[To appear in Corpus Linguistics and Linguistic Theory]

Annotating the meaning of discourse connectives in multilingual corpora

SANDRINE ZUFFEREY & LIESBETH DEGAND

Abstract
Discourse connectives are lexical items indicating coherence relations between discourse segments. Even though many languages possess a whole range of connectives, important divergences exist cross-linguistically in the number of connectives that are used to express a given relation. For this reason, connectives are not easily paired with a univocal translation equivalent across languages. This paper is a first attempt to design a reliable method to annotate the meaning of discourse connectives cross-linguistically using corpus data. We present the methodological choices made to reach this aim and report three annotation experiments using the framework of the Penn Discourse Tree Bank.

Keywords: *discourse connectives; coherence relations; multilingual annotation; annotation scheme*

1. Importance of a multilingual treatment of connectives

Discourse connectives are lexical items like *however*, *because* and *while* in English. They form a functional category including several grammatical categories such as conjunctions and adverbs, whose function is to convey coherence relations like *cause* or *contrast* between units of text or discourse (e.g. Halliday and Hasan, 1976; Mann and Thomson, 1988; Sanders, Spooren and Noordman, 1992; Knott and Dale, 1994). One of the main characteristics of discourse connectives is that they always relate two different abstract objects in discourse like events, states or propositions (Asher, 1993). This feature distinguishes discourse connectives from discourse markers like *well* and *you know* that take scope over only one abstract object.

Even though lexical or grammatical means to convey coherence relations are found in most languages (Dixon and Aikhenvald, 2009), important variations exist in the number of connectives languages display to express a given relation, even between typologically related languages. To cite a case in point, French uses mainly three different connectives to convey causal relations while Dutch has four (Degand and Pander Maat, 2003; Pit, 2007). The French connective *parce que* corresponds to *omdat* in some cases and to *doordat* in others. And the other pairs of connectives are not equivalent either. For example, the Dutch connective *aangezien* is mostly used in sentence-initial position and is perceived to be formal or even archaic by many speakers (Pit, 2007). By contrast, its French "counterpart" *puisque* is mostly used between clauses and is not associated with a formal register (Zufferey, 2012). These differences become even more noticeable when comparing the observed translations of these connectives. In a bilingual French-Dutch corpus, Degand (2004) found that while *puisque* was translated by *aangezien* in 48% of the occurrences, *aangezien* was translated by *puisque* in only 8% of the occurrences. Similarly, for the French-English pair, Zufferey and Cartoni (2012) found that while *puisque* is translated by *since* in 43.5% of the occurrences, *since* is translated by *puisque* in only 23% of the occurrences. Both studies stress that *puisque* has no equivalent connective that is as strongly associated with the communication of subjective relations. However, as observed in these studies, bilingual dictionaries treat these connectives as translation equivalents. In addition, discourse connectives are in most cases optional, as the coherence relation they convey can often also be left implicit and reconstructed by inference. From a multilingual perspective, this feature also makes cross-linguistic comparisons of connectives difficult, as languages differ in when and how they use them to mark discourse structure.

Another difficulty related to discourse connectives is that they are often polysemic and a single lexical item can be used to convey several coherence relations. For example, the connective *if* can be used to convey a conditional or a causal meaning and the connective *since* can convey a temporal or a causal meaning. Because of these numerous ambiguities and the necessity to grasp sometimes complex coherence relations, discourse connectives are a reputedly difficult class of lexical items to master. The difficulties related to the production and comprehension of connectives have been studied from many different angles. Recent research on normally developing children has for example shown that children as old as 10 years performed significantly worse than adults in a cloze task designed to assess their comprehension and use of connectives (Cain and Nash, 2011). The difficulty is even greater for second language learners, who have been repeatedly found to struggle with connectives in their L2 (Crewe, 1990; Lamiroy, 1994; Granger and Tyson, 1996; Degand and Hadermann, 2009). Connectives are also particularly challenging for translators, who have to adapt them to a new language and culture, in which textual strategies involving the use of connectives are often very different from those of the source text (Baker, 1993; Mason, 1998; Halverson, 2004).

The problem of discourse connectives is made even greater for all these populations by the inadequacy of classical tools such as dictionaries to represent their meaning, as shown above in the case of the *puisque*/*aangezien* and *puisque/since* pairs. Grammars do not fare better for this task, because connectives do not form a unified grammatical category, and their functions often lie outside the scope of individual sentences. Overall, these observations all point to the necessity to develop more adequate resources to describe the meaning of connectives and relate them to one another over various languages.

This paper is a first attempt to design a reliable method to annotate the meaning of discourse connectives cross-linguistically using corpus data. We present the methodological choices made to reach this aim and report a series of annotation experiments designed to define an appropriate taxonomy of discourse relations for multilingual purposes.

2. Representing the meaning of connectives

As argued in Section 1, connectives convey coherence relations between discourse segments. A representation of such coherence relations has been included in several well-known discourse models like Rhetorical Structure Theory (Mann and Thompson, 1988) and Segmented Discourse Representation Theory (Asher and Lascarides, 2003). However, these models have objectives that diverge from our own aims. They seek to provide a complete representation of coherence relations within a text while we want to account for the meaning of connectives only. In this respect, our objective is closer to that of the Penn Discourse Tree Bank (PDTB) developed for English (Prasad et al., 2008), because this framework takes a lexically grounded approach to discourse (even implicit relations have to be expressed in terms of a possible connective) and does not make assumptions about its global structure. In this section, we first describe the PDTB (2.1.), and explain the methodological choices that we made in order to define a hierarchy of relations applicable for multilingual annotations (2.2.).

2.1 The Penn Discourse tree Bank

The Penn Discourse Treebank (PDTB) provides a discourse-layer annotation over the Wall Street Journal Corpus. The discourse annotation consists of manually annotated senses for about 100 types of connectives, corresponding to 18,459 occurrences.

Connectives are defined in the PDTB following Asher's (1993) definition given above, i.e. as lexical items encoding a coherence relation between two abstract objects such as events, states or propositions. This definition includes a range of subordinating conjunctions (e.g. *since*, *although*, *because*), coordinating conjunctions when they are used to relate two clauses (e.g. *and*, *or*, *nor*) and adverbials (e.g. *however, for example, as a result*). These three categories are illustrated in (1) to (3). A case of coordinating conjunction not included in the category of connectives is (4), where *and* relates two noun phrases instead of two clauses, contrary to *but* in example (2). All examples come from the PDTB corpus (The PDTB Research Group, 2007: 8-9).

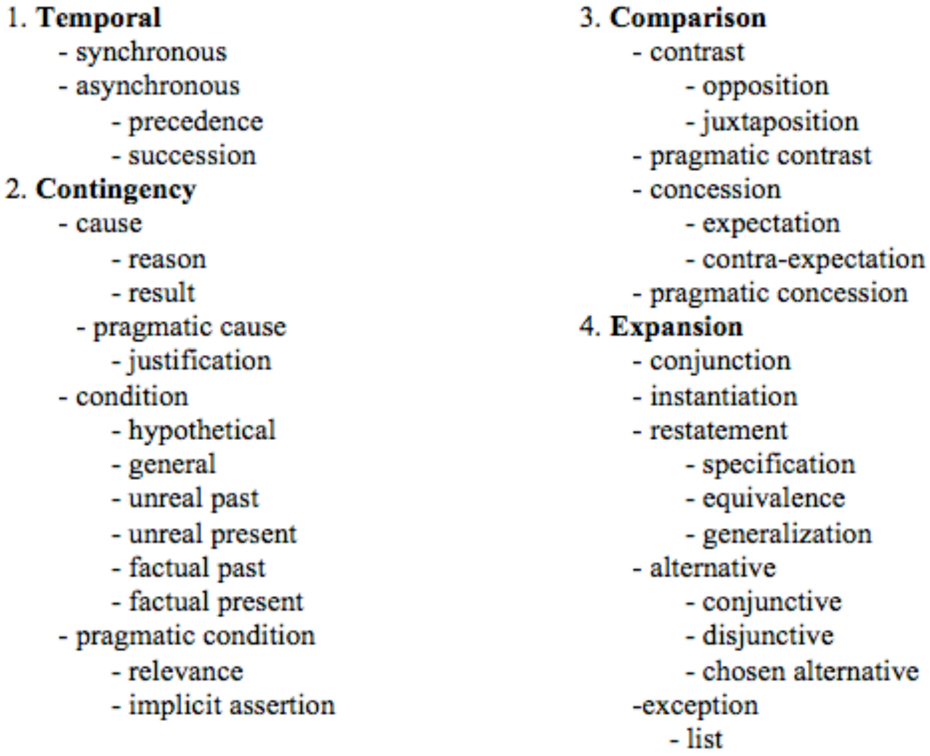
(1) The federal government suspended sales of U.S. savings bonds *because* Congress hasn't lifted the ceiling on government debt.
(2) The House has voted to raise the ceiling to $3.1 trillion, *but* the Senate isn't expected to act until next week at the earliest.
(3) Working Woman, with circulation near one million, and Working Mother, with 625,000 circulation, are legitimate magazine success stories. The magazine Success, *however*, was for years lackluster and unfocused.
(4) Dr. Talcott led a team of researchers from the National Cancer Institute *and* the medical schools of Harvard University and Boston University.

Other clausal adverbials such as *strangely* and *probably* are not included in the category of discourse connectives either, because they only take one abstract object as argument instead of two. The difference between the connective and non-connective categories of adverbials is illustrated in (5) and (6).

(5) John is very clever. He will *however* not get the job.
(6) John is very clever. He will *probably* get the job.

In (5), the adverbial *however* introduces a concession relation between the fact that John is clever with the fact that he will not get the job. These two facts represent two distinct abstract objects. By contrast, in (6) *probably* is only taking scope over one abstract object: the fact that John will not get the job, to which it adds an indication of certainty. That a consequence relation can be inferred from the juxtaposition of the two segments in (6) is not derived from the meaning of *probably* but from encyclopedic knowledge about the relation between being clever and getting a job. Similarly, discourse markers like *actually* and *you know* have not been annotated either, as their role is not to relate two abstract objects but to "signal the organizational or focus structure of the discourse. (The PDTB Research Group, 2007: 8).

The connective types annotated in the PDTB were chosen because of their high frequency in English. The annotation also includes a number of implicit discourse relations and the argument spans of connectives. The coherence relations conveyed by discourse connectives are organized in a hierarchy containing three levels of granularity (from more general to more specific senses), as reported in Figure 1. The annotators of the PDTB were allowed to freely choose tags among all levels, including the possibility to use double tags from any hierarchy levels in order to account for ambiguous cases.



1. **Temporal**
    - synchronous
    - asynchronous
        - precedence
        - succession
2. **Contingency**
    - cause
        - reason
        - result
    - pragmatic cause
        - justification
    - condition
        - hypothetical
        - general
        - unreal past
        - unreal present
        - factual past
        - factual present
    - pragmatic condition
        - relevance
        - implicit assertion
3. **Comparison**
    - contrast
        - opposition
        - juxtaposition
    - pragmatic contrast
    - concession
        - expectation
        - contra-expectation
    - pragmatic concession
4. **Expansion**
    - conjunction
    - instantiation
    - restatement
        - specification
        - equivalence
        - generalization
    - alternative
        - conjunctive
        - disjunctive
        - chosen alternative
    - exception
        - list

Figure 1. The Penn Discourse Tree Bank hierarchy of discourse relations (The PDTB Research Group, 2007: 27).

The PDTB has set the example for a number of other monolingual taxonomies of discourse relations in Czech (Zikánová, Mladová, Mírovský and Jínová, 2010), Arabic (Al-Saif and Markert, 2010), Chinese (Huang and Chen, 2011) and Hindi (Kolachina, Prasad, Sharma and Joshi, 2012). Most of these taxonomies have used the PDTB top-level classification and made a number of adjustments in the sub-levels in order to account for all the specificities of their language. In the next section, we will discuss different constraints emerging from the definition of a taxonomy designed to support multilingual annotations.

2.2 Constraints emerging from a multilingual annotation of connectives

Contrary to monolingual representations like the ones alluded to above, a taxonomy designed for multilingual purposes cannot aim for a total coverage of the specificities of every language. A balance must be reached between the generalization needed to cover multiple languages and the necessity to accurately describe the meanings of connectives in all of them. Given its successful application to a number of languages, often with only minimal changes, the PDTB appears to be a good starting point for such a comparison. In order to test the potential of generalization of the PDTB hierarchy, we have designed an original multilingual annotation experiment, described in Section 3. Based on this experiment, we propose some modifications to the PDTB hierarchy in Section 3.4. Our revised taxonomy is then tested in two additional experiments, reported in Section 4.

An important methodological choice for a multilingual comparison of connectives concerns the type of corpora used for the annotation. In order to ensure optimal comparability between languages, parallel corpora are ideal. However, big parallel corpora are rare and often limited to specific genres (see for example Granger, 2010). We argue that a parallel corpus is mandatory in order to assess the validity of a hierarchy on equivalent occurrences across languages, but once the coherence relations have been adequately defined, comparable corpora provide more flexible and accurate ways to compare connectives across languages (Evers-Vermeul, Degand, Fagard and Mortier, 2011). First, they provide a comparison between connectives that have been used in source texts only and not in translations. Previous studies have demonstrated that connectives are used differently in original texts and in translations (e.g. Degand, 2004; Cartoni, Zufferey, Meyer and Popescu-Belis, 2011; Zufferey and Cartoni, to appear). Moreover, they allow for comparisons across many different genres and are not limited by the availability of translated data. Lastly, connectives are very volatile items in translation (Halverson, 2004), and the use of parallel corpora implies that an important number of occurrences have to be discarded because they have been left out or added in the process of translation. An assessment of the magnitude of these discrepancies will be provided in the next Section.

When more than two languages are annotated simultaneously, another important issue is to define a reference against which all languages can be compared. Ideally, a language should be compared to all the others. However, because of the important variability in the use of connectives across languages, this aim is difficult to achieve in practice. If a pivot language is chosen, the occurrences of connectives to be annotated are defined according to this language, and are then selected in a similar way in all other languages. For example, if English is chosen as a pivot language, the tokens of connectives to be annotated are selected based on the English corpus and only connectives that are the translations of these tokens in the other languages are annotated. All connectives that are not translated or that are added in the target texts are discarded. This restriction allows for a more systematic comparison of the same tokens between the languages, because they are translation equivalents. We have implemented these methodological principles in the experiment described in the next Section.

3. A multilingual annotation experiment using the PDTB taxonomy

We conducted an original annotation experiment with five Indo-European languages, pertaining to the Germanic and Romance families: English, French, German, Dutch and Italian. In order to facilitate comparisons, we have decided to use English as a pivot language, as explained in Section 2. In this Section, we present the data used in this experiment (3.1.) and the annotation procedure (3.2.). We discuss its main results (3.3.) with the conclusion that some parts of the PDTB hierarchy need to be modified in order to reach a reliable annotation, optimally relevant for the cross-linguistic comparison of connectives. This new version of the hierarchy is presented in Section 3.4.

3.1 Description of data

In order to compare and annotate connectives in five languages, a small translation corpus made of four journalistic texts gathered from the Press Europe website[1] was built. The size of the corpus was around 2,500 words per language. All four texts came from different European newspapers, and the source language was different in all of them (namely: German, Romanian, Dutch and Slovak). The source languages were varied in the corpus in order not to bias the occurrences of coherence relations based on a single language and to simulate the case of a large multilingual database in which occurrences of connectives come from both original and translated texts. In the English version of the corpus, used as a pivot language for the annotation, 54 tokens of connectives were identified, corresponding to 23 different connective types. The criteria used to select tokens of connectives were similar to those applied in the PDTB project and described in Section 2.1. The list of these connectives is detailed in

Table 1.

Table 1. List of connective types in English with their token frequency.

after (1)	before (1)	in as much as (1)	though (2)
after all (1)	but (11)	meanwhile (1)	thus (2)
and (7)	despite (1)	nevertheless (3)	when (4)
as (1)	for instance (1)	so (1)	whereas (1)
as long as (1)	however (4)	then (1)	while (1)
because (2)	if (2)	therefore (2)	

3.2 Procedure

In every language, the annotation task was performed independently by two annotators. All annotators were linguists, with a special interest in discourse and having previous experience in linguistic annotation, ranging from PhD students who had completed one or several previous annotation tasks to senior researchers with up to fifteen years of annotation practice. All annotators were multilingual, and spoke at least English in addition to the language they were asked to annotate. However, they only performed annotations in their mother tongue (expect for the reference annotation in English, performed by the two authors) and did not have access to the corpus in any other language than the one they annotated, once the target connectives were identified.

The tokens of discourse connectives to be annotated were spotted on the English version of the corpus by the two authors. For every other language of the study, one annotator was asked to spot the translation equivalents. All tokens of connectives that had been translated in the target text by a connective were annotated with a discourse relation from the PDTB hierarchy by two annotators. Relations that had not been translated by a connective in the target language were not annotated.

All annotators were asked to use the definition of discourse relations provided in the PDTB annotation manual (The PDTB Research Group, 2007). As it was the case in the PDTB project, annotators were instructed to use tags from the most precise level from the hierarchy (third level) if they were confident about the relation or more generic relations in case of doubt. Annotators were also allowed to use double tags in two different cases: when they felt that the relation was ambiguous and that either one of the two chosen tags applied; when they felt that two tags had to be added in order to describe the meaning of the relation. In the first case, the two tags had to be linked with OR and in the second with AND. For example, in (7) from our corpus, the relation conveyed by *when* could arguably be either temporal or conditional. In (8) however, the relation conveyed by *as long as* both contains a temporal and a conditional meaning. The situation described in argument 1 lasts temporally only on the condition that the situation described in argument 2 holds true[2]. The meaning of *as long as* is therefore both temporal and conditional.

(7) The cliché of a Mediterranean lolling in the sun has become a mental reflex *when* trying to explain the cause of the crisis in the Eurozone.

(8) *As long as* we fail to take governments in developing countries seriously, international climate change policy is doomed to failure.

In the annotation, double tags indicating multiple meanings such as (8) were used by the annotators but tags indicating potential ambiguities as in (7) were seldom used, showing that annotators often formed one single mental representation of the meaning conveyed by connectives and were not aware of potential alternative meanings. These ambiguities were revealed when comparing several annotations of the same token.

3.3 Results

The first task given to the annotators was to identify translation equivalents between English and their own language. This first comparison provided an estimation of the magnitude of cross-linguistic divergences. In some cases, the target text did not contain any translation of the English connective or the meaning was rendered by a paraphrase. These connectives were therefore missing with respect to the English text. Annotators were also asked to count the number of connectives present in the target text (following the same criteria as those applied for English) that were not equivalents of English connectives, thus constituting additions resulting from the translation process. These connectives conveyed relations from all four top-level categories from the PDTB classification. Results from these comparisons are reported in Table 2. These results indicate that the use of a parallel corpus and a pivot language imply an important loss of connectives for the annotation. On average, this loss represents 50% of the number of occurrences that were annotated.

Table 2. Variation in the number of connectives used with respect to English corpus.

	French	**German**	**Dutch**	**Italian**
missing connectives	10	10	7	18
paraphrases	1	2	0	0
additional connectives	6	12	19	15

Another notable result from Table 2 is that paraphrases were rarely used as a translation equivalent of the lexicalized connectives from our English corpus. This does not mean however that paraphrases are not an important lexical means of communicating coherence relations. In the PDTB, a wide range of so-called ‘alternative lexicalizations’ has been identified as possible markers of such relations (e.g. Prasad, Joshi and Webber, 2010). Despite their importance for a global theory of discourse structuring devices, these lexicalizations have however not been taken into account in the pilot experiments reported in this paper.

The inter-annotator agreement was computed from a monolingual and from a cross-linguistic perspective. Percentages instead of other measures of inter-annotator agreement such as Cohen’s Kappa scores are reported throughout the paper, in order to ensure that our results are comparable with those of previous experiments conducted with the PDTB, that also report percentages. In addition, Spooren and Degand (2010) argue that Kappa scores provide an inaccurate picture of inter-annotator agreement for linguistic tasks like ours, because the observed Kappa scores almost never correspond to reliable agreements. The percentage of agreement for the two annotators working on the same language is reported in Table 3.

Table 3. Monolingual inter-annotator agreement.

	English	**French**	**German**	**Dutch**	**Italian**	**Average**
level 1	98%	95%	95%	91%	94%	95%
level 2	67%	69%	72%	60%	64%	66%
level 3	46%	47%	53%	39%	44%	46%

Results from Table 3 indicate that the level of agreement is similar across languages. In every case, the agreement is very good at the first level in the taxonomy (95% on average), medium at level 2 (66% on average) but poor at level 3 (46% on average). While agreement was computed separately for each level of annotation, agreement scores are interdependent, because disagreement at a higher level automatically leads to disagreement on a lower one. Furthermore, agreement scores were given only when alternatives were possible. For instance, the *conjunction* relation (level 2 of the level 1 *expansion* relation) does not offer any alternatives at level 3. Therefore, agreement was computed only on the first and second levels, not on the third one.

In the PDTB, the inter-annotator agreement was 92% at the top-most level and 77% at the third level of the hierarchy (Mitsalkaki, Robaldo, Lee and Joshi, 2008). The important difference with the average agreement at the third level in our experiment indicates that agreement at this level can increase with training and discussion (see also Bayerl and Paul, 2011).

The percentage of agreement for the four dimensions of level 1 is provided in Table 4.

Table 4. Monolingual inter-annotator agreement for each level 1 dimension.

	English	**French**	**German**	**Dutch**	**Italian**	**Average**
Temporal	100%	100%	86%	100%	80%	93%
Contingency	100%	92%	100%	100%	100%	98%
Comparison	95%	95%	95%	95%	94%	95%
Expansion	100%	100%	100%	71%	100%	94%

At level 1, the few disagreements observed are not always recurrent across languages, with the exception of comparison relations that lead to a similar number of disagreements across languages. At level 2 however, these disagreements are more recurrent across languages. Problematic cases mostly concern the distinction between concession and contrast, for which the annotators agree in only 50% of the relations, when the *comparison* tag is used. This agreement even drops to 40% on average at the third level (distinctions between *opposition* and *juxtaposition* for contrast and between *expectation* and *contra-expectation* for concession). Moreover, for the relations tagged as *condition*, the agreement for the third level tags (*hypothetical*, *general*, etc.) is also only 40%. Taken together, these cases represent on average 87% of the disagreements at the third level of the hierarchy. Finally, the use of the *pragmatic* tags from the PDTB scheme was very problematic, as an agreement on the use of this tag was reached only in 16% on the cases on average, and some annotators didn't use it at all. A cross-linguistic evaluation of inter-annotator agreement is reported in Table 5[3].

Table 5. Average cross-linguistic inter-annotator agreement with English.

	English/ French	**English/ German**	**English/ Dutch**	**English/ Italian**	**Average**
level 1	91%	90%	88%	85%	88.5%
level 2	67%	65%	63%	60%	64%
level 3	42%	45%	34%	35%	39%

An analysis of cross-linguistic disagreements reveals two distinct phenomena. At the top level of the hierarchy, disagreements are systemically more numerous cross-linguistically than monolingually (95% vs. 88.5% on average). This rise of disagreements always corresponds to meaning shifts due to translation. For example, the connective *when*, annotated with a temporal tag in English was once translated by *alors que*, a connective annotated with a contrast tag by French-speaking annotators. Similar cases of meaning shift occur on average in 10% of the cases in every language. This problem shows the limitations of using parallel corpora, under the assumption that connectives are translation equivalents across languages. This problem is moreover not limited to discourse connectives, translated texts differ in many respects from original ones (e.g. Baroni and Bernardini, 2006). An annotation of comparable corpora, where equivalences are established based on the similarity of coherence relations, does not run into similar problems.

For lower levels of the hierarchy, differences in the annotation could not be related to changes in translation but rather to genuine disagreements between annotators regarding the interpretation of a given relation.

The first annotation experiment described above clearly indicated that the areas of disagreements were recurrent across annotators and languages. In order to reach a more reliable annotation that can be applied cross-linguistically, some adjustments were made to the PDTB taxonomy.

3.4 Revising the PDTB taxonomy

Our goal in revising the PDTB for multilingual annotations is twofold: produce a taxonomy of discourse relations that is fine-grained enough to capture the differences of meaning between connectives across languages, and optimize inter-annotator agreement in order to produce reliably annotated data. These objectives stand in opposition, as capturing fine-grained differences of meaning requires to keep or even add many third level sense tags in the taxonomy, but these tags are precisely those producing a high number of inter-annotator disagreements. In view of these objectives, we only pruned senses that did not match differences between connectives and improved the definition of senses that were problematic for the annotators but could not be removed without producing inadequate pairings of connectives across languages.

Two examples of senses that were pruned from the taxonomy are the sub-categories of conditional and alternative relations (cf. Figure 1). In both cases, all sub-types correspond to one single connective, for example *if*, *si* or *als* for conditional relations. Removing them is therefore not detrimental for the representation of connectives' meaning. On the other hand, some sub-senses leading to an important number of disagreements have been kept in the taxonomy because they match differences between connectives. Two examples of this phenomenon are contrastive vs. concessive and pragmatic vs. non-pragmatic relations. For all these cases, we argue that inter-annotator agreement has to be improved by providing annotators with ways to operationalize the differences of meaning, as we now outline.

An important source of disagreements in our experiment was the distinction between concessive and contrastive relations, for which agreement was at chance level. Contrary to what has been done in some monolingual adaptations of the PDTB (Al Saif and Markert, 2010), we argue that this distinction cannot be removed from the taxonomy because both kinds of relations can be expressed by connectives that are not interchangeable in the languages of our study. For example, in French the connective *alors que* can only express a contrastive relation while connectives like *bien que* and *même si* can only express a concessive relation. Conversely, the third level tags from the PDTB in this category (i.e. *juxtaposition* vs. *opposition* for contrast and *expectation* vs. *contra-expectation* for concession) can be removed from the taxonomy, because they do not contribute to make additional distinctions between connectives while decreasing inter-annotator agreement from 50% to 40%.

In the literature, a series of criteria to account for the differences between concession and contrast have been identified (see Taboada and de los Ángeles Gómez-González, 2012 for a review). In order to improve inter-annotator agreement for these cases, we have operationalized the tests proposed by Lakoff (1971), who claims that contrastive relations differ from concessive relations in that they offer the possibility to: (1) reverse the two connected segments and (2) convey the relation implicitly or replace it by a neutral coordination with *and*. An additional test can be applied by using a paraphrase: a contrastive connective can always be substituted with the locution "by contrast". For example, the connective *whereas* in (9) from our corpus conveys a contrast between the percentage of civil servants in Greece and in other European countries. All three tests proposed above to assess contrastive meanings are satisfied: the connective can be removed without losing a contrastive interpretation, the order of the segments can be reversed and the connective can be replaced by the locution "by contrast".

(9) Greek civil servants account for 22.3% of the workforce, *whereas* this figure stands at 30% for France, 27% for the Netherlands, and 20% for the United Kingdom.

According to Taboada and de los Ángeles Gómez-González (2012: 22) "what is mutually exclusive in concessives is found between the propositional content of one clause and an assumption evoked in the other segment". Typically, as observed by Anscombre and Ducrot (1977), the first argument of a concessive relation leads to a certain conclusion and the second argument leads to the reverse conclusion, as illustrated in (10) from our corpus. The first segment leads to the conclusion that people sympathise with the poor but the second segment reverses this conclusion. Contrary to (9), this relation cannot be paraphrased by the locution "by contrast". In addition, the two related segments cannot be reversed without modifying the conclusion drawn from the relation and the oppositive meaning is difficult to retrieve when the connective *but* is removed. Thus, all three tests indicate that the relation is concessive.

(10) Normally, poverty should inspire feelings of compassion. But neo-liberal economic populism succeeds in extirpating such sentiments.

By integrating these linguistic tests, we hope to increase annotators' awareness of the distinctions between contrastive and concessive relations, and therefore increase the level of inter-annotator agreement.

The last major source of disagreement in our experiment concerned the use of *pragmatic* tags. Again, this distinction cannot be pruned because both types of relations are prototypically expressed by specific connectives in some languages like Dutch (see Sanders and Stukker, 2012 for a cross-linguistic illustration in the causal domain). In the PDTB taxonomy, the kind of examples grouped under this category is not always clearly defined and exemplified. For example, while pragmatic contrast is defined in the PDBT annotation manual as: "a contrast between one of the arguments and an inference that can be drawn from the other", the notion of pragmatic concession is not given any definition or example. In our revised version, the *pragmatic* tags include all occurrences corresponding to speech-act (11) and epistemic (12) uses of connectives, as defined by Sweetser and illustrated below with the causal connective *because* (1990).

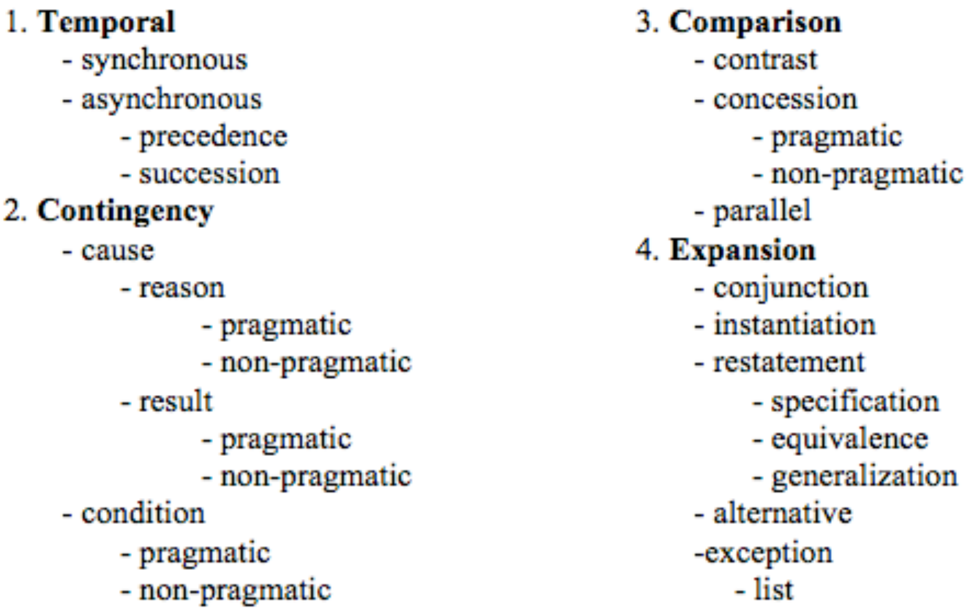
(11) Are you coming? Because we are late.
(12) Max is ill, because he did not come to work today.

Following Sanders (1997), we propose to disambiguate these two types of relations by a paraphrase test. If X causes Y to happen in the real world the relation is non-pragmatic. If X causes the speaker to claim or conclude Y the relation is pragmatic.

The pragmatic uses of connectives thus defined can occur for causal, conditional and concessive connectives. Therefore, for these tags, an additional annotation level has been added to account for the pragmatic/non-pragmatic dimension. In the case of causals, this change involved the addition of a fourth level in the hierarchy.

Finally, one single tag was added in the comparison category through the insertion of a parallel sense, in order to account for the meaning of connectives like *similarly* and *as if* that do not have a straightforward tag in the PDTB taxonomy.

All these changes lead to the revised taxonomy described in Figure 2. These changes are moreover to a large extent convergent with previous monolingual adaptations of the PDTB for typologically diverse languages like Arabic (Al-Saif and Markert, 2010) and Hindi (Kolachina et al. 2012).



1. **Temporal**
    - synchronous
    - asynchronous
        - precedence
        - succession
2. **Contingency**
    - cause
        - reason
            - pragmatic
            - non-pragmatic
        - result
            - pragmatic
            - non-pragmatic
    - condition
        - pragmatic
        - non-pragmatic
3. **Comparison**
    - contrast
    - concession
        - pragmatic
        - non-pragmatic
    - parallel
4. **Expansion**
    - conjunction
    - instantiation
    - restatement
        - specification
        - equivalence
        - generalization
    - alternative
    - exception
        - list

Figure 2: Revised taxonomy based on the results of multilingual annotation.

4. Two annotation experiments with the revised taxonomy

Given that our first experiment indicated that disagreements were not on average more numerous cross-linguistically than monolingually, we have first tested the revised version of the taxonomy with a monolingual annotation in French. This new task, described in Section 4.1, confirmed that our taxonomy was operational and provided improvements in the level of inter-annotator agreement with respect to the original PDTB taxonomy. We have therefore tested it in a larger-scale cross-linguistic annotation, described in Section 4.2 to further assess its validity and the reliability of our initial results.

4.1. A monolingual annotation experiment in French

A second corpus of 3,117 words in original French texts was assembled from the Press Europe website, following similar principles as those described in the first experiment. This second corpus contained 54 occurrences of connectives, corresponding to 20 different connective types, summarized in Table 6. Three French-speaking annotators made the annotation independently. The procedure was identical to that of Experiment 1.

Table 6. List of connective types from the second French corpus with their token frequency.

alors (1)	depuis (1)	lorsque (1)	pourtant (1)
alors que (2)	donc (1)	mais (15)	puis (1)
cependant (1)	en fait (1)	néanmoins (2)	si (4)
certes (1)	en revanche (2)	parce que (3)	tandis que (1)
de même (1)	et (10)	pendant que (1)	toutefois (4)

The inter-annotator agreement for this second annotation task is reported in Table 7.

Table 7. Inter-annotator agreement for the second annotation task with revised taxonomy.

	Annotators 1 and 2	Annotators 1 and 3	Annotators 2 and 3
level 1	94.5%	92.5%	96%
level 2	82%	79%	81%
level 3	65%	85.5%	69%
level 4	66%	100%	66%

These results indicate that the modifications made to the taxonomy did provide some improvements. Notably, the cases of disagreement between the *contrast* and *concession* tags decreased from 50% to 28% on average, with the result that pairwise agreement scores at the second level improves with respect to the first annotation (81% vs. 66% on average for the first annotation). The introduction of the pragmatic/non-pragmatic tag at the third and fourth levels did not result in lower agreement scores but did not strongly improve results either (16% of consistent use vs. 20% in Experiment 2), indicating that this distinction remains a difficult one to annotate, as was previously observed by Spooren and Degand (2010). Despite this difficulty, this distinction must be preserved in the taxonomy in order to distinguish between the meaning of some connectives, like the Dutch causal connectives *omdat* (non-pragmatic) and *want* (pragmatic).

4.2. Larger-scale cross-linguistic annotation with revised taxonomy

A third corpus was assembled from the Press Europe website, including the same five languages used in Experiment 1. This corpus of about 8,500 words per language contained in English 203 tokens of connectives corresponding to 36 different types, reported in Table 8.

Table 8. List of connective types from the third corpus with their token frequency.

after (1)	even if (4)	in short (1)	then (3)
although (6)	for example (3)	in spite of (1)	therefore (3)
and (50)	for instance (1)	indeed (1)	though (5)
as (3)	given (that) (2)	meanwhile (1)	thus (2)
as well as (1)	however (7)	now (2)	well (1)
because (5)	if (11)	or (5)	when (7)
before (4)	in fact (1)	since (1)	whether (2)
but (41)	in order to (1)	so (2)	while (9)
despite (6)	in other words (1)	that is why (1)	yet (8)

In every language, the translation equivalents were spotted and the coherence relations conveyed by these connectives were annotated with the revised taxonomy described in Figure 2. Cross-linguistic results from this third annotation task are reported in Table 9.

Table 9. Cross-linguistic inter-annotator agreement.

	English/French	English/German	English/Dutch	English /Italian
level 1	94%	93%	88%	93%
level 2	85%	74%	75%	78%
level 3	75%	66%	69%	66%
level 4	66%	93%	62.5%	70%

These results confirm the validity of our second monolingual annotation experiment, with cross-linguistic data. A paired-samples t-test was conducted to compare the percentage of agreement between our initial experiment and the new experiment involving the revised taxonomy. At level 1, the difference between the agreements (for all language pairs) reached in the first experiment (M = 88.5, SD = 2.65) and the second experiment (M = 92, SD = 2.7) is not significant t(3) = 2.11, p = 0.125. The increase is however significant at level 2 between the first experiment (M = 63.75, SD = 2.99) and the second experiment (M=78, SD = 4.97): t(3) = 6.33 (3), p < 0.01. At level 3, the difference between the first experiment (M = 39, SD = 5.35) and the second experiment (M = 69, SD = 4.24) is also significant: t(3) = 9.65, p < 0.01. The lack of improvement at level 1 was expected, as we did not make any modification at this level. The significant improvement observed at the lower levels tends to indicate that our modifications are on the right track and contribute to improve inter-annotator agreement. This experiment also confirmed that most disagreements at the first level of the taxonomy were due to meaning shifts in translation.

In this experiment, the coverage of relations and connective types was more important than in the first ones. The numbers of occurrences for level 2 relations found in the English corpus are reported in Table 10.

Table 10. No of tokens of level 2 relations in the revised taxonomy.

Level 1	Level 2	No of relations
temporal	synchronous	9
	asynchronous	10
contingency	cause	20
	condition	12
comparison	concession	69
	contrast	19
	parallel	0
expansion	alternative	7
	conjunction	46
	instantiation	3
	restatement	4
	exception	0
	list	0
Total		**203**

The more extensive coverage of connective types and relations did not reveal the need for additional distinctions in the taxonomy nor the existence of important differences between the languages. However, some relations especially in the expansion class were still underrepresented or even not represented at all in the corpus and some connectives were assessed on the basis of one single occurrence (cf. Table 8). A more extensive annotation is therefore still needed before strong conclusions can be reached for these relations.

5. Further steps for testing and implementing a taxonomy of discourse relations for multilingual purposes

Based on our initial annotation experiment, we have designed a revised version of the PDTB that seems to be operational to support a cross-linguistic annotation of discourse relations conveyed by connectives in some Indo-European languages. The coverage of this revised version is adequate, as our tokens of connectives seldom required a relation not found in the taxonomy. Arguably, this lack of problematic cases could come from the fact that the PDTB was designed for English and used to compare languages from closely related families. In addition, our experiments were still English-centered, as the annotation of connectives was dependent on their presence in the English texts. It is therefore possible that connectives specific to other languages that were not spotted because they do not have equivalents in English texts will require some additional relations. However, the fact that the PDTB taxonomy has been adapted to languages from different families such as Arabic, Chinese and Hindi without adding many new senses indicates that most senses can be carried over to languages from different families.

The next step of our experiments will be to assess whether the granularity of our revised taxonomy is precise enough to match translation equivalents across languages. In other words, to determine if all occurrences of connectives labeled with, for example, a *contrast* tag in language A really are translation equivalents of connectives annotated with the same *contrast* tag in language B. Obviously, some additional information regarding syntactic constraints (e.g. prototypical position in the sentence, verb mood, etc.) and register/modality (formal, oral, etc.) will have to be provided to prevent inadequate pairings, but we argue that this information is independent of the semantic content of connectives, conveyed by discourse relations and annotated in our experiments. Only a systematic assessment of cross-linguistic equivalences provided by the taxonomy for all relations will provide a final answer to this question. Previous contrastive works however already indicate that some additional features may be needed. For example, in the causal domain, in addition to the pragmatic/non-pragmatic tag, Zufferey and Cartoni (2012) showed that an important difference between connectives was the status of the cause segment, that can be either “given” (i.e. mutually manifest to the speaker and his audience) for connectives like *given that* and *as* or “new” for connectives like *because*. The applicability of this feature to other coherence relations should also be assessed. Another additional step in this evaluation will be the inclusion of data pertaining to different text genres. Indeed the type of connective used in a text is related to its genre, some connectives being associated with formal written mode and others exclusively used in speech, and a robust taxonomy should be applicable in all of them.

Another difficulty for the annotation of the coherence relations conveyed by connectives is that connectives can be used in some contexts to convey a different relation than the one that they prototypically convey. The most well known case of this type of underdetermination is the connective *and*, that often conveys a more specific relation than its prototypical meaning of addition, notably a temporal or a causal meaning (e.g. Spooren, 1997; Carston, 2002). This phenomenon is also applicable to other connectives, for example temporal connectives may at times convey a causal or a contrastive relation. Therefore, an important question is to define what level of meaning (semantic or pragmatic) has to be annotated. The pragmatic relation conveyed in context is more relevant to understand the contribution of a connective in a given utterance than its core semantic meaning. However, relations that differ in context from the semantic meaning of a connective give rise to an important number of disagreements between annotators, probably because they tend to rely on their perceived core semantic meaning of a connective. In order to help annotators including these pragmatic meanings derived from context, a list of such possible meanings, once derived from empirical data, could be provided to the annotators. Indeed, no connective can be used to convey all types of relations, even in a particular context. Therefore, once the range of possible inferences is established, providing annotators with such a list would help to reduce the range of possibilities and hence the number of disagreements.

5. Conclusion

In this paper, we have presented three original multilingual annotation experiments of discourse connectives, performed on parallel corpora. Our results indicate that with some adjustments designed to maximize the number of features matching distinctions between connectives, the PDTB taxonomy provided an adequate framework for multilingual annotations of discourse connectives. Our experiments also indicate that our revised version of the PDTB taxonomy remains descriptively adequate to account for the meaning of all connective types found in our corpora, but larger-scale annotations involving more relation types and connective tokens should further validate these initial conclusions.

Further work to assess the validity of this taxonomy for multilingual purposes will consist of a systematic evaluation of the cross-linguistic equivalences emerging from the use of similar tags across languages. Another important dimension will be the inclusion of implicit relations as possible translation equivalents. For example, in French a frequent clausal link to announce an explanation is the connective *en effet*. But in English, this connective is most often left out and the link is made through juxtaposition. The annotation of implicit relations will provide a systematic assessment of the variations in the explicit/implicit marking strategies between languages. Another related issue is the analysis of connectives that are added in the process of translation, that is those appearing in the parallel texts but not in the pivot language text (cf. Table 2 in Experiment 1). From a typological point of view, these connectives are interesting because they might tell us something about the type of coherence relations that are preferably marked in one language, and not in another. Here again, the use of comparable rather than parallel corpora is required in order to avoid confounding translation effects. In addition, texts from different genres should be included in future work to account for possible stylistic effects.

Acknowledgement

This work was supported by a Marie-Curie IEF grant for the DISCOM project (299185) awarded to the first author. The authors thank all the annotators who took part in the experiments described in the paper.

Bionote

Sandrine Zufferey (born 1978, PhD University of Geneva, 2007) is a post-doctoral research fellow at the Utrecht Institute of Linguistics in the Netherlands. Her research focuses on the acquisition and processing of discourse connectives. Her work also takes a cross-linguistic perspective in order to study the way specific constraints in different languages can affect cognitive processes.

Liesbeth Degand (born 1967, PhD University of Louvain, 1997) is a professor in Linguistics at the

University of Louvain (Louvain-la-Neuve, Belgium). Her main research interests go to the (corpus-based) study of discourse structure, especially discourse segmentation, and discourse markers in Dutch and French, both in oral and written language, in synchrony and diachrony, and in native as well as learner language.

Notes

---

[1] http://www.presseurop.eu/en

[2] For subordinating conjunctions, argument 2 corresponds to the argument immediately following the connective, whereas argument 1 can either precede the connective or follow argument 2. For coordinating conjunctions and adverbs, arguments are given in linear order.

[3] To compute this cross-linguistic inter-annotator agreement, we compared the means of the scores of the two annotators in the monolingual annotation experiment for Dutch, French, German, and Italian, with those for English.

References

Al-Saif, Amal and Katia Markert. 2010. *The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic*. *Proceedings of The Seventh International Conference on Language Resources and Evaluation*. 2046–2053.

Anscombre, Jean-Claude and Oswald Ducrot. 1977. Deux mais en français? *Lingua* 43. 23–40.

Asher, Nicholas. 1993. *Reference to abstract objects in discourse*. Dordrecht: Kluwer.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Baker, Mona. 1993. *In other words. A coursebook on translation*. London/New York: Routledge.

Baroni, Marco and Silvia Bernardini. 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.

Bayerl, Petra and Karsten Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics* 37(4). 699–725.

Cain, Kate and Hannah Nash. 2011. The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology* 103. 429–441.

Carston, Robyn. 2002. *Thoughts and utterances*. *The pragmatics of explicit communication*. Oxford: Blackwell.

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

Crewe, William. 1990. The illogic of logical connectives. *ELT Journal* 44. 316–325.

Degand, Liesbeth. 2004. Contrastive analyses, translation, and speaker involvement: The case of puisque and aangezien. In Michel Achard and Suzanne Kemmer (eds.), *Language, culture and mind*, 1–20. Stanford: CSLI Publications.

Degand, Liesbeth and Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the speaker involvement scale. In Arie Verhagen and Jeroen Maarten van de Weijer (eds.), *Usage-based Approaches to Dutch*, 175–199. Utrecht: LOT.

Degand, Liesbeth and Pascale Hadermann. 2009. Structure narrative et connecteurs temporels en français langue seconde. In Eva Havu et al. (eds.), *La langue en contexte*. *Actes du colloque Représentations du sens linguistique IV*, 19–34. Helsinki: Société Néophilologique.

Dixon, Robert and Alexandra Aikhenvald. 2009. *The semantics of clause linking. A cross-linguistic typology*. Oxford: Oxford University Press.

Evers-Vermeul, Jacqueline, Liesbeth Degand, Benjamin Fagard, and Liesbeth Mortier. 2011. Historical and comparative perspectives on subjectification: A corpus-based analysis of Dutch and French causal connectives. *Linguistics* 49. 445–478.

Granger, Sylviane and Stephanie Tyson. 1996. Connector usage in English essay writing of native and non-native EFL speakers of English. *World Englishes* 15. 19–29.

Granger, Sylviane. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University* 2. 14-21.

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

Halverson, Sandra. 2004. Connectives as a translation problem. In Harald Kittel et al. (eds.), *An international Encyclopedia of translation studies*, 562–572. Berlin/New York: Walter de Gruyter.

Huang, Hen-Hsen and Hsin-His Chen. 2011. Chinese discourse relation recognition. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 1442–1446. Hiang Mai: Thailand.

Knott, Alistair and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes* 18. 35–62.

Kolachina, Sudheer, Rashmi Prasad, Dipti Sharma, and Aravind Joshi. 2012. Evaluation of discourse relation annotation in the Hindi discourse relation bank. *Proceedings of LREC 2012,* 823–828. Istanbul: Turkey,

Lakoff, Robin. 1971. If's and's and but's about conjunctions. In Charles Fillmore and D. Terrence Langendoen (eds.), *Studies in linguistic semantics*, 114–149. New-York: Holt, Rinehart and Winston.

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

Mann, William and Sandra Thomson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8. 243–281.

Mason, Ian. 1998. Discourse connectives, ellipsis and markedness. In Leo Hickey (ed.), *The pragmatics of translation*, 170–184. Clevedon/Philadelphia: Multilingual Matters.

Miltsakaki, Eleni, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the Penn discourse treebank. *Lecture Notes in Computer Science* 4919. 275–286.

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

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, PLACE, 2961–2968.

Prasad, Rashmi, Aravind Joshi and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. Proceedings of COLING 2010. 2023–2031.

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

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

Sanders, Ted and Ninke Stukker. 2012. Causal connectives in discourse: A cross-linguistic perspective. *Journal of pragmatics* 44(2). 131–137.

Spooren, Wilbert. 1997. The processing of underspecified coherence relations. *Discourse Processes* 24. 149-168.

Spooren, Wilbert and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6. 241–266.

Sweetser, Eve. 1990. *From etymology to pragmatics*. Cambridge: Cambridge University Press.

Taboada, Maite and Maria de los Ángeles Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences* 6. 17–41.

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

Zikánová, Sarka, Lucie Mladová, Jiri Mírovský, and Pavlina Jínová. 2010. Typical cases of annotators' disagreement in discourse annotations in Prague dependency treebank. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 2002–2006.

Zufferey, Sandrine. 2012. 'Car, parce que, puisque' revisited: Three empirical studies on French connectives. *Journal of Pragmatics* 34. 138–153.

Zufferey, Sandrine and Bruno Cartoni. 2012. English and French causal connectives in contrast. *Languages in Contrast* 12. 232–250.

Zufferey, Sandrine and Bruno Cartoni. to appear. A multifactorial analysis of explicitation in translation. *Target*.