Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies

Bruno Cartoni[1] **, Thomas Meyer**[2]

(1)Linguistic Department, University of Geneva, 2, rue de Candolle, CH – 1211 Geneva
(2)Idiap Research Institute, Rue Marconi 19, CH – 1920 Martigny
bruno.cartoni@unige.ch, thomas.meyer@idiap.ch

Abstract

Translation studies rely more and more on corpus data to examine specificities of translated texts, that can be translated from different original languages and compared to original texts. In parallel, more and more multilingual corpora are becoming available for various natural language processing tasks. This paper questions the use of these multilingual corpora in translation studies and shows the methodological steps needed in order to obtain more reliably comparable sub-corpora that consist of original and directly translated text only. Various experiments are presented that show the advantage of directional sub-corpora.

Keywords: parallel corpora, comparable corpora, translation studies

1. Introduction

Cross-linguistic studies nowadays rely heavily on corpus data, and they often favor the use of parallel corpora for obvious practical reasons. The parallel corpora used in this context are consequently bilingual (sometimes trilingual), and their translated side is clearly identified. Recently, more and more multilingual corpora have become available but they are not primarily designed for cross-linguistic analysis. In this paper, we describe how we pre-process a large multilingual corpus, Europarl (Koehn, 2005), in order to extract specific *directional* sub-corpora. These directional corpora can then be used as comparable corpora which in turn are useful for translational studies. The paper starts with a description of the usage of various kinds of corpora in cross-linguistic studies (Section 2), before introducing the Europarl corpus in Section 3. In Section 4 we present the pre-processing and conversion of this corpus. The results of this process are summarized in Section 5. Section 6 provides examinations and comparisons of source and translated texts that are possible with directional corpora and the comparable corpora extracted from them. The latter are used to confirm hypotheses in terms of specificities of translated text that have not yet been empirically verified so far.

2. Motivation

Cross-linguistic studies, such as contrastive analysis or translation studies, rely more and more on corpus data. Most of the time, the favored corpora are bilingual, either comparable or parallel. The latter are translational corpora (with either source and target language or consisting of both directions). As in any examinations in corpus linguistics, it is crucial to take into consideration a certain number of meta-information, such as discourse genre, origin of the text and status of the languages (i.e. original vs. translated). Multilingual corpora have especially become available with research in Computational Linguistics (mostly in Machine Translation). These new resources are very valuable for cross-linguistic studies: they potentially contain many parallel corpora for different language pairs. In addition, they are very large, often of the same register and often contain comparable 'translated' corpora (translated texts in one single language, but translated from different languages).

Exploiting multilingual data for cross-linguistic studies requires methodological pre-caution, as meta-information on the data is not always directly available (see Section 4). The multilingual perspective also provokes changes in the 'classical' definition of corpora normally used in corpus studies. Section 2.1 reviews the classical distinctions existing to characterize corpus data and the discussion (see Section 5) shows how to adapt the current scheme in order to integrate new types of resources such as the comparable corpora we extract from Europarl.

2.1. Classical Corpus Typology

In past studies, and mainly since corpora have been used in cross-linguistic research, different typologies or terminological distinctions have been proposed to clarify different types of textual data. In a recent paper, Granger (2010) attempts to unify the terminology for various types of corpora used in cross-linguistic research. In this typology, Granger refuses to use the term 'parallel', because it has been ambiguously defined in the literature. It also has to be noticed that for translational corpora, Granger makes the assumption that bidirectional corpora are 'directional' in respect to source text and translated text being clearly identified (we will see that this is not the case for the multilingual corpus Europarl examined in this paper).

Granger also distinguishes monolingual and multilingual comparable corpora. For the monolingual case, she separates monolingual comparable corpora (made of translated vs. original texts and mainly used in translation studies) from comparable corpora of native vs. learner text. Among multilingual comparable corpora, she distinguishes corpora made of translated texts (that are translations in different languages of the same original texts) from comparable corpora made of original texts (that is the first original type of comparable corpora). As we will explain in the following,

multilingual corpora such as Europarl can actually contain different kinds of sub-corpora that were not included in the typology proposed above.

3. The Europarl Corpus

The Europarl corpus is a freely available corpus, composed of the proceedings of the European parliament debates. It is processed in order to be segmented and aligned by pairs of languages. It is mainly used in Statistical Machine Translation (SMT) training. More information on the construction of the corpus is given in (Koehn, 2005). When provided to the user, the corpus is made of files that contain the minutes of each day of debates (one file per day). Each deputy of the parliament can speak in his/her own language, and each statement is then translated into the other official languages of the European Union. The original language of the statements is given as meta-information, although rather scarcely. For instance, the statement presented in Figure 1 is introduced by the tag 'SPEAKER', containing a specific 'ID', the name of the deputy, and the language in which the statement has been made (the 'LANGUAGE' tag). The figure thus shows a statement originally made in Italian and translated into English. The official languages

<SPEAKER ID=6 LANGUAGE='IT' NAME='Segni'>
Madam President, coinciding with this year's first part-session of the European Parliament, a date has been set, unfortunately for next Thursday, in Texas in America, for the execution of a young 34 year-old man who has been sentenced to death. We shall call him Mr Hicks. <P>

Figure 1: Deputy statement in the Europarl corpus

have changed throughout the years in the European Parliament, as the number of members has grown. From 4 official languages in 1958 (Dutch, French, German and Italian), the corpus increased gradually with the extension of the European Union (11 languages in 1995), and contains now, in the latest version, 23 languages, but obviously not in the same proportion in terms of number of statements (see Table 3)[1].

3.1. Statistics

According to the statistics provided on the Europarl corpus website, the entire corpus contains 592'894'105 tokens and 25'601'461 sentences. As stated above, some of the deputies' statements are tagged with language information. Table 1 provides figures for these language tags. As shown, only 66.53% of the statements contain a language tag. When comparing the files in different languages, a language tag is sometimes inconsistent, i.e. it can be there in a text file of one language but not in the file of another language (we counted 6619 such divergencies). As a consequence, in total, only 118'289 statements have a proper language tag.

[1]The Europarl Corpus v6, as available on http://www.statmt.org/europarl/, contains the debates from 1996 until now.

Nbr. of statements (in all languages)	187'720
Nbr. of LANGUAGE tags	124'908
Nbr. of diverging tags	6619
Remaining trustworthy tags	118'289

Table 1: Language tags in the Europarl corpus

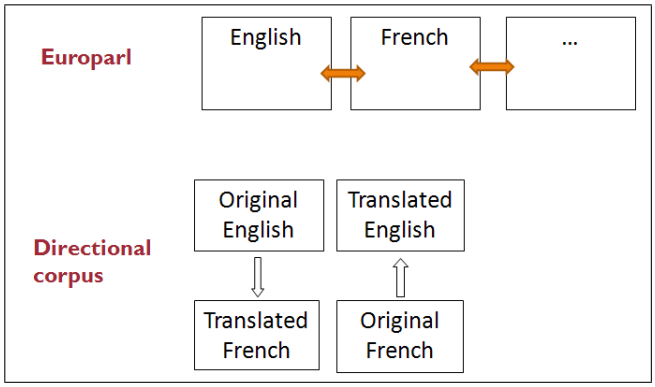

Figure 2: Directional corpora that can be obtained from Europarl

The language tags are not completely trustworthy because of mistakes, scarcity, etc. However, they *do* provide the only possible information on the original language of the statements and can therefore be used to extract 'directional corpora', as is explained in the following section.

4. Extracting Directional Corpora from the Europarl Corpus

Europarl is a multilingual corpus made of various languages translated into all other official EU languages. To this respect, it is an invaluable resource for cross-linguistic studies. But with respect to what we have seen above, the status and the use of such a corpus has to be clarified because of the different kinds of sub-corpora it might contain. This section describes how we obtained such sub-corpora and how we extracted directional corpora from the Europarl corpus.

The notion of 'directional corpus' refers to a parallel corpus where the original and the translated languages are clearly identified. Figure 2 exemplifies various directional corpora that can be extracted from the Europarl corpus.

Building directional corpora from Europarl consists in individualizing specific segment pairs that contain relevant language information.

Since language information is scarcely represented, a first step of *dissemination* needs to be performed in order to homogenize and extend the given language information across the corpus.

4.1. Dissemination of Language Tags

Language information in form of a tag can be available for a particular segment in a language file but not for the same segment of another language file. For example, the statement in Figure 1 is tagged with 'LANGUAGE=IT' (indicating that the statement has actually been produced in Ital-

ian) in the file of the English translation. This tag is however lacking for the French translation of the segment.
This is why we first gathered the given language information from all statements, in all target languages, and then 'disseminated' the information to the files where it was missing. This dissemination allows us to (i) individualize (and sometimes correct) the diverging tags – see Table 1, and (ii) to increase the number of statements in each directional pair. Table 2 below shows the increase in terms of number of statements for the English → French directional corpus.

Nbr. of statements BEFORE dissemination	19'903
Nbr. of statements AFTER dissemination	24'725
Improvement	24%

Table 2: Improvement after dissemination/correction of language tags in the Europarl corpus

This procedure allows us to obtain as many directional corpora as there are language pairs in the corpus, as exemplified in Figure 2. Still, there are two remarks to be considered on the directivity and the textual homogeneity of such directional sub-corpora.

4.2. Direct or Indirect Translation ?

Knowing whether a statement has been performed in a particular language or if it is a translation from another one does not necessarily imply that the translation has been made directly. In the context of the European parliament, because of many language pairings, this question is all the more important.
From personal discussion with a translator at the European parliament, we know that after 2003, a pivot language was used (English), which implies that all statements were first translated into English and then into the 22 other target languages. Before 2003, however, it seems that the translations were made directly from all languages into others. This is the case at least for less 'exotic' language combinations (there are probably less translators translating from Danish to Portuguese than from English into French).
In any case, this notion of direct vs. indirect translation should be taken into account when using directional corpora, especially depending on the purpose of the studies. In a translation study that aims at analyzing the translation process from one source language to one target language, such a sub-corpus might not be 'confident'. But for translation studies that aim at comparing translated texts and original texts of the same language (as we show below), the sub-corpus definitely can be of value.
Another specificity of the extracted parts has to be considered. As previously stated, Europarl is made from minutes of debates, in which every deputy speaks her/his own language. Consequently, a continuous discussion can consist of many statements in different original languages. When extracted to create parallel directional corpora, the discussion flow is broken, and the resulting corpora are made of statements that do not necessarily follow each other. Still, the extracted statements are rather long (more than one sentence), and context is preserved, though not as in the original full text.

5. Results: The Sub-corpora Extracted

Two different types of corpora can be extracted from Europarl. The first (and most obvious) type are parallel (directional) corpora, consisting of bilingual texts (original and translated texts). This first type is described in Section 5.1. A second type of corpora can also be derived from the first one: comparable corpora of various kinds, as described in Section 5.2.

5.1. Directional corpora

From the multilingual Europarl corpus, and its 592'894'105 tokens, we created, with the help of the above-mentioned dissemination procedure, an interesting set of sub-corpora that contains bi-directional parallel corpora and comparable corpora of various kinds. Table 3 presents the size of the various corpora extracted from the multilingual Europarl corpus. For this extraction, we focus only on the years 1996-1999 of the parliament debates, because they seem to be more reliable in terms of the 'directionality' of the translation (see Section 4.2).

Corpus	SL	TL
English→French	1'410'121	1'581'757
French→English	1'257'869	1'188'913
German→French	1'348'005	1'629'024
French→German	1'195'896	1'059'868
Dutch→French	846'409	946'151
French→Dutch	1'277'659	1'231'260
Italian→French	575'614	650'316
French→Italian	1'221'604	1'106'650
Spanish→French	662'788	701'551
French→Spanish	1'264'159	1'209'334

Table 3: Size of directional corpora extracted from Europarl: number of tokens in Source Language (SL) and Target Language (TL)

5.2. Comparable Corpora

As mentioned above, our extraction provides sub-corpora of various kinds. The first one, as shown in Table 3, being a number of bi-directional parallel corpora, for every language pair included in Europarl. Once the individual parallel corpora have been extracted, a second type of corpora can be considered: so-called *comparable* corpora – i.e. two different sets of texts that share common properties. Figure 3 provides a schematic view of the different comparable and parallel corpora that can be produced.
Aside from 'classical' comparable corpora made of comparable texts in different original languages, other types of comparable corpora are also available, which are of particular interest for translation studies: for example, one can examine the difference between original and translated texts (as exemplified for the French language in Figure 3)

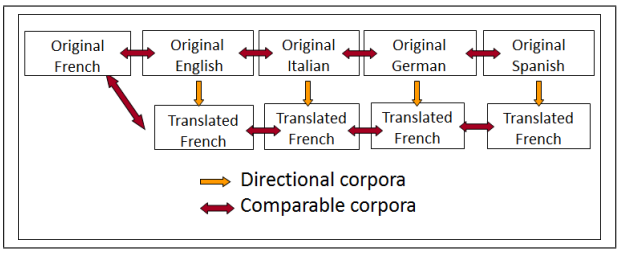

Figure 3: Different types of comparable corpora created from the directional corpora extracted from Europarl

or comparable corpora of translated texts in the same language translated from different source languages. This latter type of corpora (translated texts from different original source languages) is not mentioned in Granger's typology (see Section 2.1), and can be created from multilingual, directional corpora only.

Moreover, these comparable corpora reveal common properties of translated texts compared to other comparable corpora that generally only share the same genre, dates and topics. Here, texts are produced in the very same context and are highly similar.

6. Discussion and Possible Uses

The sub-corpora that have been extracted are new and offer interesting possibilities to confirm or test different hypotheses on so-called *translationese* (Baker, 1992; Baker, 1995) (a set of linguistic characteristics that are typical for translated texts). Indeed, the comparable corpora made of different translated texts that are all translated in the same language from different source languages, can assess and confirm some specificities of *translated text* that so far have mostly been based on intuition rather than on large-scale quantitative approaches.

In the following, we describe some experiments on comparable corpora involving French, either *translated French* translated from different languages (English, German, Italian, Spanish, Dutch), or *original French* that is built up by gathering all original French statements in opposite directional corpora (see Table 3). Table 4 summarizes figures on the number of tokens in each comparable corpus, again for years 1996-1999.

Sub-corpus	Number of tokens
Original French (OF)	2'391'806
French translated from English (EF)	1'581'757
French translated from German (DF)	1'629'024
French translated from Italian (IF)	650'316
French translated from Spanish (SF)	701'551
French translated from Dutch (NF)	946'151

Table 4: Size of comparable corpora

These experiments are based on different lexical measures, most of them being usually considered as revealing differences between translated and original language. We first show results for the well-known type/token ratio, followed by results using a lexical density measure. Finally, we summarize results on experiments on homogeneity measures applied to the comparable sub-corpora.

6.1. Type/Token Ratio

The type/token ratio is used to measure the richness of vocabulary. As stated by Baker (1995), in translation, the ratio should be higher than in original text (i.e. the vocabulary should be less rich), because it expresses *a consequence of the process of lexical simplification* (that is typical for translation). To be comparable, the type/token ratio should be measured on pieces of corpora of equal size. Table 5 shows the type/token ratio calculated over the first 100'000 token of each sub-corpus.

Corpora	Type/Token ratio
OF	0.071767
EF	0.089690
DF	0.088850
IF	0.089167
SF	0.086032
NF	0.090875

Table 5: Type/Token ratio in comparable corpora

These figures clearly show that translational corpora have a higher type/token ratio, which confirms Baker's hypothesis (Baker, 1995). Most interestingly, there is almost no difference in the ratios of the translated French corpora, whatever the source language is.

6.2. Lexical Density

As stated in (Baker, 1995), *lexical density is the percentage of lexical as opposed to grammatical items in a given text or corpus of texts*. Lexical density is related to the notion of information load. It is expected that lexical density should be smaller in a translated corpus, which would reflect that the translator tried *to control information load and to make a translated text more accessible to its new readership*. We calculated the lexical density on our sub-corpora, using all function words of the French Morphalou lexicon[2] , i.e. 262 prepositions/determiners/pronouns. The results are shown in Table 6.

Corpora	Lexical density
OF	55.33%
EF	55.59%
DF	55.20%
IF	55.11%
SF	55.05%
NF	55.64%

Table 6: Lexical density in comparable corpora

Surprisingly, there are no clear differences in original French vs. translated French. This probably reflects the homogeneity of such comparable corpora, i.e. the important

[2]A freely available morphosyntactic lexicon, available at: `http://www.cnrtl.fr/lexiques/morphalou/`

number of linguistic properties they share. However, this still questions the validity of the *lexical density* measure for distinguishing translated from original texts.

6.3. Similarities and Homogeneity Measure

Kilgarriff (2001) proposed a metric to assess the homogeneity of one corpus and/or the similarity of two corpora (both being computed the same way, homogeneity being measured by assessing the similarities of two sides of the same corpus). The metric is called *Chi-By-Degrees-of-Freedom* (CBDF). The $\chi 2$ statistic is computed over the 500 most frequent words from the two corpora to be compared. In this experiment, we limited the corpora to 200,000 words each, so that a comparison with the values given by Kilgarriff for other corpora is possible. The values are normalized by the number of degrees of freedom, which is (500–1) χ (2–1) = 499, hence the name of the measure[3]. The CBDF similarity values for 100,000-word subsets of Original French (OF), French translated from English (EF), from Italian (IF), from German (DF), from Dutch (NF) and from Spanish (SF) are shown in Table 7 below.

Taking OF vs. EF as an example, the values are computed by summing up, for all of the most frequent 500 words in OF+EF, the difference between the observed and the expected number of occurrences in each of OF and EF, more precisely (o – e)2 / e, then dividing the sum by 499. The expected number is simply the average of OF and EF occurrences, which is the best guess given the observations. The lower the result, the closer the two corpora are considered to be in terms of lexical distribution, as shown by Kilgarriff (2001).

In order to measure homogeneity, we sliced each corpus into 10 equal parts and computed the score by randomly building 10 different corpus configurations and calculating the average of the values. The similarities between sub-corpora of French translated from different source languages are shown in Table 7. The values comparing the same portion (e.g. OF/OF) indicate the homogeneity score of the respective sub-corpus.

	OF	EF	DF	IF	SF	NF
OF	**2.64**					
EF	6.00	**3.34**				
DF	5.11	4.83	**2.74**			
IF	4.88	6.30	4.99	**2.86**		
SF	5.34	5.43	5.36	4.43	**2.22**	
NF	4.62	4.29	3.14	5.22	5.43	**2.88**

Table 7: Values of CBDF ($\chi 2$ statistic normalized by degrees of freedom) for all pairs of source-specific 200,000-word subsets from Europarl. The lower the value, the more similar the subsets are.

When compared to each other, the similarity measures of the corpus pairs seem to reflect the different language families (Germanic vs. Romance) the texts are translated from. The most similar pair is Original French vs. French translated from Italian, which is not surprising given that the two languages are closely related. Also similar to OF/IF are the IF/SF pairs, the EF/DF pairs and even more noticeably the DF/NF pairs, reflecting the similarity of translations from related languages. These measures confirm that the original source language does influence the very nature of the translated text.

The measures can also be compared with similar measures performed on other corpora by Kilgarriff. For instance, the similarity score between OF and EF (6.00) is lower than all but two of the 66 pairs of corpora for which Kilgarriff has computed the CBDF value. Homogeneity values are higher than similarity values (i.e. the $\chi 2$ scores are lower). These values are again comparable, albeit clearly lower than those found by Kilgarriff, and presumably account for the lower variety of parliamentary discourse. Still, these values are similar to those of the most homogeneous subsets used by Kilgarriff, the Dictionary of National Biography (1.86) or the Computergram (2.20) (see (Kilgarriff, 2001)).

7. Conclusion

In this paper, we have shown how a multilingual corpus can be converted into specific sub-corpora that can be used in cross-linguistic studies. We have exemplified the required pre-processing that mainly consists of verifying and unifying meta-information on the corpus languages in order to clearly identify the texts of original and translated language.

The directional parallel corpora extracted with the proposed method can be used in translation studies, as exemplified in this paper, but can also be of great interest for other NLP tasks such as statistical machine translation (SMT), since the 'directionality' of the corpora in the training phase of SMT systems has an influence on the output quality (Ozdowska, 2009). These directional corpora are also in use in an ongoing SMT project (Popescu-Belis et al., 2012), both for corpus investigation and SMT training.

Among the various sub-corpora that can be extracted, we especially highlighted the comparable corpus of translated French because it contains translated French from different source languages, showing for example how the source language measurably influences the translated texts. We have further shown that the measure of type/token ratio does reveal a difference in terms of lexical richness for translated and original text, while the measure of lexical density does not seem to be appropriate in this perspective, at least in very comparable corpora.

Compared to other corpora, the Europarl corpus seems to be quite a homogeneous corpus, although some variation for translated and original text and between texts translated from different languages could be measured using a Chi-By-Degrees-of-Freedom measure.

Our experiments with the directional and comparable sub-corpora have shown the benefit of the language information based extractions for translational studies as the Europarl corpus, modulo the presented pre-processing, contains an important amount of 'highly-comparable' data.

The directional and comparable sub-corpora as presented in this paper and used for the experiments are available after

[3]This work on homogeneity and similarity was originally presented at the BUCC workshop 2011 (Cartoni et al., 2011), but new data for French translated from Dutch has been added in the experiments described in this paper.

registering at: `https://www.idiap.ch/dataset/europarl-direct`. Other language directions can be added upon requests to the authors.

Acknowledgments

We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF), under the COMTIS Sinergia Project n. CRSI22 127510 (see www.idiap.ch/comtis/).

8. References

Mona Baker. 1992. *In other words : a coursebook on translation*. Routledge, London.

Mona Baker. 1995. Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7:2:223–242.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR.

Sylviane Granger. 2010. Comparable and translation corpora in cross-linguistic research. design, analysis and applications. *Journal of Shangai Jiaotong University*.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6:1:1–37.

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

Sylwia Ozdowska. 2009. Données bilingues pour la tas français-anglais: impact de la langue source et direction de traduction originales sur la qualité de la traduction. In *Proceedings of TALN 2009*.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. In *Proceedings of LREC 2012*.