Customization of the Europarl Corpus for Translation Studies

Zahurul Islam and Alexander Mehler

AG Texttechnology
Institut für Informatik
Goethe-Universität Frankfurt
zahurul, mehler@em.uni-frankfurt.de

Abstract

Currently, the area of translation studies lacks corpora by which translation scholars can validate their theoretical claims, for example, regarding the scope of the characteristics of the translation relation. In this paper, we describe a customized resource in the area of translation studies that mainly addresses research on the properties of the translation relation. Our experimental results show that the *Type-Token-Ratio* (TTR) is not a universally valid indicator of the *simplification* of translation.

Keywords: Translation studies, Europarl corpus, translation relation

1. Introduction

In recent years, corpus based research has become very popular among scholars in the area of translation studies; it has undergone a rapid development in linguistic investigation. As Laviosa (1998, p:474) puts it: "the corpus based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation". Many translation scholars have described properties of the translation process itself as well as of the relation between source and target texts of translations. Recently, scholars in this area identified several properties of the translation relation with the aid of corpora (Baker, 1996; Olohan, 2001; Laviosa, 2002; Hansen, 2003; Pym, 2005). These properties are subsumed under four keywords: *explicitation*, *simplification*, *normalization* and *levelling out*. (Pastor et al., 2008; Ilisei et al., 2009; Ilisei et al., 2010), for example, provided empirical evidence for properties of the translation relation using a comparable corpus of English and Spanish in the medical domain. Obviously, corpora of this sort, which focus on a single language pair, are not adequate for claiming universal validity of properties of the translation relation. Currently, scholars in the area of translation studies lack corpora by which they can validate their theoretical claims, for example, regarding the scope of characteristics of the translation relation. This scope is obviously affected by the membership of the source and target languages to language families. Though the exploration of universally valid characteristics of translations is an important topic, there are not many resources for testing corresponding hypotheses. In this paper, we present a customized corpus for translation studies using the format of the *Text Encoding Initiative* (TEI) (TEI Consortium, 2008) that addresses this deficit. It can be used by translation scholars as a resource for testing their hypotheses empirically. Our experimental results in Section 5., by which we exemplify the usage of this corpus, show that, for example, the *Type-Token-Ratio* (TTR) is not a universally valid indicator of *simplification* of translations – in contrast to what has been claimed in numerous translation studies.

There are many parallel and multilingual corpora available nowadays. Most of them are not useful for translation studies immediately as they require customization. In this paper we present such a customized resource in which the languages of all source texts and their translations are specified sufficiently. Using TEI P5, we provide this corpus in a way that the features of the translation relation can be learnt automatically by using machine learning techniques. The resource that we provide is a customized version of the well-known *Europarl* corpus (Koehn, 2005). We have chosen the *Europarl* corpus because it is one of the biggest multilingual corpora that are freely available to date. The *Europarl* corpus is widely used in many corpus-based applications of natural language processing, especially in the domain of *Statistical Machine Translation* (SMT). A central feature of this corpus is that it provides information on sentence-related alignments that can be explored for characteristics of the translation relation. These characteristics can finally be used for providing valid classifications of source texts and their translations.

In summary: we provide a customized resource in the area of translation studies that mainly addresses research on properties of the translation relation. In this way, the paper presents a resource in conjunction with an evaluation of its usefulness in the area of translation studies.

The paper is organized as follows: Section 2. discusses related work followed by a brief description of the *Europarl* corpus in Section 3.. Section 4. describes the details of how we customized the *Europarl* corpus for translation studies. Experimental results on the usefulness of this corpus are presented in Section 5.. Finally, the paper concludes in Section 6..

2. Related work

At the beginning of corpus based translation studies, Baker (1995) distinguished three types of corpora that are suitable for empirical research on translations, namely: *comparable corpora*, *parallel corpora* and *multilingual corpora*[1].

[1] As Baker (1995, p:232): " sets of two or more monolingual corpora in different languages, built up either in the same or dif-

Fernandes (2006) revisited Baker's typology and rejected the necessity of *multilingual corpora* in translation studies. He claims that Baker's tripartite classification can be rearranged under the categories of *comparable* and *parallel* corpora. As a reason for this binarism, Fernandes (2006) claims that the term *multilingual* is not contrastive enough to distinguish corpora from both other categories. Moreover, he argues that corpus size is relativized by qualitative aspects, which are sometimes more relevant than quantitative ones. Fernandes (2006) also introduced the attribute *multi-directional* to denote corpora of more than two languages, where the translation direction between language pairs is not predetermined. In this line of terminology, Olohan (2004) focused on *comparable* and *parallel* corpora.

Another example is Lzwaini (2003) who presented a specialized corpus of three languages (English, Arabic and Swedish) in the domain of *Information Technology* (IT). The corpus contains user manuals in English, the online help of Windows 98 and of Microsoft Office 2000 together with their translations into Arabic and Swedish. Additionally, Lzwaini (2003) harvested bilingual text from websites of IT companies. Another example is the *Translational English Corpus* (TEC), which contains contemporary translational English (Baker, 2004). The TEC was designed for the purpose of studying translations whose target language is English. It comprises texts of four types of a variety of European and non-European source languages and contains: fiction, biography, newspaper articles and in-flight magazine all of which were translated by native speakers of English. These corpora do not require any kind of customization.

Linguistic annotation is important to build reference corpora for translation studies. Hansen and Teich (2002) show how to build a reference corpus that contains such annotations. They also discuss typical problems that occur in translations from English to German and to French.

None of these corpora contains texts of more than two languages. Thus, claims about the universal validity of properties of the translation relation cannot be tested by means of these corpora. A central deficit of these corpora is that they disregard the diversity of language families and subfamilies. Thus we are in need of a *multilingual* and *multi-directional* corpus in order to validate hypotheses in this field of research. In this paper, we describe such a multilingual and multi-directional corpus that can be used as an empirical basis of research on the characteristics of the translation relation. This approach is in support of rehabilitating the original tripartite classification of Baker (1995).

3. The Europarl corpus

The *Europarl* corpus is a *multilingual*, *parallel* corpus that has been collected from the proceedings of the *European Parliament* since 1996. Koehn (2005) built the corpus to get training data for SMT. Currently, the corpus contains about 50 million words for each of the 21 official languages of the *European Union* (EU). Language pairs are selected among these 21 languages. Translation is done by official translators of the EU, who are native speakers of the corresponding target language (van Halteren, 2008). The corpus is annotated with `<CHAPTER id>` to identify documents, with `<SPEAKER id name language>` to identify source languages and with `<P>` to segment paragraphs. The procedure of corpus collection is described in Koehn (2005). Sentences in the *Europarl* corpus are aligned using the sentence alignment algorithm described in Gale and Church (1993).

4. Translation corpus extraction

In this section, we describe the procedure of customizing the *Europarl* corpus for translation studies – see Figure 2 for a visual depiction of this procedure. Although being available since 2001, this corpus has not been used by translation scholars. A reason might be its deficient customization regarding the task of corpus-based translation studies. Our goal is to customize this corpus in a way that translation scholars can use it without further effort in preprocessing. Note that the *Europarl* corpus is diverse as it contains texts from 21 languages of 7 language (sub-)families. Table 1 shows these languages and their family memberships. Figure 2 outlines the procedure of corpus customization. The following subsections describe this procedure in more detail.

4.1. Language pair selection

The selection of language pairs is decisive when building a multilingual and multi-directional corpus that reflects the diversity of natural languages. There is neither theoretical nor practical research in the field of corpus-based translation studies on how to select such pairs for building *multilingual* parallel corpora. Our paper addresses this deficit. Our intention is to make the corpus as diverse as possible by considering a broad range of language (sub-)families. Some of the language pairs that we have chosen cross the borders of language families; some of them belong to the same family. We also paired languages with small numbers of source sentences (e.g., Lithuanian and Estonian). Not all possible language pairs are considered yet, but the number of language pairs will be extended in future experiments. Figure 1 shows the ordered languages pairs and language (sub-)families. Table 2 shows all language pairs that we have selected together with the corresponding numbers of sentences of that pair.

4.2. Sentence alignment

The *Europarl* corpus comes as plain text with additional marking of document, speaker and paragraph ids. The *Europarl* community also provides a preprocessor (e.g., a sentence splitter) together with a sentence aligner based on Gale and Church (1993). The details of the preprocessor are described in Koehn (2005). *Sentence alignment* is the first step of corpus customization that starts with reading language pairs from the input corpus (see Figure 2). Source texts and their translations are iteratively processed to align their sentences. In this stage, the sentence splitter is used to detect sentence boundaries. This step of *sentence alignment* may generate empty lines in cases where the the sentence alignment failed. We removed these lines from the output of the *sentence alignment* step. As an output, we get

ferent institutions on the basis of similar design criteria"

Language (sub-)family	Language names
Germanic	English, German, Dutch, Danish and Swedish
Romance	French, Italian, Spanish, Portuguese and Romanian
Slavic	Czech, Bulgarian, Polish, Slovak and Slovenian
Baltic	Latvian and Lithuanian
Finnic	Finnish and Estonian
Ugric	Hungarian
Hellenic	Greek

Table 1: Sub-families of languages and their members that are instantiated in the customized version of the *Europarl* corpus.

Language Pairs	Sentences	Language Pairs	Sentences
German – English	44,453	Lithuanian – Estonian	1,213
English – French	42,057	Greek – Polish	1,200
French – German	30,426	Czech – Swedish	1,154
Dutch – Italian	26,419	Hungarian – Bulgarian	948
Portuguese – Danish	12,632	Estonian – Slovak	742
Spanish – Dutch	11,694	Slovak – Latvian	319
Swedish – Finnish	10,667	Bulgarian – Romanian	120
Italian – Spanish	8,892	Latvian – Slovenian	108
Danish – Greek	4,708	Finnish – Hungarian	74
Polish – Portuguese	3,997	Slovenian – Lithuanian	47
Romanian – Czech	3,381		

Table 2: Ordered language pairs in the customized version of the *Europarl* corpus and their respective number of sentences.

for each pair of languages enumerated in Table 2 a separate parallel corpus with aligned sentences that preserve information about the corresponding source and target language.

4.3. Extracting source sentences and their translations

The next step is to extract source sentences and their translations. The *Europarl* corpus is annotated with information about the speaker and his or her native language. Note that according to van Halteren (2008) this information is not available for all sentence pairs where it may be missed on either side of the translation. To circumvent this problem of missing annotations, we solely extracted pairs of sentences for which the source language marker is available. As an output of this customization step, we got $205,245$ sources sentences together with their corresponding translations. Table 2 summarizes the results of the customization step.

4.4. TEI Generation

The next step of corpus customization is to provide the data in a machine readable way that can be easily processed by scholars in the area of translation studies. We use TEI P5 (TEI Consortium, 2008) for this task. Listing 1 shows a sample of the customized corpus. The customized corpus consists of a single file in which information about the source and target language of a translation is specified in the header of each `<TEI>` section. The complete corpus contains $205,251$ segments and, thus, $205,251$ sentences of 21 languages and their translations.

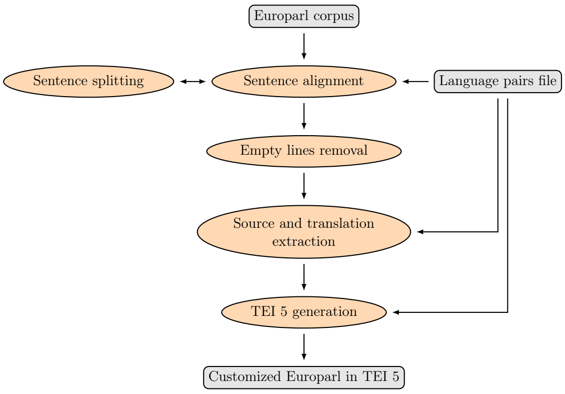

Figure 2: Building a customized version of the *Europarl* corpus: extraction steps.

5. Experiment

To measure the usefulness of our resource, we provide an experiment on testing the validity of TTR as an indicator of the simplification of translations (Baker, 1993; Baker, 1996; Hansen, 2003) (see Section 1.). Simplification describes the translator's tendency to make a translation simpler and more readable than its source. From this point of view it is hypothesized that translations tend to be more repetitive concerning lexical organization where the *Type-Token-Ratio* (TTR) is used as a quantitative measure of the lexical repetitiveness of texts. It is also hypothesized that the *average sentence length* (ASL) of translations tend to be shorter compared to their sources. Ilisei et al. (2009; Ilisei et al. (2010) provide empirical evidence on simplification as a characteristic of translations (Baker, 1993; Baker, 1996; Hansen, 2003). They examined 21 features, 9 of which including the TTR, to measure the tendency to simplification in translations. As an outcome of their study, they show that 9 features provide a significant improvement of classification in terms of higher F-scores.

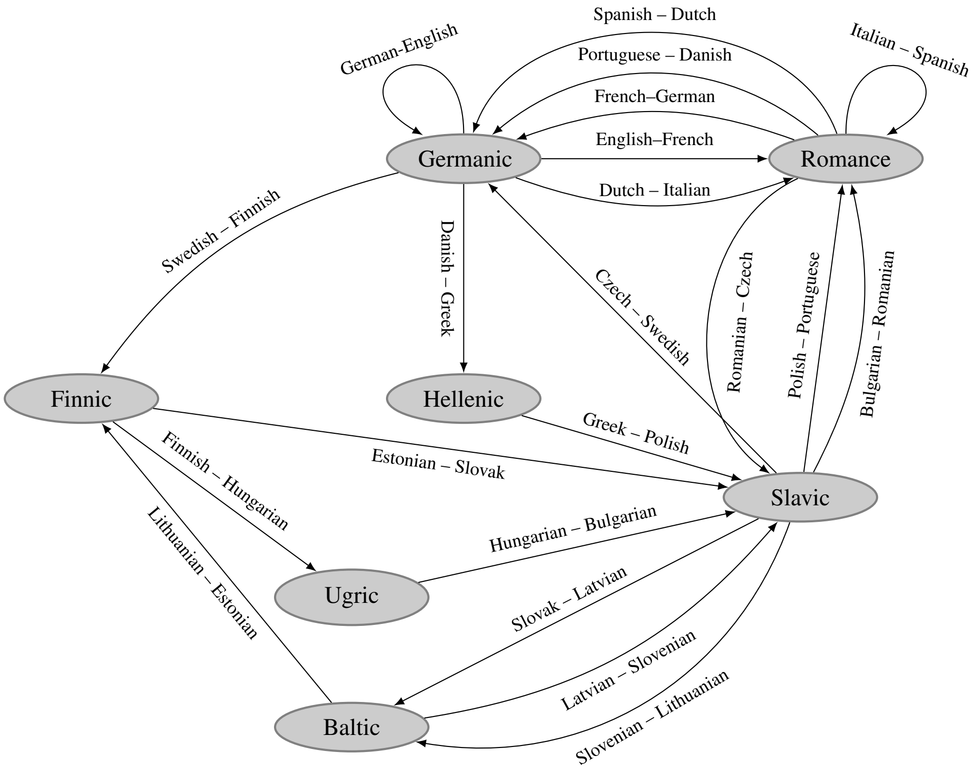

Figure 1: Language (sub-) families and language pairs

Listing 1: Corpus sample

```
<TEI>
 <teiHeader>
  <srcLang>de</srcLang>
  <trnLang>en</trnLang>
 </teiHeader>
 <documents>
     <document title=”ep−09−12−15−015”>
        <segment id=”42060”>
                  <srcSent id =”3”>Diese Antwort haben wir im Prinzip gerade erhalten .</srcSent>
                  <trnSent id =”3”>We have just received thisresponse in principle .</trnSent>
        </segment>
     </document>
  </documents>
</TEI>
```

We get similar results when examining instances of the German-English pair, that is, when making source texts and their translations input to a classification experiment based on TTR and related features. As many documents in the customized corpus contain single sentences, we performed a segment-based classification rather than a document-based one. This includes 445 segments and their translations where each segment contains 100 sentences. We use *Support Vector Machine*s (SVM) for our supervised classification. A revised implementation of the SVM is *libSVM* (Chih-chung Chang, 2011) that is part of the *WEKA* (Hall et al., 2009). We use *libSVM* with its default settings in *WEKA*. For evaluation purposes, we perform a *10-fold cross validation*. Our evaluation results in Table 3 show that the TTR is indeed a very good classificatory feature in the case of the German-English language pair, even if we combine it with ASL. We found similar results for the German-French language pair.

Feature	Accuracy	F-score
SL	75.2%	75.1%
TTR	98.5%	98.5%
SL + TTR	77.5%	77.5%

Table 3: Evaluation of German-English source and translation classification.

However, things look different if we use our customized corpus as a whole, that is, if we explore all $205,245$ segments of all 21 languages. In this case, the F-score is no longer increased when using the TTR as a classificatory feature. Table 4 shows the evaluation results. From this experiment we conclude that TTR is a useful feature for considering the German-English or German-French language pair, but fails as a universally valid indicator of the tendency to simplification in translations. This experiment exemplifies a procedure that may be followed to test related hypotheses about the expressibility of quantitative indica-

Feature	Accuracy	F-score
SL	64.4%	64.3%
TTR	52.2%	41.0%
SL + TTR	64.5%	64.4%

Table 4: Results of evaluating a classifier of source texts and their translations as collected by the customized *Europarl* corpus.

tors of characteristics of the translation relation to which our customized corpus provides indispensable data.

6. Conclusion and Future Work

The field of translation studies requires a specialized corpus that contains source and translation sentences from many languages of many language (sub-)families to validate scholars' theoretical hypotheses. In this paper, we provide a customized corpus that mainly addresses research on properties of the translation relation. In addition this paper presents a resource in conjunction with an evaluation of its usefulness in the area of translation studies.

The corpus that we provide is a customized version of the well known *Europarl* corpus. It contains $205,245$ source and translation sentence pairs from 21 languages of 7 language (sub-)families. Thus, this is a suitable resource by which translation scholars can validate their theoretical claims. Nevertheless, this corpus is opening a new window for translation scholars because they can now experiment with alignment-related features to classify sources and their translations.

As future work, there is no alternative than adding more parallel texts from different languages of different language (sub-)families to make a unique resource for translation studies. Annotation on the level of word alignment and linguistic information (e.g. POS) will be added to support the exploration of varieties of translation features.

7. Acknowledgements

We would like to thank Andy Lücking and Armin Hoenen for fruitful suggestions and comments. We also thank three anonymous reviewers. This work is funded by the LOEWE Digital-Humanities project in the Goethe-Universität Frankfurt.

8. References

Mona Baker. 1993. Corpus linguistics and translation studies - implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology. In Honour of John Sinclair*, pages 233–354. John Benjamins.

Mona Baker. 1995. Corpora in translation studies. an overview and suggestion for future research. *Target*, 7(2):223–243.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pages 175–186. Amsterdam & Philadelphia: John Benjamins.

Mona Baker. 2004. A corpus-based view of similarity and difference in translation. *International Journal of corpus Linguistics*, 9(2):167–193.

Chih-jen Lin Chih-chung Chang. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(2):27:2–27:26.

Lincoln Fernandes. 2006. Corpora in translation studies: revisting Baker's typology. *Fragmentos: Revista de Língua e Literatura*, 30:87–95.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1).

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Silvia Hansen and Elke Teich. 2002. The creation and exploitation of a translation reference corpus. In *First International Workshop on Language Resources for Translation Work and Research, 3rd International Conference on Language Resources and Evaluation (LREC-2002).*

Silvia Hansen. 2003. *The Nature of Translated Text: An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Ph.D. thesis, University of Saarland.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2009. Towards simplification: A supervised learning approach. In *Proceedings of Machine Translation 25 Years On, London, United Kingdom, November 21-22.*

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov, 2010. *Identification of translationese: A machine learning approach*, pages 503–511. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit.*

Sara Laviosa. 1998. The corpus-based approach: A new paradigm in translation studies. *journal des traducteurs / Meta: Translators' Journal*, 43(4):474–479.

Sara Laviosa. 2002. *Corpus-based translation studies. Theory, findings, applications*. Amsterdam/New York: Rodopi.

Sattar Lzwaini. 2003. Building specialised corpora for translation studies. In *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics*.

Maeve Olohan. 2001. Spelling out the optionals in translation:a corpus study. In *Corpus Linguistics 2001 conference. UCREL Technical Paper number 13. Special issue.*

Maeve Olohan. 2004. *Introducing Corpora in Translation Studies*. London/New York: Routledge.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08).*

Anthony Pym. 2005. Explaining explicitation. In *New Trends in Translation Studies. In Honour of Kinga Klaudy*, pages 29–34. Akadémia Kiadó.

TEI Consortium. 2008. TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/Guidelines/P5/.

Hans van Halteren. 2008. Source language markers in europarl translations. In *COLING*, pages 937–944.