EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates

Alina Karakanta, Mihaela Vela, Elke Teich
Department of Language Science and Technology, Saarland University
alina.karakanta@uni-saarland.de, {m.vela, e.teich}@mx.uni-saarland.de

Abstract

Multilingual parliaments have been a useful source for monolingual and multilingual corpus collection. However, extra-textual information about speakers is often absent, and as a result, these resources cannot be fully used in translation studies. In this paper we present a method for processing and building a parallel corpus consisting of parliamentary debates of the European Parliament for English into German and English into Spanish, where original language and native speaker information is available as metadata. The paper documents all necessary (pre- and post-) processing steps for creating such a valuable resource. In addition to the parallel corpora, we collect monolingual comparable corpora for English, German and Spanish using the same method.

Keywords: parallel corpus, comparable corpus, European Parliament, metadata, multilingual

1. Introduction

Multilingual parliaments have been a useful source for monolingual and multilingual corpus collection. However, it is often the case that the compilation of these corpora is not transparent or that useful information about speakers and the status of a given speech (original vs. translation) is absent. Consequently, parliamentary corpora cannot be directly used for research on translation.

An important and probably the earliest attempt to create a parallel corpus from parliamentary proceedings is the Canadian Hansard corpus. It consists of transcripts of debates of the Canadian Parliament (annotated with metadata about the original language) in the two official languages of Canada, English and French. Similarly, several attempts have been made to collect and structure the proceedings of the European Parliament. One of the most popular collections of European parliamentary proceedings is EuroParl (Koehn, 2005), which has been widely used for machine translation[1] and cross-lingual research (Cartoni et al., 2013). It consists of transcribed and revised spoken utterances by speakers of the European Parliament (EP), translated into several languages. Although the monolingual subcorpora of EuroParl often include metadata about the original language of the sentences, this information is not always consistent and it is completely absent from the bilingual corpora. For this reason, EuroParl might be suitable for training MT systems, but for other tasks manipulation of the data is often required. For this reason, other projects have focused on correcting and structuring EP proceedings for linguistic applications (cf. Corrected and Structured EuroParl corpus (Graën et al., 2014), European Comparable and Parallel Corpora (Calzada Pérez et al., 2006), Digital Corpus of the European Parliament (Hajlaoui et al., 2014), Talk of Europe – Travelling CLARIN Campus/LinkedEP (van Aggelen et al., 2017)).

For translationese research, parliamentary proceedings have to be structured as parallel corpora where the translation direction is known. Most of the previous projects on this field rely on the "language" tag to extract sentences produced in the original language from EuroParl (Lembersky et al., 2012b), even though this information is scarce and sometimes inconsistent, as shown by Cartoni and Meyer (2012). Rabinovich et al. (2015) compile a cross-domain corpus for translationese research annotated with metadata about the translation direction. In later work, Rabinovich et al. (2017) attempt to preserve the traits of the original author in the extracted corpus in order to measure the signals left by the author's gender in original and translated text. Nisioi et al. (2016) create a monolingual English corpus of native, non-native and (human) translated texts extracted from the EP proceedings. The corpus is a subset of the corpus collected by Rabinovich et al. (2015) and preserves, similar to our corpus, metadata about the speaker.

Contrary to these approaches, we provide a complete pipeline to collect and compile European Parliament debates into a high-quality, metadata-rich corpus, with accurate speaker and language information, useful for a variety of natural language processing (NLP) tasks.

The paper is structured as follows. Section 2. presents the motivation for building such a resource. Section 3. describes the processing steps, including crawling the web, sorting and filtering the crawled data. In this section we also give an overview on the metadata as well as the corpus structure and statistics. Section 4. discusses possible applications of such a corpus in the field of translation studies and Section 5. provides a brief summary and conclusion.

2. Motivation

Motivations for building a resource as the one described lie in the intended context of use. Machine translation can profit from such a resource, since it has been shown that for statistical machine translation (SMT) direction-aware translation models yield better translation quality than models trained on texts in the opposite direction (Kurokawa et al., 2009; Lembersky et al., 2012a).

Translation studies, in particular research on the specific properties of translations, is a research field that can profit from such a resource. Research on (human and machine) translations has shown that translations exhibit specific properties, such as simplification, explicitation, normaliza-

[1] http://statmt.org/moses/

tion, shining-through etc., also known as "translationese" (cf. Baker (1995; Laviosa (1998; Teich (2003; Volansky et al. (2015)). The only factor taken into consideration in this kind of studies is, by now, translation direction. As shown by Koppel and Ordan (2011) translationese research should incorporate other relevant factors, too, including information on the speaker (native vs. non-native) or production mode (written vs. spoken).

Other NLP research fields such as gender identification (Koppel et al., 2002) or topic detection (Yang et al., 2011; Blei, 2012) might also benefit from metadata-rich corpora. For example, information about the affiliation of a speaker to a specific party or parliamentary group interconnected with information about the country they represent, allows for detecting common (or different) topics at party, group, national or European level.

3. Corpus Processing

In this section, we describe a pipeline for building a comparable/parallel corpus from European Parliament debates. It is based on meta-information on the proceedings and the Members of the European Parliament (MEP). Our final goal is to obtain:

(i) a **parallel corpus** where the source language (SL) sentences come from native SL speakers and are aligned to sentences in the required target language (TL) and

(ii) a **comparable monolingual corpus** of the target language, where the sentences come from native TL speakers.

The process of building the corpus can be described in the following steps:

1. Download proceedings in HTML
2. Download MEPs' metadata in HTML
3. Extract MEPs' information in a CSV file
4. Model proceedings as XML
5. Filter out text units not in the expected language
6. Add MEPs' metadata to proceedings
7. Add sentence boundaries
8. Annotate token, lemma, Part-of-Speech
9. Separate originals from translations and filter by native speakers
10. Extract text into raw format
11. Sentence-align the resulting corpus

Even though this is an end-to-end pipeline, some steps are independent from each other. For example, step 11 applies only to create a raw sentence-aligned parallel corpus, suitable for MT experiments, while step 8 is optional and can be applied at any point.

3.1. Crawling the Data

The data to compile the corpus was collected from the official website of the European Parliament[2]. A typical URL for the proceedings of a given day consists of the base URL, a date and the language version. To date our method provides support only for English, German and Spanish, but it can be easily localized by simply translating the roles (e.g. president, commissioner) in the required language. It is also possible to determine a specific date range.

```
language_version = en #choose language
if list_of_dates is True:
    read(list_dates)
else:
    generate_range_of_dates(list_dates)
for date in list_dates:
    generate(URL)
    request(URL)
    if URL is True:
        download(document)
    else:
        proceed_next_date(date)
```

Figure 1: Pseudocode for crawling the proceedings

Following the process shown in Figure 1, we collected URLs with dates between 20/07/1999 and 18/01/2018. The format of the obtained data is HTML, which allows us later to preserve meta-information as XML. In addition to the proceedings, the European Parliament website maintains a database with all MEPs[3]. We obtained MEPs' information, such as basic information about the speaker and their history record, also in HTML.

3.2. Metadata

There are two types of metadata collected for the purpose of this corpus:

(a) **Proceedings' metadata**
Proceedings' metadata is basic metadata about the parliamentary session. As depicted in Figure 2, a session is divided into several sections, i.e. agenda items, which are then subdivided into interventions. Information is also obtained about the speakers and the source language of the text. Lastly, the metadata contains the actual text of the proceedings as paragraphs.

(b) **MEPs' metadata**
Basic metadata is extracted about each MEP, such as nationality, political affiliation with the European Parliament and with the national parties. As shown in Figure 3, the information is split into 3 categories:

- meps.csv: basic information about the MEP
- national_parties.csv: political affiliation of the MEP in his/her country
- political_groups.csv: political affiliation at the European Parliament

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

[3]http://www.europarl.europa.eu/meps/en/map.html

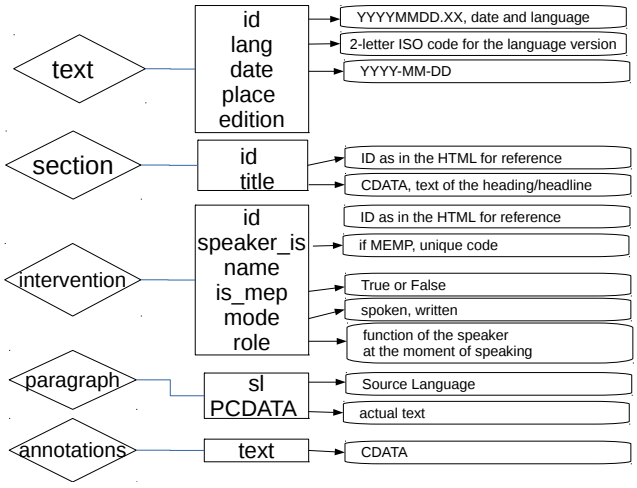

Figure 2: Metadata structure for the proceedings. The words in the diamonds represent the tags and the words in the squares the attributes under each tag. The third column contains the description of each attribute.

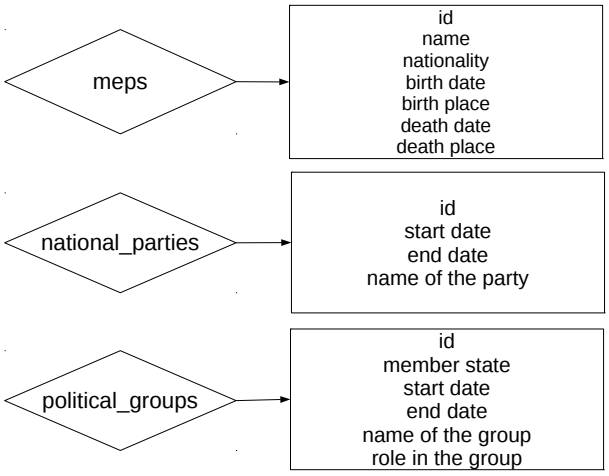

Figure 3: Metadata structure for information about MEPs. The words in the diamonds represent the tags and the words in the squares the attributes under each tag.

It should be noted that not all speakers before the European Parliament are MEPs. There are also members of other European institutions, representatives of national institutions, guests, etc. There is currently no metadata for them, but the information about these speakers is extracted from the proceedings. For each proceeding in XML we retrieve all interventions whose speaker is an MEP. Then we add relevant speaker's metadata to the intervention.

3.3. Sorting the Data

Since our goal is to achieve maximum quality of the data obtained, we employ a series of sorting and filtering techniques to clean the data and preserve the utterances that best serve our tasks.

As a first step, we filter out text units not in the expected language. Interventions sometimes remain untranslated and thus their text appears in their original language. In order to avoid this noise, we identify the most probable language of each text unit and remove the paragraphs which are not in the expected language (e.g. Bulgarian fragments found in the English version) using the Python language identifiers *langdetect*[4] and *langid*[5] and a series of heuristics.

Secondly, we filter out interventions to preserve only sentences by native speakers. A native speaker is defined here as someone holding the nationality of a country with the source language as official language. For English we filter MEPs whose origin is United Kingdom, Ireland or Malta, for German Germany, Austria, Belgium, Luxembourg and Italy, and for Spanish Spain.

An optional step is to perform Part-of-Speech tagging and lemmatization using TreeTagger (Schmid, 1994).

3.4. Sentence Alignment

The creation of a parallel corpus requires sentenced-aligned data with one sentence per line. For this task we employ hunalign (Varga et al., 2005), an automatic sentence aligner. First, we split the text into sentences using NLTK's Punkt tokenizer. Then, we extract the text from the XML files and write to files one sentence per line based on intervention, in the format of filename.intervention_id.lang e.g. *1999720.2-202.en*. This is particularly important for alignment quality as each intervention is a small text unit and aligning a few sentences per time yields higher accuracy than aligning a full text. Before aligning the sentences, we tokenized the text using Moses tokenizer with the specific setting for each language. Then, the interventions are aligned. Since we wish to obtain the highest possible quality, we set a confidence threshold of 30 for the aligned sentences and rerun the alignment based on the dictionary built in the first alignment round. A numerical ladder file is created, based on which we perform the final alignment on the untokenized files. Finally, the resulting alignments are concatenated in one file for the source and one for the target language to create the parallel corpus.

3.5. Corpus Structure and Statistics

The corpus is structured according to the steps followed for its compilation. For every step, the files generated are stored in a specified folder so that they can be used for any suitable task. At the time of compilation, the corpus consists of 1077 files for English, German and Spanish, while the parallel and the comparable corpora are in a one-file raw format that can be used directly for training an MT system. The final corpus structure is shown in Table 1.

The statistics for the comparable and the parallel corpus for the three supported languages are presented in Table 2 and Table 3. In Table 2, the language identifier method filters out texts not in the required language, while still preserving a large amount of data. The application of factors relevant for translation, both for the comparable and the parallel corpora provides us useful information about the language preferences of the speakers in the Parliament. Of course, neither all sentences in a specific language are produced by native speakers of this language, nor all sentences are translated into all languages. For this reason, filtering

[4]https://pypi.python.org/pypi/langdetect

[5]https://pypi.python.org/pypi/langid

Directory	Description
html	The crawled proceedings and MEPs' information in HTML
metadata	MEPs' metadata in CSV
txt	Raw text of the proceedings
xml	Proceedings transformed from HTML to XML
xml_langid	Proceedings in XML where the text not in the expected language is filtered out
xml_metadata	Proceedings in XML with added MEPs' metadata
xml_sentences	Proceedings in XML where text is split into sentences
xml_translationese	Proceedings in XML filtered by factors relevant for translation – original, translation, native speaker For each language *a*, it contains · the originals in *a*, · the originals in *a* only by native speakers, · all translations from any language into *a* and · all translations into *a* from a specific *SL* where the speakers are native speakers of the *SL*
xml_ttg	PoS-tagged and lemmatized proceedings in XML
raw_parallel	For each language the corresponding parallel corpora
raw_comparable	For each language the comparable corpus of original texts by native speakers

Table 1: Corpus structure

	EN		DE		ES	
	words	sents	words	sents	words	sents
html	95.21 M	5.11 M	91.48 M	5.25 M	97.08 M	5.19 M
xml	95.60 M	5.11 M	92.43 M	5.27 M	97.33 M	5.17 M
langidfilter	65.55 M	3.23 M	40.23 M	2.63 M	51.32 M	2.49 M
translationese_orig	19.69 M	0.84 M	11.74 M	0.68 M	10.75 M	0.37 M
translationese_native	8.67 M	0.37 M	7.86 M	0.42 M	5.66 M	0.18 M

Table 2: Statistics of the comparable corpora after every processing step

	EN→DE		EN→ES	
	words	sents	words	sents
all	42.08 M/38.93 M	1.91 M	42.11 M/44.21 M	1.87 M
translationese_orig	6.43 M/6.22 M	296.7 K	5.75 M/6.18 M	249 K
translationese_native	3.18 M/3.10 M	137 K	2.93 M/3.15 M	125 K

Table 3: Statistics of the parallel corpora after every processing step

sentences produced in the original language (non-translated texts) shows that only 20%-30% of the sentences in the supported languages are originals, while around 50% of the originals are produced by native speakers. In spite of this, the pipeline described above still provides us with a high quality and significant in size dataset, useful for a variety of applications.

4. Possible Applications

A corpus as described in this paper is a valuable resource for various kinds of applications. One application is machine translation, for which a metadata-rich corpus allows a more principled data selection, which in some cases has been shown to be more beneficial than using all the data available both for phrase-based as well as neural machine translation (Axelrod et al., 2011; Gascó et al., 2012; van der Wees et al., 2017).

Another application is human translation, e.g. modelling translational choice. Using the EuroParl-UdS, in our ongoing research we employ the noisy channel model as commonly applied in machine translation. According to Equation 1)

$$\arg\max_t p(t|s) = \arg\max_t p(s|t)p(t) \quad (1)$$

translation is described by maximizing the product of the probability of a TL expression t given a SL expression s by maximizing

(i) the probability of a SL expression s given a TL expression t and

(ii) the probability of a TL expression t on its own, i.e. without being conditioned by s.

This matches exactly the human translator's goal of reaching a high level of translation adequacy by maximizing the

fidelity to SL (i.e. high likelihood that the SL expression is a match for a particular TL expression) and the conformity with TL expectations (i.e. high probability of the chosen translation solution in the context of the TL) and can therefore be taken as a basis for modelling human translational choice (Teich and Martínez Martínez, forthcoming). Furthermore, we employ the corpus in studies of translation entropy, comparing the range and distribution of translation options in professional productions in EuroParl-UdS with learner translations for analysis of translation difficulty in different translation learner groups (Martínez Martínez and Teich, 2017).

5. Summary and Conclusion

We have presented an approach to building and processing parallel corpora consisting of parliamentary debates of the European Parliament (EP) harvesting valuable metadata such as speaker status and translation direction. Existing corpora built from EP proceedings do not contain such metadata, which impedes their use in translation studies or variationist linguistic analysis. We have shown our approach at work for English into German and English into Spanish parallel corpora as well as corresponding monolingual comparable corpora, but the approach is generic and can be applied to any language pair.

A metadata-rich resource such as the EuroParl-UdS is valuable for various NLP tasks and it is crucial for the advancement of insights into the process of human translation, where we need to know as much as we can about the production conditions, including the status of a given text (original vs. translation) and information about the speaker. In addition, our complete and fully documented pipeline can be easily used to compile metadata-rich or raw, parallel and comparable corpora for various linguistic applications. The corpus is available at CLARIN-PID[6] under licence CC-BY-SA-NC-4.0; the scripts are available on GitHub at https://github.com/hut-b7/europarl-uds.

6. Acknowledgements

We would like to thank our colleague José Manuel Martínez Martínez for his support - by providing us his scripts (https://github.com/chozelinek/europarl) for crawling the data - while building this resource.

7. References

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–245.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Calzada Pérez, M., Marín Cucala, N., and Martínez Martínez, J. M. (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos.

Cartoni, B. and Meyer, T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *LREC, 2012*, page 6, Istanbul.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27(1):23–42.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does More Data Always Yield Better Translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 152–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the europarl corpus for linguistic applications.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86, Phuket, Thailand. AAMT.

Koppel, M. and Ordan, N. (2011). Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326. Association for Computational Linguistics.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT-Summit XII*, pages 81–88.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4):557–570.

Lembersky, G., Ordan, N., and Wintner, S. (2012a). Adapting translation models to translationese improves smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265. Association for Computational Linguistics.

Lembersky, G., Ordan, N., and Wintner, S. (2012b). Language Models for Machine Translation: Original vs. Translated Texts. *Comput. Linguist.*, 38(4):799–825, December.

Martínez Martínez, J. and Teich, E. (2017). Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In L. Cercel,

[6] http://hdl.handle.net/21.11119/0000-0000-D5EE-4

et al., editors, *Kreativität und Hermeneutik in der Translation*, pages 403–426.

Nisioi, S., Rabinovich, E., Dinu, L. P., and Wintner, S. (2016). A Corpus of Native, Non-native and Translated Texts. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Rabinovich, E., Wintner, S., and Lewinsohn, O. L. (2015). The Haifa Corpus of Translationese. *CoRR*, abs/1509.03611.

Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., and Wintner, S. (2017). Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees.

Teich, E. and Martínez Martínez, J. (forthcoming). Translation, entropy, cognition. In Arnt Lykke Jakobsen et al., editors, *Routledge Handbook of Translation and Cognition*.

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

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2):271–281.

van der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1400–1410.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

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

Yang, T.-I., Torget, A. J., and Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, Stroudsburg, PA, USA. Association for Computational Linguistics.