Dialogue and Discourse 4(2) (2013) 65-86 doi: 10.5087/dad.2013.204

Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique

Bruno Cartoni BRUNO.CARTONI@UNIGE.CH
Département de Linguistique
Université de Genève
Rue de Candolle 2
CH-1211 Genève 4

Sandrine Zufferey S.I.ZUFFEREY@UU.NL
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht

Thomas Meyer THOMAS.MEYER@IDIAP.CH
Idiap Research Institute
Centre du Parc
Rue Marconi 19
CH-1920 Martigny

Editors: Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

The various meanings of discourse connectives like *while* and *however* are difficult to identify and annotate, even for trained human annotators. This problem is all the more important since connectives are salient textual markers of cohesion and need to be correctly interpreted for many Natural Language Processing applications. In this paper, we suggest an alternative route to reach a reliable annotation of connectives, by making use of the information provided by their translation in large parallel corpora. This method thus replaces the difficult explicit reasoning involved in traditional sense annotation by an empirical clustering of the senses emerging from the translations. We argue that this method has the advantage of providing more reliable reference data than traditional sense annotation.

Keywords: discourse relations, connectives, annotation methods, parallel corpora, translation

1 Introduction

Many natural language processing (NLP) tools rely on annotated data, that is linguistic data enriched with meta-information. For most part, this information requires manual annotation, often performed by more than one human annotator, in order to ensure optimal reliability. This paper reports a set of experiments performed for the annotation of discourse connectives in the context of a project that aims at improving machine translation systems.

One of the main problems for current machine translation systems comes from lexical items that cannot be resolved by looking at individual sentences, such as pronouns, discourse connectives and verbal tenses. The goal of the Swiss COMTIS project[1] is to

[1] http://www.idiap.ch/comtis

 Submitted 02/12; Accepted 11/12; Published online 04/13

extend the current statistical machine translation paradigm by modeling these inter-sentential relations (Popescu-Belis et al. 2011; 2012). This project addresses several types of cohesion markers, but the experiments reported in this paper are limited to discourse connectives. We particularly focus on the challenging task of annotating the meaning of connectives, and advocate the use of a method called translation spotting. This method is based on the collection of a large amount of translations of connectives in a target language in order to capture the different meanings of a given connective in the source language.

The paper is organized as follows. First, we briefly define the category of discourse connectives, emphasizing their importance for textual coherence and discussing the challenges they raise for machine translation (Section 2). We go on to compare in Section 3 two techniques used in the literature to annotate the meaning of connectives, namely sense annotation (3.1) and translation spotting (3.2) and discuss their potential advantages and limitations. In Section 4, we sequentially test these methods through a series of annotation experiments, with the conclusion that translation spotting adds improvements with respect to sense annotation. We go on to show in Section 5 that translation spotting can also be used to identify fine-grained differences between connectives conveying the same meaning (i.e., a causal relation). Section 6 discusses the advantages and limitations of the translation spotting method and Section 7 summarizes our conclusions.

2 Discourse Connectives: a Challenge for Machine Translation

Discourse connectives, such as the words *because* and *while* in English or *parce que* and *mais* in French 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). Even though most languages possess such a set of items, they vary tremendously in the number of connectives they have to express relations and in the use they make of them.

Moreover, a well-known property of discourse connectives is that they are often multifunctional and can convey several coherence relations. In some cases, various relations are conveyed by the same occurrence of a connective. For example, in French, the connective *tant que* (roughly corresponding to the English *as long as*) intrinsically conveys both a temporal relation and a conditional meaning in all its occurrences. In other cases, a connective can potentially convey several relations, but a single occurrence conveys only one of these relations. In such cases, a specific occurrence can be ambiguous between several rhetorical relations. To cite a case in point, the English connective *since* can convey a causal meaning but also a temporal one. In French however, these two meanings require distinct translations: *depuis que* for the temporal meaning and *car* or *puisque* for the causal one. From a machine translation perspective, the main challenge raised by discourse connectives is to be able to assign them a correct meaning in order to translate them appropriately. For example, in order to translate (1) correctly, a system has to recognize that *since* here has a temporal meaning and not a causal one, and should therefore be translated by *depuis que* as in (2) and not by the causal connective *car* as in (3), as was produced by a web-based translation engine.

1. I have been having fun **since** this conference started.
2. J'ai eu beaucoup de plaisir **depuis que** la conférence a commencé.
3. *****J'ai eu plaisir **car** cette conférence a commencé.

In order to disambiguate discourse connectives for machine translation (and more specifically for statistical machine translation (SMT), the COMTIS project proposes to pre-process their occurrences and label them with meaning tags, thus enabling the SMT system to make the correct choice in the target language. In other words, the training data should contain occurrences of *since* labeled as either *causal* or *temporal*, in order to help the SMT system to learn how these two uses of the connective should be translated in different contexts.[2] This labeling of connectives is achieved automatically using machine learning, with algorithms trained on manually annotated reference data (Meyer & Popescu-Belis 2012). Afterwards, the same classifier is applied when translating a new sentence.

In this approach, the automatic disambiguation of connectives thus requires the manual annotation of a large amount of data. In this paper, we discuss the problems raised by this manual annotation. We present the different techniques that have been applied in the COMTIS project in order to achieve reliable and tractable results. First, a classical sense annotation approach has been used, which consists in asking human judges to annotate manually a set of data with several possible senses for each connective. The rather low inter-annotator agreement resulting from this annotation led us to investigate another technique based on translation spotting. These two approaches are described in turn in the next sections.

3 State-of-the-Art Methods for the Annotation of Connectives

This section presents two methods used to annotate discourse connectives: sense annotation (Section 3.1) and translation spotting (Section 3.2). Section 3.3 provides an overview of the resources created using translation spotting.

3.1 Sense Annotation

A classical annotation method for connectives consists in asking several human annotators to assign a label from a list of senses to occurrences of a given connective. Usually, such annotations are performed by more than one annotator, and an evaluation step assesses the reliability of the annotation by measuring the inter-annotator agreement. This assessment is needed in order to ensure that the annotation is valid (Arstein & Poesio 2008). As stated by Spooren and Degand (2010: 253) "ideally coders work completely independently and agree substantially". But in many cases, this goal cannot be met. Spooren and Degand suggest various solutions in order to improve the level of agreement, such as increasing the amount of training for the annotators, or discussing the disagreements between annotators in order to reach a consensus. In a meta-analysis of factors influencing inter-annotator agreement on three different types of linguistic data, Bayerl & Paul (2011) found eight factors with a significant impact on agreement scores, among which were the amount of training, the homogeneity of the group of annotators and number of linguistic categories to be annotated. Even though this meta-analysis did not include linguistic phenomena related to discourse, these factors confirm that Spooren and Degand's suggestions should have a positive impact on inter-annotator agreement.

One of the most important resources containing sense annotation for discourse connectives is the Penn Discourse Treebank (PDTB) (Prasad et al., 2008).[3] The PDTB provides a discourse-layer annotation over the Wall Street Journal Corpus (WSJ) containing the same sections as have already been annotated syntactically in the Penn

[2] The COMTIS project focuses on French and English, but the methodology developed for the disambiguation of connectives can be extended to other languages.

[3] The current version 2.0 is available through the Linguistic Data Consortium at: http://www.ldc.upenn.edu. A website with an extensive bibliography, tools and manuals can be found at: http://www.seas.upenn.edu/~pdtb

Treebank. The discourse annotation consists of manually annotated senses for about 100 types of explicit connectives, implicit discourse relations and their argument spans. For the total size of the WSJ corpus of about 1,000,000 tokens, there are 18,459 annotated instances of explicit connectives and 16,053 instances of annotated implicit discourse relations. The senses that discourse connectives can signal are organized in a hierarchy containing three levels of granularity, with four top level senses (Temporal, Contingency, Comparison and Expansion) followed by 16 subtypes on the second level and the 23 detailed sub-senses on the third level. The annotators of the PDTB were allowed to freely choose senses among all levels, including the possibility to annotate double sense labels (from any hierarchy levels) to account for ambiguous cases. This is why, in principle, 129 sense combinations are possible. A similar methodology has been implemented to annotate discourse relations in many other languages such as Hindi, Czech, Arabic and Italian (see Webber & Joshi 2012 for a review). In addition, Zufferey et al. (2012) conducted multilingual annotation experiments in five Indo-European languages. In all these studies, similar cases of inter-annotator *disagreement* were reported. These results indicate that the methodology and results from the PDTB can be to a large extent replicated in other languages.

Among the 100 different explicit connectives found in the PDTB, we calculated that 29 of them were annotated only with one sense for all their occurrences, covering 412 occurrences. These connectives can therefore be treated as non-ambiguous. Among the remaining 71 connectives, we counted that 52 connectives were annotated with two labels belonging to different top-level categories in the hierarchy. For example, the connective *while* was annotated with the label *concession* (belonging to the comparison class), and with the label *synchrony* (belonging to the temporal class). We reasoned that connectives like *while*, with several senses belonging to different top-levels categories, represented an important ambiguity that needed to be resolved for translation purposes. We therefore concentrated our annotation effort on connectives belonging to this category.

In the PDTB, problems related to inter-annotator agreement have been resolved by choosing the first common label in the hierarchy above the ones that were annotated. For example, when one annotator had labeled an occurrence of *while* as *expectation*, and another annotator had labeled it as *contra-expectation* (both labels come from the most detailed third level of the hierarchy), this disagreement was resolved by going up to the second level of the hierarchy and choosing the tag *concession*, covering the two chosen tags. Detailed information on the performance of the annotators is given in Miltsakaki et al. (2008). The inter-annotator agreement for the four top-level senses in the PDTB is high, at 92%. For the most detailed third level however, performance drops to 77%, showing the difficulty of such a fine-grained annotation.

Performance on specific discourse connectives is only given for the early stages of the PDTB corpus annotation. For example, in Miltsakaki et al. (2005), some information is provided on the annotation of *while* with its four main senses, that were described at the time of that paper as: temporal, concessive, contrast and comparison. For 100 tokens of *while* and two annotators, 20 sentences were judged to be uncertain. Out of the 80 remaining sentences, there was 84% of agreement and 16% of disagreement. When all 100 sentences are taken into account, the overall agreement reaches only 67%.

In short, sense annotation such as the one performed in the PDTB is not always straightforward for the annotators and different annotators do not consistently annotate many fine-grained distinctions.

3.2 Translation Spotting

Translation spotting is an annotation method that makes use of the translation of specific lexical items in order to disambiguate them. For example, an occurrence of *since* translated by *puisque* in French indicates that this occurrence of *since* has a causal rather

than a temporal meaning, because the French connective *puisque* is unambiguous while the English *since* is not. Table 1 presents an excerpt of parallel sentences from Europarl containing *since* in English and the translation spotting, done manually. For one single item in the source language, translation spotting has to be performed over a large set of bilingual sentence pairs, in order to cover many possible correspondences in the target language.

	English Sentence	**French Sentence**	**Transpot**
1	In this regard the technology feasibility review is necessary, **since** the emission control devices to meet the ambitious NOx limits are still under development.	À cet égard, il est nécessaire de mener une étude de faisabilité, **étant donné que** les dispositifs de contrôle des émissions permettant d'atteindre les limites ambitieuses fixées pour les NOx sont toujours en cours de développement.	étant donné que
2	Will we speak with one voice when we go to events in the future **since** we now have our single currency about to be born?	Parlerons-nous d'une seule voix lorsque nous en arriverons aux événements futurs, **puisqu**'à présent notre monnaie unique est sur le point de voir le jour?	puisque
3	In East Timor an estimated one-third of the population has died **since** the Indonesian invasion of 1975.	Au Timor oriental, environ un tiers de la population est décédée **depuis** l'invasion indonésienne de 1975.	depuis
4	It is two years **since** charges were laid.	Cela fait deux ans que les plaintes ont été déposées.	paraphrase

Table 1: Example of translation spotting for *since*

The term *translation spotting* was originally coined by Véronis & Langlais (2000) to designate the automatic extraction of a translation equivalent in a parallel corpus. In our experiments however, the spotting was done manually in order to get fully accurate reference data. Indeed, some attempts have been made to perform translation spotting automatically (Simard, 2003), but they proved to be particularly unreliable when dealing with connectives: Danlos and Roze (2011) assessed the translation spotting performed by TransSearch (Huet et al. 2009), a bilingual English-French concordance tool that automatically retrieves the translation equivalent of a query term in target sentences, and found that for the French connectives *en effet* and *alors que*, the tool spots an appropriate English translation for 62% and 27.5% of the cases respectively. Compared to the general performance of the TransSearch tool for the rest of the lexicon (around 70% of accurate transpots), these results are particularly low. Danlos & Roze (2011) suggest that one possible explanation is the important number of possible translations that can be found for connectives, ranging from no translation to paraphrases and syntactic constructions, which therefore are difficult to spot automatically.

The theoretical idea behind translation spotting is that differences in translation can reveal semantic features of the source language (e.g. Dyvik, 1998; Noël, 2003). In these studies, translation is used to elicit some semantic feature of content words in the source language. Yet, Behrens & Fabricius-Hansen (2003) convincingly showed that using translated data can also help to identify the semantic space of the coherence relation of *elaboration*, conveyed with one single marker in German (*indem*) but translated in various ways in English (*when, as, by + ing, -ing*). Of course, translated texts do not faithfully

reproduce the use of language in source texts as translation has a number of inherent features (e.g. Baker, 1993). Translated data can therefore only be used to shed light on the source language, and investigation should be based on the source language side of parallel data only (see Section 4.1 for details on our corpus data).

When performed manually, translation spotting provides very reliable results and has a number of advantages over sense annotation. First, it relies on the decision made by the translator, who is an expert in his/her own language, and who makes translation choices according to the entire context of use (i.e., knowledge of the whole text) and his/her professional training in the target language. Second, the task is easier to explain to human annotators, and disagreements are rather few. By contrast, the disagreements for some sense tags can be really high for some distinctions such as *concession* and *contrast* (Zufferey et al. 2012). Third, the different labels are not set *a priori*, and the wide variety of translations provides an overview of the possible means to translate a connective. Finally, this task gives an interesting view of the number of discrepancies between the two languages, when there are no one-to-one translation equivalences, a very frequent situation for connectives. This last advantage is less important for annotation but has important implications for other NLP tasks relying on aligned data.

However, translation spotting also has a number of limitations. The most important one is that it provides a direct disambiguation only when the language of translation is less ambiguous than the source language for a given linguistic item, and only one translation is possible for each meaning of the source language. In addition, even in a large corpus, there is no guarantee that all possible senses of a connective will be covered. Another limitation is the necessity to include data from several genres in order to cover a larger range of connective uses, as the functions of connectives are variable across text types (Sanders 1997). In the specific context of the COMTIS project however, the parallel corpus used for translation spotting is the same corpus as the one used to build the language model for machine translation. Consequently, ambiguities that are found in the annotation are precisely those that have to be dealt with for machine translation.

In order to solve part of these limitations, we suggest adding a second step of analysis to translation spotting. This step consists in grouping items of the target language that share the same meaning. For example, in Table 1, the translation spotting of *since* in sentences 1 and 2 are clustered, because both *étant donné que* and *puisque* convey a causal meaning in French, while the two others (*depuis* and the paraphrase *cela fait X que*) convey a temporal meaning. But clustering is not always an easy task for all meaning differences. In order to perform it in the most reliable way, we propose an empirical method involving an interchangeability test. This test is performed by asking human judges to decide which connective can be replaced by another one from the list of possible translations. It takes the form of a sentence completion task. This additional step allows for the separation of translations that are equivalent and reflect the same meaning in the source language and translations that are not equivalent (or interchangeable) and reflect two different meanings of the connective in the source language. For example, a translation spotting performed for the English connective *although* resulted in three main translations in French: *pourtant, bien que* and *même si*. However, an interchangeability test performed on a set of French sentences revealed that *bien que* and *même si* were interchangeable (provided that the mood of the verb is unmarked), as they both reflect a concessive meaning of *although*, while *pourtant* cannot be used in place of the other two connectives, as it reflects the contrastive meaning of *although*. Thus, through this sentence completion task, equivalent translations could be reliably identified and the two meanings of *although* were reliably coded in the source language. Additional examples of such tests are presented in Section 4.4.

3.3 Resources Created in the COMTIS Project

In the COMTIS project, translation spotting was so far performed on seven English connectives, reported in Table 2. In this table, a priori meanings correspond to the possible meanings of connectives identified in reference data and a posteriori meanings correspond to the meaning tags assigned after the clustering phase described above. The number of sentences for the resources created through translation spotting is often lower than the number of sentences that were spotted, due to cases of zero translations or ambiguous connectives, for which no specific meaning can be identified. Translation spotting was made with English-French parallel sentences. Additional spottings of connectives are in progress for other language pairs.

Connective	***A priori* meanings**	***A posteriori* Meanings**	**No. of annotated sentences**	**Resources created (in sentences)**
while	contrast, concession, comparison, temporal	contrast/temporal, concession, contrast, temporal_duration, temporal_punctual, temporal_conditional	499	294
although	contrast, concession	contrast, concession	197	183
though	contrast, concession	contrast, concession	200	155
even though	contrast, concession	contrast, concession	212	191
since	causal, temporal	causal, temporal, temporal/causal	423	423
yet	adverb, concession, contrast	adverb, concession, contrast	509	403
meanwhile	contrast, temporal	contrast, temporal	131	131
Total			**2171**	**1780**

Table 2: Resources created in the COMTIS project through translation spotting

4 Experiments Comparing Sense Annotation and Translation Spotting

We have discussed in Section 3 two possible methods for assigning a meaning to ambiguous connectives. In this section, we will test them through a series of annotation experiments using a convergent methodology and the same annotators in both cases. These experiments will provide a comparative evaluation of their advantages and limitations.

4.1 Data and Methodology

For our experiments we used the Europarl corpus (Koehn, 2005), a multilingual corpus made of the minutes of the debates of the European parliament. This corpus contains 23 languages in parallel: each speaker speaks in his/her own language, and every statement is translated into the other official languages.

The Europarl corpus is a 506-fold parallel corpus (23*23-23), but this does not mean that all parallel data contains an original text and its translation. A statement made in German will be translated both into English and French, and the two resulting texts are therefore two parallel translations. Moreover, the two directions of translation cannot be

considered as equivalents. Previous studies (Degand, 2004; Cartoni et al. 2011) revealed that one of the variation factors for the use of connectives is the status of the text, and more specifically whether it is an original text or a translation. Consequently, the use of parallel data in the study of discourse connectives requires identifying clearly the source and the target languages. The Europarl corpus contains this information in the meta-data structure, but pre-processing steps are required to extract parallel texts, where original and translated languages are clearly identified. These steps are described in Cartoni et al. (2011) and Cartoni & Meyer (2012).

The annotators that did the sense annotation experiments described in Section 4.2 were native French speakers with high proficiency in English. These annotators have been trained in two steps. First, they received written explanations about the discourse relations that they were going to annotate with examples of these relations. After reading instructions, they were asked to annotate a set of 50 sentences. A first evaluation was performed on this annotation, by computing an inter-annotator agreement score, and by looking more precisely at cases where the annotation diverged. In a second phase, the annotators received additional explanations about the discourse relations, focusing on the cases where disagreements were found. In some cases, a think-aloud protocol was also used (Ericsson & Simon 1980), by asking each annotator individually to verbalize the reasoning leading to their final decision while they were annotating a couple of sentences. This provided an efficient correction for the annotators in case an incorrect criterion was used and could be identified.

4.2. A Sense Annotation Experiment in English and French

The annotation of connective senses has been tested on one English connective (*while*) and one French connective (*alors que*) that share the property of conveying a contrastive meaning in part of their occurrences.

According to the LEXCONN database of French connectives (Roze et al. 2010), the connective *alors que* can convey a temporal-background meaning (4) in addition to its contrastive meaning (5).

4. En mai, **alors que** je me trouvais encore à Pau, je suis tombé malade.
   In May, CONNECTIVE *I was still in Pau, I got sick.*
5. J'aime beaucoup Molière, **alors que** Corneille m'ennuie profondément.
   I like Molière very much, CONNECTIVE *Corneille bores me dreadfully.*

According to Miltsakaki et al. (2005), the English connective *while* can signal four different senses.[4] First, *while* can indicate a temporal meaning (TEMP), referring to a duration in time, i.e., the synchronous overlapping of two events, as in example (6). The second sense is a comparison (COMP) with a juxtaposition of two or more alternatives, as in example (7). The third label is concession (CONC), where one argument of the sentence is an expectation, which is then violated or negated by the second argument of the sentence, as in example (8). The fourth sense marks a strong contrast (CONT), for example between two extremes (antonyms) of a gradable scale, as in example (9).

6. That impressed Robert B. Pamplin, Georgia-Pacific's chief executive at the time, whom Mr. Hahn had met **while** fundraising for the institute.
7. Between 1998 and 1999, loyalists assaulted and shot 123 people, **while** republicans assaulted and shot 93 people.

[4] The PDTB in its current version uses slightly different and up to 21 different senses (combinations) for *while*.

8. **While** the pound has attempted to stabilize, currency analysts say it is in critical condition.
9. **While** Georgia-Pacific's stock has outperformed the market in the past two years, Nekoosa has lagged the market in the same period.

For the English and the French connectives, we have asked two human annotators to annotate occurrences with the meaning described above. In French, they annotated 423 sentences containing *alors que*, extracted from the French part of the Europarl corpus. Annotators were asked to decide between two labels: "B" for background or "C" for contrast. Two additional labels were provided: one that could be used to indicate that the annotator could not decide which meaning the connective conveyed ("U") and one serving to annotate strings of characters that did not correspond to the connective *alors que* but to another uses of this string of words, as in (10) from the corpus. Such cases were annotated with "D" for discarded.

10. On verrait **alors que** le fédéralisme européen, qu'on nous propose tout à coup comme la panacée, a constitué, dès ses balbutiements, la cause même du mal que l'on dénonce.
    *We would **then** see **that** European federalism, while is all of a sudden being proposed as a cure-all, has from its earliest days been the very cause of the wrong we are condemning.*

The results of this annotation are reported in the Table 3, a contingency table showing the agreements and disagreements between the two annotators.

		Annotator1				
		B	**C**	**D**	**U**	**Total**
Annotator2	**B**	86	109	0	7	202 (47.8%)
	C	12	181	0	6	199 (47%)
	D	0	0	20	0	20 (4.7%)
	U	0	2	0	0	2 (0.5%)
	Total	98 (23.2%)	292 (69%)	20 (4.7%)	13 (3.1%)	423 (100%)

Table 3: Contingency table for the annotation of *alors que*

The agreement of the two annotators on this task was calculated with Cohen Kappa's score (Carletta 1986) and reached 0.428. This represents 67.8% of cases of observed agreement. When looking more closely at the results, we noticed that there was no disagreement on the simplest category D (discard) that was correctly annotated in all 20 occurrences, thus confirming that the two annotators were reliable. They almost never used the label "U", which means that they were rather confident about their choices. Moreover, the cases of disagreements between B and C seem to indicate that the two annotators did not adopt the same strategy in case of uncertainty. There were, for example, an important number of cases (109), where the first annotator consistently chose the contrastive meaning, while the second annotator chose the background meaning, but not the other way round (12 cases only). In other words, ambiguous cases were consistently classified with B by one annotator and C by the other. We will argue in Section 6 that such occurrences may correspond to natural ambiguities, for which a double label tag should be assigned.

In English, 300 sentences containing *while* were extracted from the English part of Europarl and annotated by the same annotators. Guidelines taken from the PDTB

annotation manual (The PDTB Research Group, 2007) were provided to explain the different meanings conveyed by *while*. Annotators had to decide between these four labels, plus one label if they could not decide ("U"). The inter-annotator agreement (Cohen's Kappa score) was 0.426, a rather similar value to the one obtained for *alors que* described above. This corresponds to an agreement for 61.3% of the sentences, a slightly lower value than the 67% obtained by Miltsakaki et al. (2005). The contingency table for *while* is presented in Table 4.

		Annotator 1					
		COMP	CONC	CONT	TEMP	U	Total
Annotator 2	COMP	**13**	1	2	2	0	18 (6%)
	CONC	15	**101**	1	21	1	139 (46.3%)
	CONT	8	22	**5**	8	1	44 (14.7%)
	TEMP	9	9	6	**64**	5	93 (31%)
	U	0	2	1	2	**1**	6 (2%)
	Total	45 (15%)	135 (45%)	15 (5%)	97 (32.3%)	8 (2.7%)	300 (100%)

Table 4: Contingency table for the annotation of *while*

The distribution of annotations reported in Table 4 is rather unbalanced. Annotators seem to reach some agreement for *concession* and *temporal* senses but overall the four labels are mixed, and no particular preference is observed for alternative tags. Contrary to *alors que* (see Table 3 above), for which one annotator clearly tended to choose a different strategy than the other, no emergence of a consistent strategy is found in this case. The larger range of possible meanings probably caused this important number of divergences.

In sum, these annotation experiments highlighted the difficulties of labeling the meanings of discourse connectives, even when only a binary distinction was necessary. In both cases, the inter-annotator agreement remained low, with a Kappa score never reaching 0.5. In the domain of computational linguistics, the threshold of acceptable agreement is highly debated (Arstein & Poesio 2008), but following Krippendorff's scale assessing inter-annotator agreement (Carletta 1996: 52), these Kappa scores do not indicate reliable coding. Following the scale by Landis & Koch (1977), a value of 0.4 is considered to reflect a moderate agreement. In all cases, this score does not appear to be reliable enough to provide reference data for training automated classifiers, as it is aimed in the COMTIS project.

4.3. A Translation Spotting Experiment with the Connective *While*

As mentioned above, the connective *while* can convey four major meanings: temporal, concessive, contrastive and comparative. As we have seen with the sense annotation experiment, the distinction between these four meanings is hard to make in a systematic and reliable way for human annotators. We therefore tried to separate these senses in the source language through translation spotting.

We used 508 bi-sentences extracted from the Europarl corpus for the English-French pair, and we extracted sentences that were originally produced in English. Two human annotators (the same annotators who did the sense annotations) were then asked to identify the connective that was used in the target French text in order to translate *while*. If it was not translated by a French connective, they were allowed to assign different tags for the use of a present participle, a paraphrase, or no translation at all. The table below provides details about the different means used to translate *while* in French.

	No.	%		No.	%
alors que	91	18.24%	mais	4	0.80%
gerund	85	17.03%	malgré	3	0.60%
paraphrases	72	14.43%	quoique	3	0.60%
si	54	10.82%	pendant que	2	0.40%
zero translation	41	8.22%	alors même que	1	0.20%
tandis que	39	7.82%	aussi	1	0.20%
même si	33	6.61%	avant que	1	0.20%
bien que	26	5.21%	contre	1	0.20%
s'il est vrai que	14	2.81%	en même temps que	1	0.20%
tant que	10	2.00%	étant donné que	1	0.20%
pendant	5	1.00%	quand	1	0.20%
puisque	5	1.00%	s'il est exact que	1	0.20%
lorsque	4	0.80%	**Total**	**499**	**100%**

Table 5: Translation equivalents of *while* found in the corpus

Although the task might seem trivial, the two annotators provided a different translation spotting for 150 sentences out of the 508.[5] Most of these cases were due to a disagreement about what counted as a paraphrase. For example, one annotator treated the string of words *s'il est vrai que* as a paraphrase and the other as a connective. This disagreement is easily correctible, and further training has consistently increased the level of agreement. In subsequent tasks, the annotators agreed in 91.5% of the cases when transpotting other connectives like *whereas*, and in 93% of the cases for *although*.

4.4. Interchangeability Tests as a Second Step for Translation Spotting

As can be seen in Table %, a wide range of French connectives is used to translate *while*, reflecting the numerous meanings that this connective can convey. In order to deduce its meanings based on the translations, an additional task of clustering is needed, which involves analyzing the French connectives used in the translations. In order to do so, we performed an interchangeability test on French connectives, taking the form of a sentence completion task. Such a task consists of taking a bunch of sentences from our parallel data containing a specific connective (the connective used in the translation), erase it and ask human annotators to decide, from a list of connectives, which one would fit, without paying attention to the verb mood, which may be influenced by the connective. This kind of test allows making a decision with no theoretical *a priori*. The only *a priori* decision that we made was to separate the translations from Table 5 into two sub-groups: the temporal connectives on one side and all the others on the other side.

Among the 6 most frequent French connectives used to translate *while* (*alors que, si, tandis que, même si, bien que, s'il est vrai que*), we proposed a set of sentences with blanks to fill in to three annotators. For each of the sentences (numbered 1 to 24), Table 6 provides the connectives that were used in the text, followed by the connectives chosen by the annotators (the numbers in brackets correspond to the number of times the connectives have been chosen). Only connectives that were chosen several times are reported.

[5] Among the 508 occurrences of *while*, 499 were connectives. The other occurrences were nouns as in "for a while" or "a while ago", and have been excluded from the count.

Sentence	Connective used in translation	Chosen connectives (number of times / 3 annotators)
1	alors que	alors que (3), si (3), s'il est vrai que (3), tandis que (2)
2	alors que	alors que (3)
3	alors que	alors que (3), tandis que (3)
4	alors que	même si (3), bien que (2)
5	bien que	bien que (3), même si (2)
6	bien que	bien que (3), même si (2), s'il est vrai que (2)
7	bien que	bien que (3), même si (2)
8	bien que	bien que (2), même si (3), si(2), s'il est vrai que (2)
9	même si	même si (3), bien que (3), si(2), s'il est vrai que (2)
10	même si	même si (3), bien que (3), s'il est vrai que (3), si(2)
11	même si	même si (3), bien que (2)
12	même si	même si (3), bien que (3)
13	si	s'il est vrai que (3), même si (3), si(2), bien que (2)
14	si	s'il est vrai que (3), si(3), même si (3), bien que (2)
15	si	s'il est vrai que (3), si(3), même si (2)
16	si	s'il est vrai que (3), même si (3), si(2), bien que (2)
17	s'il est vrai que	s'il est vrai que (3), même si (3), si(2), bien que (2)
18	s'il est vrai que	s'il est vrai que (3), même si (2), bien que (2)
19	s'il est vrai que	s'il est vrai que (3), même si (3), bien que (3)
20	s'il est vrai que	s'il est vrai que (3), même si (3), si(2), bien que (2)
21	tandis que	alors que (3), tandis que (2), si (3)
22	tandis que	alors que (3), tandis que (3)
23	tandis que	alors que (3), tandis que (3)
24	tandis que	alors que (3), tandis que (3)

Table 6: Interchangeability test for non-temporal uses of *while*

Through this test, two clusters of connectives are clearly emerging: one with a concessive meaning containing *même si*, *bien que*, *si* and *s'il est vrai que*, and another one with a contrastive meaning containing *alors que* and *tandis que*. However, this also shows that *alors que* can also have a concessive meaning, as in sentence 4, where it's been interchanged in majority with *même si* and *bien que*. Within these two clusters, there seems to be some more subtle clusters between *même si* et *bien que* on one side, and *si* and *s'il est vrai que* on the other side. This is confirmed in the descriptive reference work LEXCONN (Roze et al. 2010) that assigns the connective *si* both a *concessive* and a *condition* meaning. This latter meaning was never annotated in the English reference for *while* (the PDTB), but will also emerge from the interchangebility test described below. Finally, the meaning of *comparison* was not found in this test. It also shows that the connectives used in the translation were always the first choice of the annotators as well, with the noticeable exception of *tandis que* that the annotators seem to avoid using.

The same test was also performed for the French connectives conveying a temporal meaning *pendant que, tant que, lorsque*. Results are reported in Table 7.

Sentence	Connective used in the translation	Chosen connectives (number of times / 3 annotators)
1	lorsque	lorsque (3)
2	lorsque	lorsque (3)
3	lorsque	lorsque (3), pendant que (2)
4	lorsque	pendant que (3)
5	pendant que	pendant que (3)
6	pendant que	pendant que (3)
7	tant que	tant que (3)
8	tant que	tant que (3)
9	tant que	tant que (3)
10	tant que	tant que (3)

Table 7: Interchangeability test for temporal uses of *while*

This test, contrary to the one above for concessive/contrastive meanings, shows no cluster with more than one connective. Apart from a few exceptions, it seems to show that there are three connectives with a specific meaning that cannot be expressed by another connective. For example, the connective *tant que*, that can roughly be translated into English by *as long as*, indicates duration in time as well as condition: the duration lasts only while the event mentioned in the segment following the connective unfolds. The connective *pendant que* conveys both a notion of contrast and simultaneity with another event. This connective indicates that a contrastive and temporal meaning can coexist in some connectives, with the consequence that some uses of *while* could be tagged as both temporal and contrastive. Finally, *lorsque* only indicates temporal simultaneity.

The interchangeability tests allow the clustering of French connectives that convey the same meaning, and consequently narrow the different possible meanings of English *while*. The translation spotting and interchangeability tests also revealed that there were more fine-grained features to the temporal uses of *while* (simultaneity, condition, etc.). These specificities of *while* with a temporal meaning are more specific than the labels used in the PDTB, where the temporal category is only sub-divided into *synchronous* and *asynchronous*. In this particular case, the translation reveals fine-grained distinctions of meaning in the source language, as it was the case in studies focusing on content words, mentioned in Section 3.2.

Table 8 summarizes the different meanings that have been highlighted by clustering French connectives. Only French connectives that were used more than once have been included in the analysis.

Meaning	%	French connectives
concession	25.45	si (54), même si (33), bien que (26), s'il est vrai que (14)
contrast	7.89	tandis que (39)
contrast/temporal	18.24	alors que (91)
temporal/condition	2	tant que (10)
temporal/comparison	1.4	pendant que (7)
temporal/simultaneity	0.8	lorsque (4)

Table 8: Meanings of *while* emerging from translation spotting

These meanings are then reported on the corresponding occurrence of English *while*, that receives the labels inferred from the translation. This annotated data (294 occurrences of *while* in total) is then used to train classifiers based on machine learning algorithms, in order to automatize the annotation procedure (Meyer and Popescu-Belis, 2012). From the 294 instances, 14 are kept as a held-out test, while the other 280 are used for training a

Maximum Entropy classifier, using the Stanford NLP package (Manning and Klein, 2003). In both, the training and the test sets, features from syntactical parsing (Charniak and Johnson, 2005) are extracted: POS tags and syntactical ancestor categories for the connective, the surrounding words and words at the beginning and end of the clauses. Further features are gained in form of punctuation patterns, antonyms from WordNet and temporal ordering of events obtained from a TimeML parser (Verhagen et al., 2005). Using these features, the 6 listed senses (see *a posteriori meanings* in Table 2) for the connective *while* can be disambiguated, in the held-out set, with an accuracy of about 65%, meaning that the classifier predicts the correct sense in two thirds of all cases. Meyer and Popescu-Belis (2012) have also shown that such a classifier can be used to automatically label the large training data for machine translation. As a consequence, such an SMT system translates discourse connectives more correctly. They further validate the method by automatically classifying up to 12 other temporal-contrastive connectives with larger training sets and by integrating these classifiers into SMT as well.

These experiments show that investigations based on translation spotting over large parallel data can uncover unexpected meanings of the connectives used in the source language. As explained in the next section, this technique can also be used to uncover more fine-grained differences of usages within a single rhetorical relation.

4.5. Comparison and Evaluation

In this section, we systematically compare the translation spotting technique with sense annotation in terms of the sense tags they provide. For the French connective *alors que*, we have compared the sense annotation resulting from translation spotting and clustering with the labels assigned directly by annotators in the sense annotation. This enabled us to check whether the results of the two techniques provided consistent results or not.

As a first comparison, we only used the 267 occurrences for which the two annotators had agreed on the label (background or contrast), and compared this label with the English connectives used to translate *alors que*. Results are presented in Table 9 (only connectives appearing with a frequency of >5% are reported).

Background label		
Transpot	No.	%
when	24	27.91%
while	10	11.63%
at a time when	9	10.47%
as	7	8.14%
zero translation	7	8.14%
whilst	6	6.98%
although	5	5.81%

Contrast label		
Transpot	No.	%
whereas	50	27.62%
when	28	15.47%
while	26	14.36%
although	19	10.50%
zero translation	13	7.18%
whilst	11	6.08%

Table 9: Translation equivalents according to the meaning of *alors que*

When the two annotators agreed on a *background* meaning for *alors que*, a majority of connectives chosen by the translator also have a background meaning (like *when, at a time when*). In the second half of the table, among the occurrences of *alors que* that were labeled as *contrast* by the two annotators, the main connective used can only have a contrastive meaning (*whereas*) while all the other connectives used in translation are ambiguous and can have several labels, amongst which a contrastive meaning is always found in reference data (such as *while*).

In addition, when looking at the 134 occurrences where the annotators disagreed, we notice that 60 of them were translated by unambiguous connectives in English: 51 *alors que* are translated by a clearly contrastive English connective (such as *although, whereas, but…*) and 9 occurrences are translated with clearly temporal English connective (*at the time when, now that*). This confirms that translation spotting can provide disambiguation when annotators cannot. The remaining 74 occurrences are translated by ambiguous connectives in English (*when, while, whilst*). In those cases, the ambiguity is kept in translation.

In sum, this comparison shows that the results from translation spotting are often similar to the sense labels assigned by annotators and can also provide results for an important number of cases of which annotators do not reach agreement. In addition, this technique has the advantage of providing a better way to deal with ambiguity than sense annotation. In many cases, ambiguity is revealed in translation spotting by the choice of a target language connective that can also have the same multiple meanings, as it is the case for the pair of *while* and *alors que*. In consequence, ambiguity can naturally be preserved and dealt with in such cases. On the other hand, while annotating the senses of a connective from a monolingual perspective, our experiments have shown that annotators often feel compelled to choose between various possible meanings. This can lead to arbitrary choices between two values that can in fact coexist naturally. This problem was accounted for in the PDTB by allowing any combination of labels from the sense hierarchy in order to annotate double sense tags to certain occurrences of discourse connectives. However, this technique does not ensure that annotators will identify all the meaning components of a connective, and use several tags instead of one.

5 Translation Spotting for the Identification of Sub-Senses of Connectives

Until now, we have shown that connectives can often convey more than one rhetorical relation and argued that disambiguating these different meanings in context represented a difficult task of manual annotation. In this section, we will concentrate on a different fact: most rhetorical relations can be conveyed in many languages by a whole array of different connectives. For example, a causal relation can be conveyed in French by *parce que*, *car*, *puisque*, *étant donné que, comme, vu que*, etc. (for recent surveys of cross-linguistic comparisons involving causality, see Sanders & Stukker, 2012; Sanders & Sweetser, 2009). The point is that all these connectives are not always interchangeable and therefore cannot be treated as equivalents. Zufferey (2012), for example, showed through a sentence completion task and an acceptability judgment task that the connectives *puisque* and *car* were almost never interchangeable, contrary to what previous theoretical studies had concluded (e.g. Lambda-l Group 1975, Roulet et al. 1985). The main consequence of this finding for machine translation is that assigning a *cause* label to a connective does not ensure that a correct translation will be achieved, since all connectives conveying a causal meaning are not interchangeable. In a nutshell, this observation means that at least in some cases, a more fine-grained annotation scheme than simple rhetorical relations such as *cause*, *concession*, *temporal*, etc. is needed to ensure an optimal translation of connectives. In the PDTB, *cause* is not the most fine-grained level, but its main subdivision between *reason* and *result* serve to separate connectives like *because* and all the French connectives listed above, that have a consequence-cause order of the segments, from connectives like *so* that have a reversed order (cause-consequence).

In this section, we will limit ourselves to giving a flavor of the kind of information that is needed in order to translate causal connectives accurately (see Zufferey & Cartoni 2012, for a detailed presentation of these criteria). Our aim is to show that translation spotting is also a very relevant annotation technique at this finer level of granularity.

One of the main criteria dividing the category of causal connectives is the subjective or objective nature of the causal relation described. In some cases like (11), the causal

relation relates events in the world and is therefore objective, while in other cases like (12) the causal relation involves the speaker's own reasoning or speech act and is therefore more subjective (e.g. Sanders, 1997; Degand & Pander Maat 2003).

11. The snow is melting, because the temperature is rising.
12. John was tired, because he fell asleep.

In English, this difference is not visible in terms of connectives, as *because* can convey both objective and subjective relations (Sweetser 1990). However, in many other languages like Dutch (Pit 2007), German (Sanders & Stukker 2012) and French (Zufferey 2012; Degand & Fagard 2012), different connectives are used to express both kinds of relations. For example, in written French, objective uses are prototypically translated by *parce que* while subjective uses are translated by *car*. This means that in order to translate occurrences of *because* accurately in a number of languages, the degree of subjectivity of the causal relation has to be taken into account. In this case, translation spotting provides an immediate solution for the annotation of occurrences of *because*, in order to provide training data for machine learning algorithms. The translation choices indeed provide this information, as can be seen in Table 10, which presents the translation spotting of 196 parallel sentences containing *because.*

	No.	**%**		**No.**	**%**
car	76	38.78%	vu que	1	0.51%
parce que	63	32.14%	dès lors que	1	0.51%
paraphrases	27	13.78%	*gerund*	1	0.51%
zero translation	8	4.08%	:	1	0.51%
dans la mesure où	6	3.06%	en effet	1	0.51%
puisque	3	1.53%	sans quoi	1	0.51%
en effet	3	1.53%	compte tenu que	1	0.51%
étant donné que	1	0.51%	du fait que	1	0.51%
à défaut	1	0.51%	**Total**	**196**	

Table 10: Translation spotting of the English connective *because*

The two main translations of *because* in French are *car* and *parce que*. It can be assumed that the translations by *car* correspond to the subjective uses of *because* while the translations by *parce que* correspond to its objective uses. In order to verify this claim, we asked two experts to annotate 100 sentences containing the connective *because* with the objective/subjective trait. Results indicate that 90% of the *because* sentences translated by *car* were annotated as subjective. Similarly, 85% of the *because* sentences that were annotated as objective by the annotators were translated by *parce que* rather than *car*.[6]

In sum, this example shows that translation spotting can also be used for very fine-grained distinctions, as long as they are visible in the translations. This comparison also confirms that the information provided by the translations coincides with sense annotation made by experts and is therefore reliable, as discussed in Section 4.5.

[6] In contemporary spoken French, *parce que* is the only connective used for both kinds of relations and in writing, *parce que* can also convey subjective relations in some cases.

6 Discussion

The various annotation tasks presented in this paper confirm that the meanings of discourse connectives are difficult to annotate for human judges. Arguably, this difficulty is at least partially related to the taxonomy of discourse relations that the annotators are instructed to apply. Some fine-grained distinctions are indeed difficult to annotate reliably; for example it is only at the top level of their taxonomy (containing only four generic classes) that the PDTB annotators reached a reliable, even though not perfect, agreement level (92%) (Miltsakaki et al. 2009). However, this kind of general annotation is not precise enough for many applications, including those involving a form of cross-linguistic mapping.

Another problem related to this type of annotation is that there is no consensus in the literature about what an optimal taxonomy of discourse relations should consist of (see e.g. Hovy 1990 for a discussion of this problem). The ideal granularity of the taxonomy is probably not universal but strongly depends on the goal of the annotation. In the case of the COMTIS project underlying this study, the annotation of discourse connectives served the goal of pre-processing for machine translation systems, enabling a disambiguation of the meaning of connectives, leading to an accurate translation choice. As we have shown in this paper, for this purpose a fine-grained taxonomy is required, in order to capture the sometimes subtle differences of meanings between connectives. As our experiments on *alors que* and *while* have demonstrated, this fine-grained annotation is not reliably achieved by human annotators, even when a careful and time-consuming training procedure has been implemented. This led us to consider an alternative route to sense annotation, making use of the information provided by the translation and the intuitive knowledge that native speakers have about the possibility to use a connective in a given sentence (cf. the sentence completion tasks that are part of the second step of our method).

From a theoretical perspective, there seems to be a justification of the acute difficulty of annotating connectives, compared to other lexical items. Many studies on discourse connectives have argued that these lexical items encode procedural rather than conceptual information (e.g. Blakemore, 2002; Moeschler, 2002; Wilson, 2011). In other words, their role in the sentence is to instruct the addressee about the way some of the arguments are related. For example, the connective *therefore* instructs the hearer to look for a consequence between the segment preceding the connective and the one following it. This property of discourse connectives can at least partially explain why their meanings are often difficult to pin down by human annotators. Indeed, procedural meaning is not as easily accessible to conscious introspection as conceptual information (Blakemore, 2002). However, speakers have a very reliable ability to intuitively judge the acceptability in a given context. Just like it is the case for syntax, this intuitive ability is dependent on the language faculty and is not accompanied by a form of declarative knowledge. This difference explains why the task of sense annotation is often difficult for annotators while the sentence completion tasks involved in the translation spotting technique are rather straightforward. Thus, the translation spotting technique avoids one of the main problems related to discourse connectives: the difficulty to reason explicitly about their meaning in context. This task is replaced by several more manageable ones for annotators: identifying a translation and, in the second phase of clustering, using a set of connectives to fill in blanks in sentences. The clustering of senses inferred from these interchangeability tests provides a more reliable indication on the meaning of connectives than the application of a pre-defined set of tags indicating coherence relations, which are often difficult to define and identify. Moreover, the clustering of senses is also more flexible, as tags are defined according to the meaning of connectives in translation, rather than beforehand. Finally, because the annotation tasks involved in translation spotting are rather easy, this technique provides an interesting way to gather rapidly an important amount of data.

This paper has also shown that a cross-linguistic perspective provides some new insights on the possible meanings of connectives in a given language. For instance, the translation of *while* by *tant que* in French indicated that this connective could establish a *condition* meaning. This tag was however not assigned to *while* in the PDTB. Moreover, we saw in Section 5 that looking at translations could also be used to investigate some very fine-grained properties of connectives conveying the same rhetorical relation (i.e., causality). All these observations confirm that looking at a language through the mirror of another language can bring new insights on the meaning of these lexical items, even from a monolingual perspective.

The translation spotting method also has some obvious limitations. First and foremost, it relies on the choices made by the translator. Even with professional translators as the ones involved in our corpus, the translation choice for one particular occurrence of a connective is the result of a specific interpretation and incorrect translations, or at least translations involving meaning shift, cannot be excluded. However, we argue that the important amount of parallel sentences investigated should flatten this bias. Consequently, translation spotting can be expected to be a reliable method only when applied over a large amount of data. This requirement is another limitation of this method.

Another potential problem comes from the fact that it is dependent on the presence of multiple translations in the target language. Indeed, a connective could have many theoretical senses in one language but all these senses could be covered by one single connective in the target language. Whether this limitation is a problem or not depends on the expected generalization of the annotation. If the aim of the annotation is to provide an accurate translation in a given target language, this ambiguity can be carried over without producing translation errors. However, this technique will not provide indications on the different meanings of this connective that could be reused for a different target language.

Moreover, when an ambiguity is repeatedly preserved across languages, the status of this ambiguity should be questioned. For example, it is possible that sometimes *background* and *contrast* are two values of a connective that are denoted at the same time in a given occurrence, just like some other connectives require several labels to account for their meaning. The fact that a connective covering these two meanings is also used in the translation (as in the example of the pair made of *alors que* and *while*) might mean that the value "background-contrast" can be treated as a single unit, or a somehow underspecified value. In other words, the possibility that connectives can sometimes convey two compatible but different rhetorical relations in a single occurrence has to be taken into account, as it is the case in the PDTB where annotators are allowed to use double tags for single connective occurrences. Another example of such a double meaning can be observed in some occurrences of *since*, where a temporal and a causal meaning both seem to be conveyed simultaneously. Further confirmation for the existence of such double sense labels can be obtained from experiments with automated sense classifiers and machine learning. Before training the classifiers, the cases where human annotators disagreed can be resolved by assigning double labels, for instance, when one annotator used a *temporal* sense for an occurrence of *since*, and the other annotated a *causal* sense, this disagreement can be resolved by assigning a label *temporal-causal* (similarly, *background-contrast* for the French connective *alors que*). For *since*, an automated classifier using three labels (*temporal, causal* and *temporal-causal*) almost reaches the same performance as one that uses *temporal* and *causal* only. For *alors que* a three-way classifier (including *background-contrast*) even reaches higher performance than the two-way one – which is quite surprising, as usually, more classes means more difficulties for automated tools to disambiguate them (Meyer et al. 2011). This might provide further evidence for the existence and usefulness of double sense labels for discourse connectives.

7 Conclusion

In this paper, we demonstrated through several annotation experiments that annotating the senses of discourse connectives is a difficult task for which human annotators do not reach a truly reliable agreement. We proposed the use of an alternative technique to perform this annotation, making use of the clues provided by the translation of the connective in a target language. When the target language does not provide a direct disambiguation, all translations are clustered into different senses based on the possibility to replace the various connectives in the target language. The clusters are formed based on native speakers' judgments about the possibility to use connectives interchangeably in a sentence. This technique therefore provides a more reliable way than traditional sense annotation to label connectives with their meaning in context.

This technique also opens new avenues for further cross-linguistic research on discourse relations and connectives. The approach proposed in this paper offers an interesting and easy way to gather contrastive data that can be extended to larger-scale contrastive analyses. As demonstrated in the case of *while* and the category of causal connectives, the systematic comparison of a large amount of correspondences in translated corpora can provide a complete picture of the equivalences between languages, and provide useful indications about the granularity of discourse relations that are required to describe them cross-linguistically. If extended to a larger set of languages and connectives in a variety of genres, this method would allow for more empirically grounded generalizations about discourse relations in the world's languages. In particular, the fact that one particular occurrence can convey two discourse relations simultaneously, and that this double meaning is repeatedly found in other languages might reflect some general tendencies about the cognitive similarity of some discourse relations.

Acknowledgements

This study was funded by the Swiss National Science Foundation through the COMTIS Sinergia project (www.idiap.ch/comtis). The authors would like to thank the annotators for their careful and meticulous work.

References

Ron Artstein and Massimo Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555-596.

Mona Baker (1993). Corpus linguistics and translation studies: Implications and applications. In M.Baker et al. (Eds), Text and Technology: In honor of John Sinclair. John Benjamins, Amsterdam/Philadelphia.

Petra Bayerl and Karsten Paul (2011). What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics* 37(4):699-725.

Bergljot Behrens and Cathrine Fabricius-Hansen (2003). Translation equivalents as empirical data for semantic/pragmatic theory. In Jaszczolt K, Turner Jen (editors), *Meaning through Language Contrast.* Amsterdam: Benjamins. 463-477.

Diane Blakemore (2002) *Meaning and relevance: the semantics and pragmatics of discourse markers.* Cambridge University Press, Cambridge, USA.

Jean Carletta (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Bruno Cartoni and Thomas Meyer (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of LREC 2012*, pages 2132-2137, Istanbul, Turkey.

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*, pages 78-86, Portland, USA.

Eugene Charniak and Mark Johnson (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL) pages 173–180. Ann Arbor, MI.

Laurance Danlos and Charlotte Roze (2011). Traduction (automatique) des connecteurs de discours. In *Proceedings of TALN 2011*, Montpellier, France.

Liesbeth Degand (2004). Contrastive analyses, translation, and speaker involvement: The case of puisque and aangezien. In M. Achard and S. Kemmer (Eds.), *Language, culture and mind*, pages 1-20, Stanford: CSLI Publications.

Liesbeth Degand and Benjamin Fagard (2012). Competing connectives in the causal domain: French *car* and *parce que*. *Journal of Pragmatics* 44(2): 154-168.

Liesbeth Degand and Henk Pander Maat (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In A. Verhagen and J. Maarten van de Weijer (editors), *Usage-based approaches to Dutch*, pages 175-199, LOT, Utrecht.

Helge Dyvik (1998). A translational basis for semantics. In Johansson, Stig & Signe Okselfjell (eds) *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, pages 51-86, Amsterdam: Rodopi.

K. Anders Ericsson and Herbert Simon (1980) Verbal reports as data. *Psychological Review* 87(3):215-251.

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

Ed Hovy (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*. Pittsburgh, Pennsylvania.

Stéphane Huet, Julien Bourdaillet and Philippe Langlais (2009). Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. In *Proceedings of TALN'09*, Senlis, France.

Philipp Koehn (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, September 13-15, pages 79-86, Phukhet, Thailand.

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

Lambda-l, Groupe (1975). Car, parce que, puisque. *Revue Romane* 10:248-280.

Richard J. Landis and Gary G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

William Mann and Sandra Thomson (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, pages 19-45, Dallas, USA.

Christopher Manning and Dan Klein (2003). Optimization, MaxEnt Models, and Conditional Estimation without Magic. Tutorial at HLT-NAACL and 41st ACL conferences. Edmonton, Canada and Sapporo, Japan.

Thomas Meyer and Andrei Popescu-Belis (2012). Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL 2012 Workshop on Hybrid Approaches to Machine Translation (HyTra)*, Avignon, France, pp. 129-138.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey and Bruno Cartoni (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine

Translation. In *Proceedings of 12th SIGdial Meeting on Discourse and Dialog*, pages 194-203, Portland, USA.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the TLT 2005 (4th Workshop on Treebanks and Linguistic Theories)*, Barcelona, Spain.

Eleni Miltsakaki, Livio Robaldo, Alan Lee and Aravind Joshi (2008). Sense Annotation in the Penn Discourse Treebank. In Alexander Gelbukh (editor), *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*, pages 275-286, Springer Berlin / Heidelberg.

Jacques Moeschler (2002). Connecteurs, encodage conceptuel et encodage procédural. *Cahiers de linguistique française* 24:265-292.

Dick Noël (2003). Translations as evidence for semantics: An illustration. *Linguistics* 41(4):757-785.

The PDTB Research Group (2007). The Penn Discourse Treebank 2.0 Annotation Manual. IRCS Technical Reports Series, 99p.

Mirna Pit (2007). Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast*, *7*(1), 53–82.

Andrei Popescu-Belis, Bruno Cartoni, Andrea Gesmundo, James Henderson, Cristina Hulea, Paola Merlo, Thomas Meyer, Jacques Moeschler and Sandrine Zufferey (2011). Improving MT coherence through text-level processing of input texts: the COMTIS project. In *Proceedings of Tralogy 2011 (Translation Careers and Technologies: Convergence Points for the Future)*, Paris, France.

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*, pages 2716-2720, Istanbul, Turkey.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber (2008). The Penn Discourse Treebank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Morocco.

Eddy Roulet, Antoine Auchlin, Jacques Moeschler, Christian Rubattel and Marianne Schelling, (1985). *L'articulation du discours en français contemporain*. Peter Lang, Berne, Switzerland.

Charlotte Roze, Laurance Danlos and Philippe Muller (2010). LEXCONN: a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

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

Ted Sanders, Wilbert Spooren and Leo Noordman (1992). Towards a taxonomy of coherence relations. *Discourse Processes* 15, 1-36.

Ted Sanders and Ninke Stukker, (2012). Causal connectives in discourse: a cross-linguistic perspective. *Special issue of Journal of Pragmatics* 44 (2):131-137.

Ted Sanders T. and Eve Sweetser (2009). *Causal categories in discourse and cognition*. Walter de Gruyter, Berlin, Germany.

Michel Simard, (2003). Translation spotting for translation memories. *HLT-NAACL 2003, Workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond*.

Wilbert Spooren and Liesbeth Degand (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6 (2): 241-266.

Marc Verhagen, Inderjeet Mani, Roser Sauri, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, John Phillips, James Pustejovsky (2005). Automating

Temporal Annotation with {TARSQI}. *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics* (ACL), Demo Session (pp. 81–84). Ann Arbor, USA.

Jean Véronis and Philippe Langlais (2000). Evaluation of parallel text alignment systems: The arcade project. In *Parallel Text Processing*. Kluwer Academic Publishers, Text Speech and Language Technology Series: 369-388.

Bonnie Webber and Aravind Joshi (2012). Discourse Structure and Computation: Past, Present and Future. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Jeju, Republic of Korea. 42–54,

Deirdre Wilson (2011). The conceptual-procedural distinction: Past, present and future. In Escandell-Vidal, V. et al. (editors), *Procedural Meaning: Problems and Perspectives*, Bingley: Emerald Group Publishing. 3-31.

Sandrine Zufferey (2012). *Car*, *parce que*, *puisque* revisited. Three empirical studies on French connectives. *Journal of Pragmatics* 44(2), 138-153.

Sandrine Zufferey and Bruno Cartoni (2012). English and French causal connectives in contrast. *Languages in Contrast* 12(2):232-250.

Sandrine Zufferey, Liesbeth Degand, Andrei Popescu-Belis and Ted Sanders (2012). Empirical validations of multilingual annotation schemes for discourse relations. *Eighth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*, pages 77-84, Pisa, Italy.