study, and automatic metrics using BERT representations, such as MoverScore, is related to the result of the binary classification task with the two kinds of models. When the lexical diversity of HT is lower than MT, the MT BLEU score is the lowest and when the lexical diversity of HT is very close to MT, the MT BLEU score is the highest. The evaluation results given by MoverScore are basically inversely proportional to the classification accuracy of the BERT model. Therefore, we suggest the difference in lexical diversity between MT and the reference be given more attention in MT evaluation with automatic metrics.

We are aware that there are other possible factors that may account for the phenomenon that HT is more likely to be classified as MT than the other way around. In our experiment, we only manipulate one factor. In future work, we intend to further study the independent patterns of MT compared with HT and investigate if the differences between MT and HT are related to the quality of MT. As differences in lexical diversity may influence automatic metrics for MT evaluation in different ways, we plan to explore this phenomenon with other metrics, such as COMET (Rei et al., 2020).

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *RANLP 2017: The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*,

pages 21–28. Association for Computational Linguistics.

Mona Baker et al. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and Technology: In honour of John Sinclair*, 233:250.

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.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation Studies in Scandinavia*, 1:88–95.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

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.

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.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Maja Popović. 2020. On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 365–374.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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.

Jonathan Slocum. 1985. A survey of machine translation: Its history, current status and future prospects. *Computational linguistics*, 11(1):1–17.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218.

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

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at wmt 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

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.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Quantitative Evaluation of Alternative Translations in a Corpus of Highly Dissimilar Finnish Paraphrases

Li-Hsin Chang, Sampo Pyysalo, Jenna Kanerva, and Filip Ginter
TurkuNLP Group
Department of Computing
Faculty of Technology
University of Turku, Finland
{lhchan, sampyy, jmnybl, figint}@utu.fi

Abstract

In this paper, we present a quantitative evaluation of differences between alternative translations in a large recently released Finnish paraphrase corpus focusing in particular on non-trivial variation in translation. We combine a series of automatic steps detecting systematic variation with manual analysis to reveal regularities and identify categories of translation differences. We find the paraphrase corpus to contain highly non-trivial translation variants difficult to recognize through automatic approaches.

1 Introduction

The study of translation language for Finnish has largely focused on individual linguistic features. The debate on the existence of translation universals sparked the well-developed research line of comparing translated and original language. Examples of such studies include the comparison of nonfinite structures in translated and original Finnish (Puurtinen, 2003; Eskola, 2004), and investigation of subject changes in translations using a French-Finnish parallel corpus (Huotari, 2021). Variation in alternative translations is less studied. Paloposki and Koskinen (2004) qualitatively compare the degree of domestication in language use in Finnish first translations and retranslations. While this study is done qualitatively, several paraphrase corpora with translated language have been released more recently, enabling research from a quantitative prospective. Such corpora include Opusparcus (Creutz, 2018) and TaPaCo (Scherrer, 2020), both constructed automatically using language pivoting and containing Finnish subsets.

Recently, the Turku Paraphrase Corpus has become available (Kanerva et al., 2021), consisting of paraphrase pairs, of which the vast majority are manually selected from the OpenSubtitles[1] dataset. The construction of the paraphrase corpus capitalizes on the fact that many movies and TV shows have multiple independently produced translations. The selection is carried out manually, comparing side-by-side the two lexically maximally distant subtitle versions for each movie or TV show and selecting instances of paraphrases. Upon selection, the candidate pairs are assigned to a category such as *paraphrase in any context* or *paraphrase in this context but not universally*, etc. The Turku paraphrase corpus is substantial in size, with 45,000 manually extracted, naturally occurring paraphrase pairs (a paraphrase pair henceforth refers to two segments of text, each about a sentence long or slightly longer), and a further 7,900 pairs created by editing an extracted pair so as to obtain a fully context-independent paraphrase.

Due to the way in which it was constructed, the corpus is directly applicable to the study of translation language and in particular to the analysis of variation in translation. The unique value of the corpus for this purpose is that it consists mostly of fully manually selected translation variants focused on lexically and structurally dissimilar pairs. These are very difficult to extract automatically: automatic methods can reliably identify only simple variation, while lexically and structurally substantially different pairs are very difficult to automatically distinguish from non-paraphrases, i.e. phrases that are not alternative translations.

In this paper, we will characterize the paraphrase corpus in terms of translation language, focusing especially on the types of variation (e.g. synonym usage, redundancy or verbosity) occurring in the data. Our aim is to establish whether the corpus can be of utility to translation language modelling and machine translation system evaluation. To this end, we will focus on two main ques-

[1] http://www.opensubtitles.org

tions: (a) how easily could the translation pairs be extracted automatically, and (b) what are the main types of variation exhibited by the pairs.

2 Corpus statistics and pre-processing

The full corpus includes 45,000 naturally occurring paraphrases and 7,900 pairs obtained by rewriting a previously extracted example. The source of these paraphrases is in the vast majority of cases alternative translations of subtitles, with a small section originating from news headings. To construct a lexically and structurally diverse paraphrase corpus, the annotators were instructed to only select non-trivial paraphrase candidates, avoiding simple, uninteresting changes such as minor differences in inflection and word order.[2] For the analysis in this paper, we use the training section of the corpus, restricting further exclusively to examples originating from OpenSubtitles. This gives 34,561 naturally occurring paraphrase pairs and 5,445 rewritten paraphrases. Each naturally occurring paraphrase pair in the corpus have a numerical label manually assigned by an annotator from the following set: `4`: universally paraphrase regardless of context, `3`: paraphrase in the given context but not universally, `2`: related but not paraphrase. Additionally, those annotated as `4` can be assigned one or several flags which sub-categorize different types of paraphrases: > or <: universal paraphrase in one direction but not the other, `s`: substantial difference in style, `i`: meaning-affecting difference restricted to a small number of morphosyntactic features. By contrast to the original paraphrases, the rewrites are always full, universally valid paraphrases, i.e. label `4`. The rewriting process strives to change as little of the original sentences as possible: these include simple fixes such as word or phrase deletion, addition or re-placement with a synonym or changing an inflection, while more complicated changes are avoided. The rewrites are thus an efficient way to obtain full paraphrases in terms of corpus creation. The label distribution of the Turku paraphrase corpus subset used for later analysis is shown in Table 1.

For the purpose of the subsequent analysis, we parse the paraphrases using the Turku Neural Parser Pipeline (Kanerva et al., 2018, 2020), a state-of-the-art parser producing POS and morphological tags, word lemmas, as well as dependency trees in the Universal Dependencies scheme (Nivre et al., 2016). We use the model trained on UD_Finnish-TDT v2.7 corpus, which utilizes the pre-trained FinBERT language model in tagging and dependency parsing (Virtanen et al., 2019).[3]

Universal paraphrases	14,986
Label 4	8,578
Label 4s	963
Rewrites	5,445
Context-dependent paraphrases (Label 3 or has <, >, or `i` flags)	24,757
Related but not paraphrase	263
Total	**40,006**

Table 1: Label distribution of paraphrases from the subset of alternative subtitle translations in Turku paraphrase corpus training set.

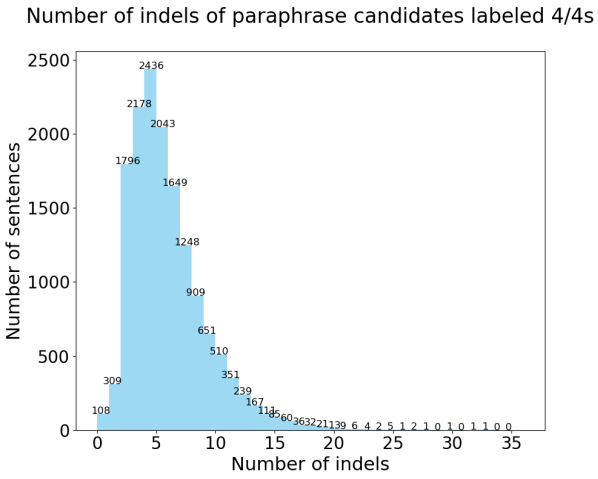

Figure 1: Distribution of the number of lemma indels for universal paraphrases labeled `4`/`4s` including rewrites.

3 Analysis of variation

3.1 Automatic categorization

To investigate and categorize the paraphrase pairs by the form of variation, we calculate the difference in the set of lemmas (i.e. insertions/deletions of lemma, henceforth lemma indels) for each pair, excluding punctuation characters from the analysis. Figure 1 shows the distribution of the number of lemma indels for all universal paraphrases showed in Table 1 (paraphrases with labels `4` and `4s` including rewrites), i.e. all pairs

[2] Finnish has relatively free word order and reordering can be trivially detected automatically.

[3] Model available at `https://turkunlp.org/Turku-neural-parser-pipeline/models.html`

Ratio	Word	Indel	Total
0.45	tosi (really)	64	143
0.41	lakata (stop)	51	125
0.39	ikävä (unfortunate)	55	142
0.38	tahtoa (want)	83	216
0.37	ihan (quite)	145	391
0.35	todella (really)	201	572
0.34	kai (perhaps)	107	311
0.34	aivan (exactly)	117	343
0.34	kyllä (truly)	158	465
0.34	ikinä (never)	127	374

Table 2: Most overrepresented words varying between different translations (minimum occurrence in corpus=50)

equivalent in meaning regardless of their context. As a result of excluding trivial paraphrase candidates, less than 1% (108 pairs) out of 14,986 pairs have zero lemma indels. Such pairs are formed purely by word reordering and/or changes in inflection. We next investigate paraphrase pairs that can be accounted for by automatic synonym substitutions. We combine two resources to build a synonym dictionary for lemmas. The first resource is `Word2Vec` embeddings (Mikolov et al., 2013) for lemmas trained from Suomi24 discussion fora texts[4]. For each lemma, we take at most 15 closest lemmas in the vector space as synonyms using the `gensim` library (Řehůřek and Sojka, 2010). In addition, we supplement our synonym dictionary with Finnish WordNet (Lindén and Niemi, 2014) using the NLTK library (Bird et al., 2009). Out of the 14,878 pairs of paraphrases with lemma indels, 951 pairs (∼6%) have all of their lemma indels accounted by synonyms. An additional 7370 pairs (∼49%) have lemma indels partially accounted by synonyms. The synonym dictionary only takes into account one-to-one synonyms. As a consequence, one-to-many synonyms and phrasal paraphrases are not included.

Table 2 shows the lemmas that are most overrepresented among the inserted or deleted words relative to their overall frequency. We find emphasizers (e.g. *tosi (really)*), particles (e.g. *kyllä (truly)*), auxiliary verbs, other functional words, and a small number of very common synonym pairs among the most frequently varying words.

To further focus on meaningful variation, we disregard all words with a dependency relation deemed functional in the Content-Word Labeled Attachment Score (CLAS) (Nivre and Fang, 2017), which is developed to evaluate dependency parsing with focus on content-bearing words.[5] After disregarding these functional words, we are able to account for the variation in a further 82 paraphrase pairs. All of the above mentioned findings are summarized in Table 3. As the variation in 13,608 pairs (i.e. full 90% of the data) is not accountable by using the above automatic categories, we characterize these manually.

4/4s	14986
Word reordering	1
Same lemma, same order	27
Same lemma, different order	80
CLAS	82
Synonym	945
Synonym + CLAS	243
Others	13608

Table 3: Automatic classification of universal paraphrases labeled `4/4s` including rewrites.

3.2 Manual categorization

In the manual categorization, we sample 100 paraphrase pairs among those paraphrases where the variation is not fully explainable using the automatic metrics defined above. Each paraphrase pair is annotated in terms of 8 different variation categories: *word-to-word*, *word-to-phrase* and *phrase-to-phrase* synonyms indicating a straightforward single word synonym replacement, a single word replaced with a synonymous phrase, or a phrase replaced with a synonymous phrase, *redundancy or verbosity* for including additional words not strictly essential for the meaning, *explicit pronouns* for explicitly including pronouns visible otherwise in the verb inflection, *emphasizer* for including additional emphasis words (such as very), *figurative language/idioms*, and *uncertainty or hedging* where both statements express hedging with different markers.

For each paraphrase pair a set of categories explaining the variation is annotated. In Table 4 we

[4] `dl.turkunlp.org/finnish-embeddings/finnish_s24_skgram_lemmas.bin`

[5] These dependency relations are `aux` (auxiliary), `aux:pass` (passive auxiliary), `case` (pre/postposition), `cc` (coordinating conjunction), `clf` (classifier), `cop` (copula), `det` (determiner), `mark` (marker), `punct` (punctuation), `cc:preconj` (preconjunct), and `cop:own` (copula in possessive clauses).

Category	Count	Ratio
Word-to-word synonym	61	34%
Word-to-phrase synonym	33	18%
Phrase-to-phrase synonym	22	12%
Redundancy or verbosity	21	12%
Explicit pronouns	16	9%
Emphasizers	14	8%
Figurative language/idioms	9	5%
Uncertainty or hedging	3	2%

Table 4: Manual analysis results

plot the frequency of each category, showing the straightforward single word synonym replacement being by far the most frequent category, occurring in 61% of the paraphrase pairs. However, albeit word-to-word replacement being frequent, it rarely accounts for the whole variation in the pair. Only 12% of the paraphrases include word-to-word synonyms as sole variation category, other instances occurring in combination with at least one additional variation category.

3.3 Amount of Non-elementary Variation

We measure the proportion of non-elementary variation in the alternative translations in terms of percentage of text (in terms of alphanumeric characters) in the manually extracted paraphrase pairs, out of the total amount of the source material that the annotators processed. The proportion is 15.8%, meaning that approximately every sixth line was considered to be dissimilar in an interesting manner by the annotators, enough to be included in the paraphrase corpus. The remaining 84% of the text is reported by the corpus creators to be for the most part elementary variation, text without correspondence in the other subtitle version, conflicting erroneous translations, and rarely pairs that are meaningless without deep understanding of their broader context.

3.4 Language pivoting

To establish the proportion of the manually extracted paraphrase pairs that could be identified through their source text, as well as to establish the feasibility of automatically aligning the paraphrase pairs with their English source, we use the OpenSubtitles section of the OPUS machine translation dataset and identify those pairs in our dataset that have at least one common English source segment in the English–Finnish OpenSubtitles section of OPUS. We normalize both Finnish and English texts by lowercasing and dropping all non-alphanumeric characters so as to maximize the recall.

Such language pivoting is a common technique for mining cases of translation variation. Language pivoting targets candidates, where the same source-language segment is translated into two different target-language segments, using a corpus of aligned bilingual document pairs. The candidates are typically further filtered by various means to remove spurious alignments and other pairs which are not equivalent in meaning, despite sharing the same aligned source-language segment.

We find that 2,136 pairs were matched, a mere 6% of all categories of paraphrase in the corpus (barring rewrites). Full 94% of the paraphrase pairs cannot be reached through simple language pivoting at least on the level of full segments. Further, while the average length of texts found through pivoting is 3.8 tokens, the average length of texts in the data is 8.4 tokens. The pivoting thus unsurprisingly biases towards short segments, that are more likely to be appropriately aligned and identified. Clearly, in order to align the paraphrase pairs with their (mostly English) source, a manual annotation step will be necessary.

4 Discussion, Conclusions and Future Work

In this paper, we have presented a quantitative analysis of a large, manually extracted paraphrase dataset from the point of view of translation language, and especially its non-elementary variation. Our findings are two-fold. Firstly, we demonstrated that in the case of OpenSubtitles — a very widely used corpus in machine translation — the proportion of non-elementary variation in alternate translations is relatively small, at 16% of the text. Secondly, we have shown that the paraphrase corpus contains highly non-trivial translation variants that are difficult to account for through simple heuristics and can thus serve for further study in translation language without biasing the results towards simpler examples that can be identified automatically.

The corpus in its current form can serve as a resource for evaluating robustness of different evaluation metrics. Quora Question Pairs (QQP)[6] and

[6] data.quora.com/First-Quora-Dataset-\
Release-Question-Pairs

the QQP subset of Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) have been used to evaluate the robustness of machine translation and image captioning metrics (Zhang et al., 2020). QQP is a collection of question headings from the Quora forum labeled as either duplicate or not, while PAWS is an adversarial dataset automatically generated from QQP and Wikipedia to contain highly lexically similar paraphrases and non-paraphrases. Based on our findings, the Turku paraphrase corpus serves as an interesting resource to be used in a similar manner to evaluate metric robustness. An obvious direction for future work is to align, through a combination of heuristics and manual annotation, the paraphrase pairs with their English source. This would result in a test set suitable for evaluation of machine translation systems in terms of their rephrasing ability, as well as for research on MT system evaluation methodology in presence of substantial rephrasing.

Acknowledgments

The research presented in this paper was partially supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG). The research was also supported by the Academy of Finland and the DigiCampus project. Computational resources were provided by *CSC — the Finnish IT Center for Science*. We thank Veronika Laippala for her advice from a linguistic point of view.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Mathias Creutz. 2018. Open Subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sari Eskola. 2004. Untypical frequencies in translated language: A corpus-based study on a literary corpus of translated and non-translated Finnish. In Anna Mauranen and Pekka Kujamäki, editors, *Translation Universals: Do they exist?*, pages 83 – 99. John Benjamins Publishing Company.

Léa Huotari. 2021. *Effet du prototype sur le changement de sujet en traduction : Étude d'un corpus bidirectionnel littéraire français↔finnois*. Ph.D. thesis, University of Helsinki.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23nd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2020. Universal Lemmatizer: A sequence to sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation. *Language Resources and Evaluation*, 48(2):191–201.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden.

Outi Paloposki and Kaisa Koskinen. 2004. A thousand and one translations: Revisiting retranslation. In Gyde Hansen, Kirsten Malmkjaer, and Daniel Gile, editors, *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*, pages 27 – 38. John Benjamins Publishing Company.

Tiina Puurtinen. 2003. Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals? In Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson, editors, *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, pages 141 – 154. Brill.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 6868–6873.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

A Example instances of manual analysis categories

	Translation$_1$	**Translation**$_2$
Word - word	Vasta ammuttu	Ammuttu hiljattain
	Olen pistämättömän hygieeninen.	Olen moitteettoman hygieeninen.
	Etkö mennyt poliisin luo?	Et mennyt poliisin puheille?
	[...] on luultavasti uusi identiteetti.	[...] on varmasti uusi henkilöllisyys.
Word - phrase	Anteeksi odotus.	Anteeksi, että kesti.
	En edes osaa näytellä.	En edes tiedä miten näytellä.
	On niin paljon valinnanvaraa.	On niin paljon mistä valita.
	Useimmat teistä tietävät [...]	Suurin osa teistä tietää, [...]
Phrase - phrase	Andrew ehti ensin.	Andrew oli vain nopeampi.
	Iän myötä [...]	Mitä vanhemmaksi tulin, sitä [...]
	Miksi hän tekee niin? Etkö ole utelias?	Etkö halua tietää miksi hän tekee niin?
	kuuluuko seuralaisennekin tilin osakkaisiin?	Kuuluuko tili myös seuralaisellenne?
Figurative	Olen täysin hereillä, [...]	Olen pirteä kuin peipponen [...]
	Ole nyt vain hiljaa.	Pidä nyt vain pääsi kiinni.
	Teitkö sen tasataksesi tilit?	Teitkö sen päästäksesi tasoihin?
	Tiedä häntä.	En minä tiedä.
Emph.	Jopa runoja.	Runojakin.
	En tiennyt koko säännöstä.	En edes tiennyt säännöstä.
	Mitä täällä tapahtui?	Mitä ihmettä täällä on tapahtunut?
	[...] näen asiat selvemmin.	[...] näen kaiken aina selvemmin.
Verbosity/ redund.	Voin kertoa teille, että [...]	Se mitä voin kertoa teille, on että [...]
	Se, ketä etsit, on kuollut!	Se ihminen jota etsit on kuollut!
	Mihin voin laittaa tämän? Pedille.	Minne voin laskea tämän? Voit laittaa sen sängylle.
	Hae ensiapupakkaus vessan kaapista.	Hae ensiapupakkaus. Se on vessan kaapissa.
Hedge	[...] herättävätkö ne liikaa huomiota.	[...] että ne saattavat kiinnittää liikaa huomiota.
	Vihaan [...] luultavasti ehkä enemmän [...]	Vihaan [...] ehkä enemmänkin [...]
	Lapset taisivat [...]	Näyttää siltä, että lapset [...]

Table 5: Examples of manual analysis categories. English translations in Table 6.

	Translation$_1$	**Translation**$_2$
Word - word	Recently shot	Just shot
	I am spotless clean	I am perfectly clean
	Didn't you approach the police?	Didn't you talk to the police?
	[...] is likely a new identity.	[...] is surely a new ID.
Word - phrase	Sorry the wait.	Sorry, that it took long.
	I can't even perform.	I don't even know how to perform.
	The choice is so varied.	The choice is very broad.
	Most of you know [...]	The biggest part of you know, [...]
Phrase - phrase	Andrew made it there first.	Andrew was simply faster.
	With age [...]	The older I became, [...]
	Why is he doing so? Aren't you curious?	Don't you want to know why he is doing so?
	Does your colleague also belong among the stock holders?	Does the stock belong also to your colleague?
Figurative	I am fully awake, [...]	I'm astir as a bird [...]
	Be quiet now.	Keep your mouth shut.
	Did you do it to even the score?	Did you do it to get equal?
	God knows.	I don't know..
Emph.	Quite the poem.	A poem.
	I didn't know of the rule as such.	I really didn't know of the rule.
	What happened here?	What on earth happened here?
	[...] you see things more clearly.	[...] you always see everything more clearly.
Verbosity/ redund.	I can tell you that [...]	What I can tell you is that [...]
	The one you are looking for is dead!	The person you are looking for is dead!
	Where can I put this? On the bed.	Where can I lay this down? You can put it on the bed.
	Fetch the first aid kit from the cupboard in the washroom	Fetch the first aid kit. It is in a cupboard in the washroom.
Hedge	[...] do they attract too much attention.	[...] that they may attract too much attention.
	I hate [...] presumably maybe more [...]	I hate [...] maybe even more [...]
	The kids might [...]	It seems that the kids [...]

Table 6: Examples of manual analysis categories, best-effort translation to English.