How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives

Bruno Cartoni
Linguistics Department
University of Geneva
2, rue de Candolle
CH – 1211 Geneva 4

Sandrine Zufferey
Linguistics Department
University of Geneva
2, rue de Candolle
CH – 1211 Geneva 4

{bruno.cartoni|sandrine.zufferey}@unige.ch

Thomas Meyer
Idiap Research Institute
Rue Marconi 19
CH – 1920 Martigny

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19
CH – 1920 Martigny

{thomas.meyer|andrei.popescu-belis}@idiap.ch

Abstract

In this paper, we question the homogeneity of a large parallel corpus by measuring the similarity between various sub-parts. We compare results obtained using a general measure of lexical similarity based on χ^2 and by counting the number of discourse connectives. We argue that discourse connectives provide a more sensitive measure, revealing differences that are not visible with the general measure. We also provide evidence for the existence of specific characteristics defining translated texts as opposed to non-translated ones, due to a universal tendency for explicitation.

1 Introduction

Comparable corpora are often considered as a solution to compensate for the lack of parallel corpora. Indeed, parallel corpora are still perceived as the gold standard resource for many multilingual natural language processing applications, such as statistical machine translation.

The aim of this paper is to assess the homogeneity of the widely used Europarl parallel corpus (Koehn 2005) by comparing a distributional measure of lexical similarity with results focused on a more specific measure, the frequency of use of discourse connectives. Various perspectives can be taken to assess the homogeneity of this corpus. First, we evaluate the (dis)similarities between translated and original language (Experiment 1) and then the (dis)similarities between texts translated from different source languages (Experiment 2).

Analyzing the use of discourse connectives such as *because* and *since* in English highlights important differences between translated and original texts. The analysis also reveals important differences when comparing, for a given language, texts that have been translated from various source languages. The different distribution of connectives in original vs. translated French, as well as across varieties of French translated from various source languages (English, German, Italian and Spanish), are all the more intriguing that they are not matched by a distributional difference of the general vocabulary in these corpora. We will indeed show that a well-known method (Kilgarriff 2001) designed to compare corpora finds that the original French and the various translated portions of Europarl are rather similar, regardless of their source language.

The paper is structured as follows: we first present related work on the characterization of translated text (Section 2). In Section 3, we argue that analyzing discourse connectives sheds new light on text (dis)similarity. Section 4 presents the Europarl parallel corpus and its sub-parts that have been used in our studies, as well as the methodology and measures that have been applied to assess text similarities. Section 5 presents our main findings and Section 6 discusses our results, drawing methodological conclusions about the use of parallel corpora.

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 78–86,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011. ©2011 Association for Computational Linguistics

2 Previous Work

Existing studies on translated corpora are mainly designed to automatically identify the presence of so-called “translationese” or “third code”, in other words, a text style deemed to be specific to translated texts, as in (Baroni and Bernardini 2005) or in (Ilisei et al. 2010). In the literature, many possible characteristics of translationese have been identified, such as those listed in (Baker 1996): translations are simpler than original texts (Laviosa-Braithwaite 1996); translations are more explicit than original texts due to an increase of cohesion markers (Blum-Kulka 1986); and the items that are unique in the target system (i.e. that do not have exact equivalents in the source language) are under-represented in translations (Tirkkonen-Condit 2000).

In the field of natural language processing, several studies on parallel corpora have shown that when building a statistical machine translation system, knowing which texts have been originally written in a given language and which ones are translations has an impact on the quality of the system (Ozdowska 2009). A recent study using machine learning has confirmed the universal of simplification as a feature of translated texts (Ilisei et al. 2010).Corpora can be compared using similarity measures. Most of these measures are based on lexical frequency. Kilgariff (2001) provides a comprehensive review of the different methods for computing similarity.

In this study, we chose to use the CBDF measure (Chi-by-degrees-of-freedom), as proposed in (Kilgariff 1997), to assess the similarity of our sub-corpora, as explained in Section 4.3. We compare this measure with another marker of text diversity (connectives), as explained in the following section.

3 Discourse Connectives as Markers of Text Diversity

Discourse connectives like *but*, *because* or *while* form a functional category of lexical items that are very frequently used to mark coherence relations such as *explanation* or *contrast* between units of text or discourse (e.g. Halliday & Hassan 1976; Mann & Thomson 1992; Knott & Dale 1994; Sanders 1997). One of the unique properties of discourse connectives is that the relation they convey can in many cases be inferred even when they are removed, as illustrated in (1) and (2):

1 Max fell because Jack pushed him.

2 Max fell. Jack pushed him.

The causal relation conveyed by *because* in (1) is also inferable when the connective is absent by using world knowledge about the possible relation between the fact of pushing someone and this person’s fall in (2). In other words, contrary to most other lexical items, connectives can be used or left out without producing ungrammatical results or losing important aspects of meaning. At a macro-textual level, it is however clear that a text containing no connective at all would become rather difficult to understand. Several psycho-linguistic studies have indeed stressed the role of connectives for processing (Millis & Just 1994; Noordman & Blijzer 2000). But the point we want to make here is that in most texts or discourses, some coherence relations are conveyed by the use of connectives while others are not, depending on what the author/speaker feels necessary to mark explicitly.

Another consequence of the fact that connectives are optional is that their use in translation can vary tremendously between the source and the target texts. Studies that have examined at the use of connectives in translation have indeed found that connectives were often removed or added in the target texts, and that the type of coherence relation conveyed was sometimes even modified due to the actual choice of connectives in the target system (Altenberg 1986; Baker 1993; Lamiroy 1994; Halverson 2004). For all these reasons, discourse connectives appear to be particularly interesting to investigate in relation to corpus homogeneity.

In this study, we focus more particularly on the category of causal connectives, that is to say connectives such as *because* and *since* in English. This particular category seemed especially appropriate for our purposes for a number of reasons. First, causal connectives form a well-defined cluster in many languages and can be studied comprehensively. Second, causal relations are amongst the most basic ones for human cognition and in consequence causal connectives are widely used in almost all text types (Sanders & Sweetser 2009). Lastly, causal connectives have been found to be more volatile in translation than other categories, such as for example concessive connectives like *but*, *however*, etc. (Halverson 2004; Altenberg 1986).

From a quantitative perspective, function words are usually very frequent whereas most content words tend to be in the tail of the distribution. This provides another reason to treat connectives as a key feature for assessing text similarities.

4 Corpora and Methodology

4.1 Corpora

Our analysis is based on the Europarl corpus (Koehn 2005), a resource initially designed to train statistical machine translation systems. Europarl is a multilingual corpus that contains the minutes of the European Parliament. At the parliament, every deputy usually speaks in his/her own language, and all statements are transcribed, and then translated into the other official languages of the European Union (a total of 11 languages for this version of the corpus – version 5). Based on this data, several parallel bilingual corpora can be extracted, but caution is necessary because the exact status of every text, original or translated, is not always clearly stated. However, for a number of statements, a specific tag provides this information.

From this multilingual corpus, we extracted for our first experiment two parallel and "directional" corpora (En-Fr and Fr-En). By "directional" we mean that the original and translated texts are clearly identified in these corpora. Namely, in the English-French subset, the original speeches were made in English (presumably mostly by native speakers), and then translated into French, while the reverse is true for French-English. Still, for many applications, these would appear as two undifferentiated subsets of an English-French parallel corpus.

Since language tags are scarcely present, we automatically gathered all the tag information in all the language-specific files, correcting all the tags and discarding texts with contradictory information. Therefore, these extracted directional corpora are made of discontinuous sentences, because of the very nature of this multilingual corpus. In one single debate, each speaker speaks in his/her own language, and when extracting statements of one particular language, discourse cohesion across speakers is lost. However, this has no incidence at the global level on the quantitative distribution of connectives.

We have focused our investigation on the years 1996 to 1999 of the Europarl corpus. Indeed, statistical investigations and information gathered at the European Parliament revealed that the translation policy had changed over the years. The 1996-1999 period appeared to contain the most reliable translated data of the whole corpus.

For Experiment 1, we extracted two parallel directional corpora made of two languages – French and English – in order to compare translated and original texts in both languages, as shown in Figure 1.

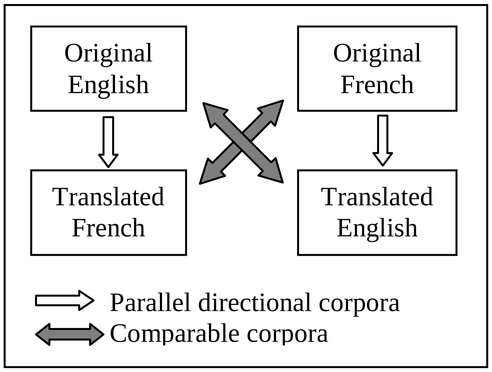

Figure 1: Parallel and comparable corpora extracted from Europarl

Table 1 gives the number of tokens in the English-French and in the French-English parallel directional corpora.

Parallel corpus	Token in ST	Token in TT
English-French (EF)	1,412,316	1,583,775
French-English (FE)	1,257,879	1,188,923

Table 1: Number of tokens in Source Texts (ST) and Translated Texts (TT) of the parallel directional corpora.

Following the same methodology, we extracted for Experiment 2 other parallel directional corpora, again with French as a target language (also from the 1996-1999 period), as shown in Figure 2. Table 2 presents the sizes of these four additional comparable corpora.

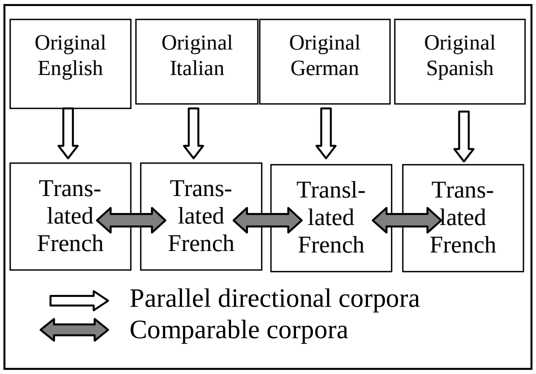

Figure 2: Parallel and comparable corpora for Translated French

Parallel corpus	Token in ST	Token in TT
German-French (DF)	1,254,531	1,516,634
Italian-French (IF)	552,242	624,534
Spanish-French (SF)	597,607	633,918

Table 2: Number of tokens in Source Texts (ST) and Translated Texts (TT) of the three additional parallel directional corpora of translated French.

These parallel directional corpora have been used as comparable corpora in our study because they are written in the same language and are of the same genre, but do not have the same "status", since some are original texts while others are translations, as shown in . Moreover, for comparison purposes, we have also used a sub-part of Europarl which was originally produced in French (noted OF), corresponding to the French part of the French-English corpus described in Table 1

All the experiments described below are based on these comparable corpora, i.e. on the translated vs. original corpus (for French and English) and on the different corpora of translated French (with Italian, English, Spanish and German as source languages).

4.2 First Measure: CBDF Measure

Following a proposal by Kilgarriff (2001), who criticizes a number of simpler techniques, we have measured corpus similarity by computing the χ^2 statistic over the 500 most frequent words from the two corpora to be compared, which were limited to 200,000 words each, so that comparison with the values given by Kilgarriff was possible. The value was normalized by the number of degrees of freedom, which is $(500-1) \times (2-1) = 499$, hence its name. As shown by Kilgarriff with artificially designed corpora, for which the similarity level was known in advance, the χ^2 statistic is a reliable indicator of similarity. Moreover, Kilgarriff (2001: Table 10, page 260) provides a table with the χ^2 values for all 66 pairs of 200,000-word corpora selected from 12 English corpora, which we will use for comparison below. The table also lists internal homogeneity values for each corpus, obtained by averaging the χ^2 statistic over each 200,000-word corpus split several times in half. In fact, as the same method is used for computing both similarity and homogeneity, only 100,000-word fragments are used for similarity, as stated by Kilgarriff.

The CBDF similarity values between 100,000-word subsets of Original French (OF), French translated from English (EF), from Italian (IF), from German (DF), and from Spanish (SF) are shown in Table 4 below. Taking OF vs. EF as an example, these 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 $(o - e)^2 / e$, and 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).

For measuring homogeneity, we sliced each corpus in 10 equal parts, and computed the score by randomly building 10 different corpus configurations and calculating the average of the values.

4.3 Second Measure: Counting Connectives

As explained above, we focused our experiments on comparing frequencies of causal connectives. For French, our list of items included *parce que*, *puisque*, *car*, and *étant donné que*. For English, we included *because*, *since*, and *given that*[1]. In the case of *since*, we manually annotated its two meanings in order to distinguish its causal uses from its temporal ones, and retained only its causal uses in our counts.

To count the number of occurrences for each causal connective in each sub-part of the corpus, we first pre-processed the corpora to transform each connective as one word-form (e.g. *étant donné que* became *étantdonnéque*, and *puisqu'* became *puisque.*). Then, we counted each connective, and normalized the figures to obtain a ratio of connectives per 100,000 tokens.

Moreover, when comparing French sub-corpora translated from different source languages, we also computed the rank of each connective in the frequency list extracted from each corpus. Comparing these ranks provided important information about their respective frequencies.

We have found that the frequency of each connective does not vary significantly throughout the corpus (years 1996-1999), which tends to prove that the use of connectives does not depend crucially on the style of a particular speaker or translator.

5 Results

This section presents the results of the CBDF measure for each corpus (Section 5.1), and shows how the frequencies of connectives reveal differences between translated and original texts (Section 5.2) and between texts translated from various source languages (Section 5.3).

5.1 Text Similarity according to CBDF

For Experiment 1, we have compared the differences between original and translated texts, for English and French. The values of CBDF similarity resulting from this comparison are shown in Table 3. Compared to the different scores computed by Kilgarriff, these scores indicate that the two pairs of corpora are both quite similar.

	CBDF
Original English – Translated English	13.28
Original French – Translated French	12.28

Table 3: CBDF between original and translated texts

The similarities between sub-corpora of French translated from different source languages (Experiment 2) are shown in Table 4. The values comparing the same portion (e.g. OF/OF) indicate the homogeneity score of the respective sub-corpus.

	OF	EF	DF	IF	SF
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*

Table 4: Values of CBDF (χ^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.

Looking at the values in Table 4, we can see that the similarity score between OF and EF is 6.00, which, compared to Kilgarriff's values for British corpora, is lower than all but two of the 66 pairs of corpora he compared. Most of the values observed by Kilgarriff are in fact between 20 and 40, and the similarity we found for OF vs. EF is, for instance, in the same range as the one for the journal *The Face* vs. *The Daily Mirror*, a tabloid, and higher than the similarity of two broadsheet newspapers (i.e., they get a lower CBDF value). Therefore, we can conclude that OF and EF are very similar from a word distribution point of view.

As for the other pairs, they are all in the same range of similarity, again much more similar than the corpora cited in Kilgarriff's Table 10. Regarding internal comparisons, OF/EF appears as the second most dissimilar pair, preceded only by IF/EF (French translated from Italian vs. from English). 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 and EF/DF pairs, reflecting the similarity of translations from related languages.

[1] The English causal connective *for* is more difficult to address because of its ambiguity with the homographic preposition. However, on a sample of 500 tokens of *for* randomly extracted from Europarl, we found only two occurrences of the connective *for*, leading us to exclude this connective from our investigation.

Homogeneity values are higher than similarity values (the χ^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 subset used by Kilgarriff, the *Dictionary of National Biography* (1.86) or the *Computergram* (2.20).

Figures on the distribution of connectives, presented in the next section, tend to show that these sub-corpora are however not as similar as they may seem at a first view.

5.2 Text Similarities Measured with the Use of Causal Connectives: Experiment 1

In Experiment 1, we highlight the differences in the use of causal connectives between original English and translated English. Figure 3 shows the discrepancy between the use of the same connectives in original and translated texts. Among these connectives, *since* is the only truly ambiguous word. We have therefore also evaluated the proportion of causal uses of *since* among all the uses of the word *since*. In original English, this proportion is 31.8% and doubles in translated English to reach 67.7%.

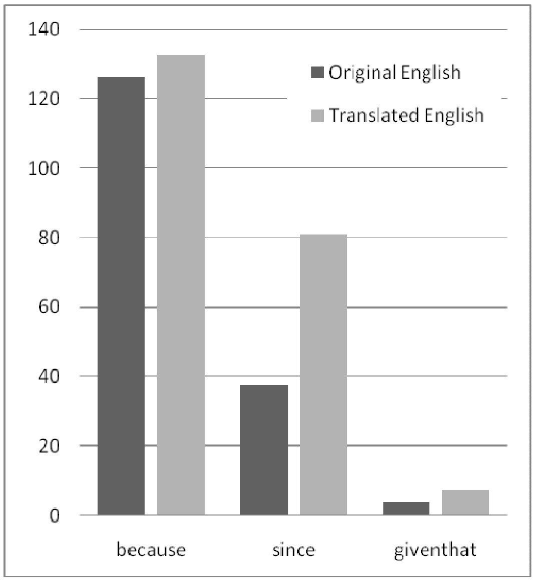

Figure 3: Ratio connectives/100,000 tokens in original and translated English.

These figures show that original and translated texts differ, at least in terms of the number of causal connectives they contain. While *because* seems equally used in original and translated English, *since* and *given that* are used three times more frequently in translated than in original texts. This variability is also noticeable when comparing original and translated uses of French connectives, as shown in Figure 4.

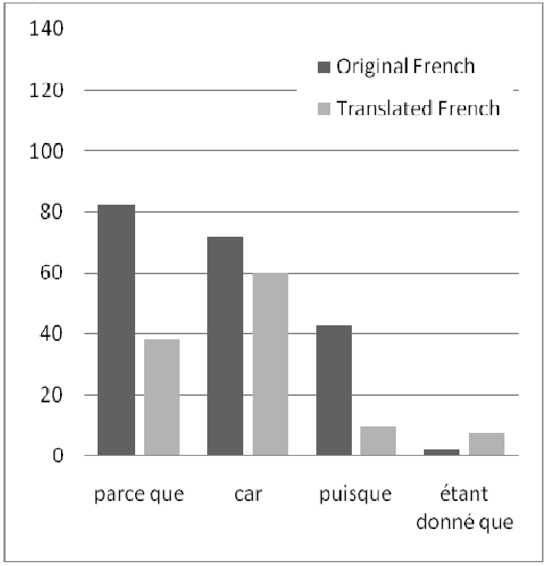

Figure 4: Ratio connectives/100’000 tokens in original and translated French.

For French, while *car* seems to be equally used in both sub-parts of the corpus, *parce que* is used twice less frequently in translated than in original texts. This discrepancy is even bigger in the case of *puisque*, which is used five times less frequently in translated than in original texts. The reverse phenomenon is observed for *étant donné que*, which is used four times more frequently in translated than in original texts.

By looking at the translation of every connective, we were able to count the number of connectives inserted in the target language, that is to say when there was a connective in the target system but no connective in the original text. Conversely, we have also counted the number of connectives removed in the target text, when a connective in the source language was not translated at all. Overall, we found that connectives were inserted much more often than removed during the process of translation. In the case of English as a target language, 65 connectives were inserted while 35 were removed. In the case of French, 46 connectives were inserted while 11 were removed.

5.3 Text similarities measured by the use of causal connectives: Experiment 2

When comparing the number of occurrences of French causal connectives across texts translated from different languages, the differences are striking. Indeed, every source language seems to increase the use of one specific connective in the French translations.

Figure 5 presents the ratio of connectives per 100’000 token. The data compares the use of connectives in French translated from English, Italian, Spanish and German.

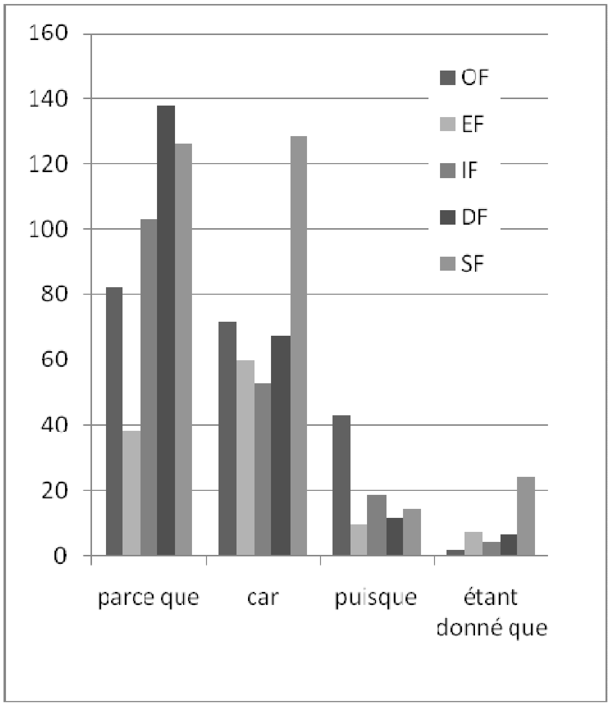

Figure 5: Connectives per 100,000 tokens in French texts translated from various source languages (for each connective, from left to right OF, EF, IF, DF, SF)

Table 5 provides the *rank* of every connective in the word frequency list (sorted by decreasing frequency) computed for each sub-corpus. Grey cells indicate the most frequent connective in each sub-corpus.

	OF	**EF**	**IF**	**DF**	**SF**
parce que	**115**	292	**99**	159	87
car	136	**172**	201	**82**	**85**
puisque	235	1070	601	886	790
étant donné que	3882	1368	2104	1450	459

Table 5: Rank of the connectives in word frequency list for each corpus. Note that the order varies with the source language.

These figures show that the distribution of every connective differs radically according to the source language. Every source language seems to increase the use of one specific connective. When German is the source language, *car* is used twice more often than when English or Italian are the source languages. When Italian is the source language, *parce que* is used twice as often and when English is the source language, *étant donné que* is again used twice as often. Overall, *puisque* is the only connective that does not seem to be enhanced by any of the source languages, which confirms some prior linguistic analyses of this item, showing that *puisque* does not have exact equivalents in other close languages (Degand 2004; Zufferey to appear).

6 Discussion

We have compared the use of discourse connectives in different sub-parts of the Europarl parallel corpus with the use of general vocabulary, as computed by a measure of lexical homogeneity. Our main finding is that even though the lexical measure showed the similarity of these sub-parts, the use of discourse connectives varied tremendously between the various sub-parts of our corpus.

One of the reasons why connectives show more variability than many other lexical items is that they are almost always optional. In other words, as argued in Section 3, for every individual use of a connective, the translator has the option to use another connective in the target language or to leave the coherence relation it conveys implicit. Coherence marking is therefore a global rather than a local textual strategy.

Given that connectives can be used or left out without producing ungrammatical results, studying their variability between comparable corpora provides interesting indications about their global homogeneity. The significant variability that we report between comparable (monolingual) sub-parts of the Europarl corpus indicates that they are not as homogeneous as global lexical measures like the CBDF tend to indicate. In other words, the various sub-parts of the corpus are not equivalents of one another for all purposes, and should not be used as such without caution. These differences were noticeable both by the different number of every connective used in every sub-part of the corpus, but also by the rather different frequency rank that was measured for every one of them in these same sub-parts.

From a translation perspective, our study also provides some further confirmation for the existence of specific characteristics that define translated texts (i.e. “translationese” or “third code”). More specifically, our study corroborates the explicitation hypothesis (Blum-Kulka 1986), positing that translated texts are more explicit than original ones due to an increase of cohesion markers. Connectives are part of the lexical markers that contribute to textual coherence, and we found that they are indeed more numerous in translated than in original texts. For English as a target language, translators have inserted twice as many connectives as they have removed. For French, this proportion raises to four times more insertions than omissions.

However, our data also indicates that the source language has an important influence on the nature of its translation. Indeed, for the use of connectives, we report important variations between texts translated into French from various source languages. More interestingly still, every source language triggered the use of one specific connective over the others. This connective was always specific to one particular source language.

It is also noteworthy that the similarity between texts translated into French, as measured with the CBDF, is greater when the source languages are typologically related. In our corpora of translated French, we found that texts were more similar when comparing the portion translated from Spanish and Italian (Romance languages) and when comparing texts translated from English and German (Germanic languages). This result makes intuitive sense and provides further confirmation of the reliability of this measure to assess global similarity between portions of texts.

7 Conclusion

The Europarl corpus is mostly used in NLP research without taking into account the direction of translation, in other words, without knowing which texts were originally produced in one language and which ones are translations. The experiments reported in this paper show that this status has a crucial influence of the nature of texts and should therefore be considered. Moreover, we have shown that translated texts from different source languages are not homogeneous either, therefore there is no unique translationese, and we identified some characteristics that vary according to the source language.

Our study also indicates that global measures of corpus similarity are not always sensitive enough to detect all forms of lexical variation, notably in the use of discourse connectives. However, the variability observed in the use of these items should not be discarded, both because of their rather frequent use and because they form an important aspect of textual strategies involving cohesion.

Acknowledgments

This study was partially funded by the Swiss National Science Foundation through the COMTIS Sinergia project (www.idiap.ch/comtis). The authors would particularly like to thank Adam Kilgarriff for his explanations regarding the CBDF measure.

References

Altenberg Bengt. 1986. Contrastive linking in spoken and written English. In Tottie G. & Bäcklund U. (Eds.), English in Speech and writing: a symposium. Uppsala, 13-40.

Baker Mona. 1993. In Other Words. A coursebook on translation. Routledge, London/New York.

Baker Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Somers H. (Ed.) Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager. John Benjamins, Amsterdam, 175-186.

Baroni Marco and Bernardini Silvia. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3). 259-274

Degand Liesbeth. 2004. Contrastive analyses, translation and speaker involvement: the case of *puisque* and *aangezien*. In Achard, M. & Kemmer, S. (Eds.), Language, Culture and Mind. The University of Chicago Press, Chicago, 251-270.

Halliday Michael and Hasan Ruqaiya. 1976. *Cohesion in English.* Longman, London

Halverson Sandra. 2004. Connectives as a translation problem. In Kittel, H. et al. (Eds.) An International Encyclopedia of Translation Studies. Walter de Gruyter, Berlin/New York, 562-572.

Ilisei Iustina, Inkpen Diana, Corpas Pastor Gloria and Mitkov Russlan. 2010 Identification of Translationese: A Machine Learning Approach. In Gelbukh, A. (Ed), Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, 503-511

Kilgarriff Adam. 2001. Comparing Corpora. *Intl. Journal of Corpus Linguistics* 6(1): 1-37.

Kilgariff Adam. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Fifth ACL Workshop on Very Large Corpora*, Beijing.

Knott Alistair and Dale Robert. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes* 18(1), 35-62.

Koehn Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation, *MT Summit 2005*.

Lamiroy Beatrice. 1994. Pragmatic connectives and L2 acquisition. The case of French and Dutch. *Pragmatics* 4(2), 183-201.

Laviosa-Braithwaite Sara. 1996. The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation. PhD Thesis, Manchester, UMIST.

Mann William and Thomson Sandra. 1992. Relational Discourse Structure: A Comparison of Approaches to Structuring Text by 'Contrast'. In Hwang S. & Merrifield W. (Eds.), Language in Context: Essays for Robert E. Longacre. SIL, Dallas, 19-45.

Millis Keith & Just Marcel. 1994. The influence of connectives on sentence comprehension. *Journal of Memory and Language* 33 (1): 128-147.

New Boris, Pallier Christophe, Brysbaert Marc, Ferr Ludovic and Holloway Royal. 2004. Lexique~2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36 (3): 516-524.

Noordman Leo and de Blijzer Femke. 2000. On the processing of causal relations. In E. Couper-Kuhlen & B. Kortmann (Eds.) Cause, Condition, Concession, Contrast. Mouton de Gruyter, Berlin. 35-56.

Ozdowska Sylvia. 2009. Données bilingues pour la TAS français-anglais : impact de la langue source et direction de traduction originales sur la qualité de la traduction. Proceedings of Traitement Automatique des Langues Naturelles, TALN'09, Senlis, France.

Sanders Ted. 1997. Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes* 24: 119–147.

Sanders Ted and Sweetser Eve (Eds) 2009. Causal Categories in Discourse and Cognition. Mouton de Gruyter, Berlin.

Tirkkonen-Condit Sonja. 2000. In search of translation universals: non-equivalence or « unique » items in a corpus test. Paper presented at the UMIST/UCL Research Models in Translation Studies Conference, Manchester, UK, April 2000.

Zufferey Sandrine to appear. “Car, parce que, puisque” Revisited. Three empirical studies on French causal connectives. *Journal of Pragmatics*.