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.).

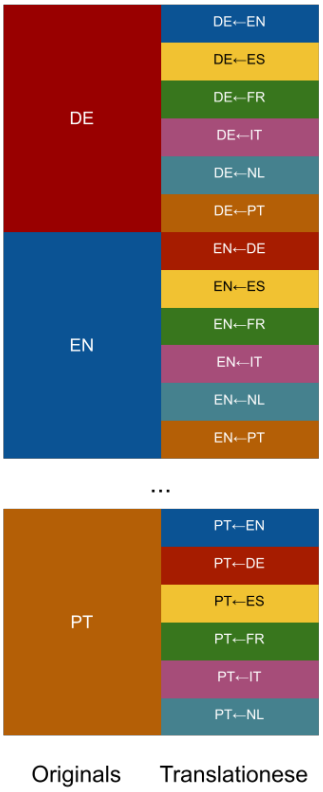

Figure 2: Structure of a 7-language comparable corpus for translationese classification.

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).