[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).

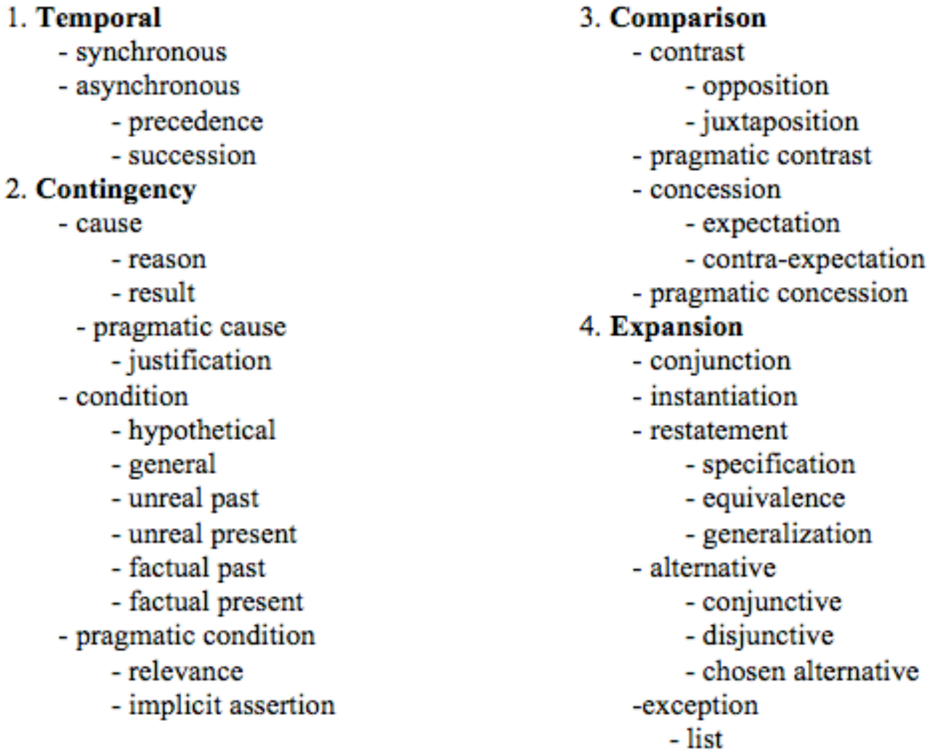
(1) The federal government suspended sales of U.S. savings bonds *because* Congress hasn't lifted the ceiling on government debt.
(2) The House has voted to raise the ceiling to $3.1 trillion, *but* the Senate isn't expected to act until next week at the earliest.
(3) Working Woman, with circulation 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)
after all (1)
and (7)
as (1)
as long as (1)
because (2)
before (1)
but (11)
despite (1)
for instance (1)
however (4)
if (2)
in as much as (1)
meanwhile (1)
nevertheless (3)
so (1)
then (1)
therefore (2)
though (2)
thus (2)
when (4)
whereas (1)
while (1)

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).

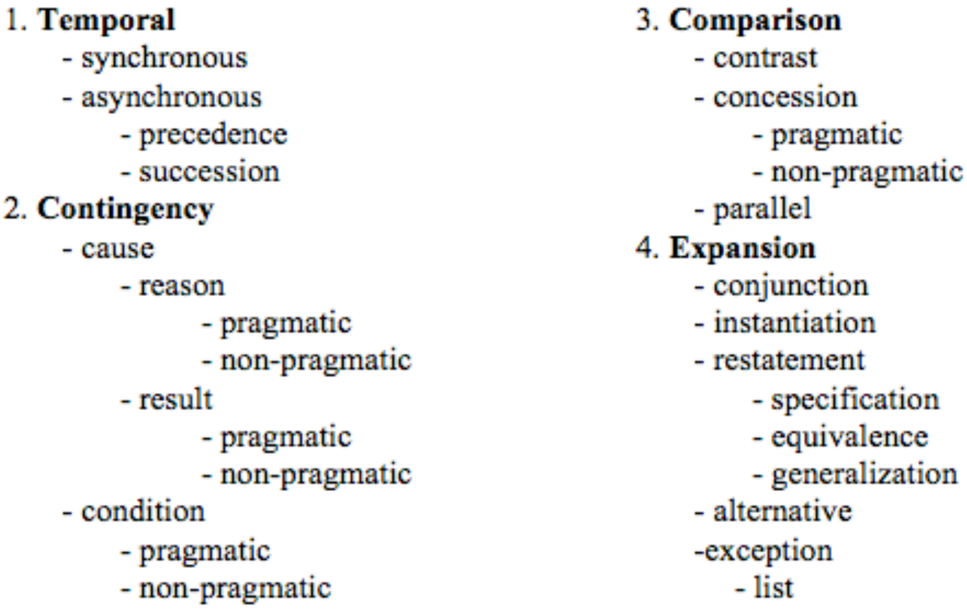
(11) Are you coming? Because we are late.
(12) Max is ill, because he did not come to work today.

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*.

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

Received 02/12; First letter 10/12; Accepted 12/12;
Final revision 02/13; Published online 04/13

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

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.	%
alors que	91	18.24%
gerund	85	17.03%
paraphrases	72	14.43%
si	54	10.82%
zero translation	41	8.22%
tandis que	39	7.82%
même si	33	6.61%
bien que	26	5.21%
s'il est vrai que	14	2.81%
tant que	10	2.00%
pendant	5	1.00%
puisque	5	1.00%
lorsque	4	0.80%
mais	4	0.80%
malgré	3	0.60%
quoique	3	0.60%
pendant que	2	0.40%
alors même que	1	0.20%
aussi	1	0.20%
avant que	1	0.20%
contre	1	0.20%
en même temps que	1	0.20%
étant donné que	1	0.20%
quand	1	0.20%
s'il est exact que	1	0.20%
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.	**%**
car	76	38.78%
parce que	63	32.14%
paraphrases	27	13.78%
zero translation	8	4.08%
dans la mesure où	6	3.06%
puisque	3	1.53%
en effet	3	1.53%
étant donné que	1	0.51%
à défaut	1	0.51%
vu que	1	0.51%
dès lors que	1	0.51%
gerund	1	0.51%
:	1	0.51%
en effet	1	0.51%
sans quoi	1	0.51%
compte tenu que	1	0.51%
du fait que	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.

Dialogue and Discourse 1 (2010) 1-100
doi: 10.5087/dad.2010.001

Received 2/09; First letter 7/09; Accepted 7/09;
Final revision 12/09; Published online 2/10

A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations

Merel C.J. Scholman M.C.J.SCHOLMAN@COLI.UNI-SAARLAND.DE
Utrecht Institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK
Utrecht, The Netherlands

Jacqueline Evers-Vermeul J.EVERS@UU.NL
Utrecht Institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK
Utrecht, The Netherlands

Ted J.M. Sanders T.J.M.SANDERS@UU.NL
Utrecht Institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK
Utrecht, The Netherlands

Editor: Barbara Di Eugenio

Abstract

Over the last decade, annotating coherence relations has gained increasing interest of the linguistics research community. Often, trained linguists are employed for discourse annotation tasks. In this article, we investigate whether non-trained, non-expert annotators are capable of annotating coherence relations. For this goal, substitution and paraphrase tests are introduced that guide annotators during the process, and a systematic, step-wise annotation scheme is proposed. This annotation scheme is based on the cognitive approach to coherence relations (Sanders et al., 1992, 1993), which consists of a taxonomy of coherence relations in terms of four cognitive primitives. The reliability of this annotation scheme is tested in an annotation experiment with 40 non-trained, non-expert annotators. The results show that two of the four primitives, *polarity* and *order of the segments*, can be applied reliably by non-trained annotators. The other two primitives, *basic operation* and *source of coherence*, are more problematic. Participants using an explicit instruction with substitution and paraphrase tests show higher agreement on the primitives than participants using an implicit instruction without such tests. We identify categories on which the annotators disagree and propose adaptations to the manual and instructions for future studies. It is concluded that non-trained, non-expert annotators can be employed for discourse annotation, that a step-wise approach to coherence relations based on cognitively plausible principles is a promising method for annotating discourse, and that text-linguistic tests can guide annotators during the annotation process.

Keywords: discourse annotation, corpora, coherence relations, interrater reliability

1 The complexity of discourse annotation

The advent of linguistic corpora has had a large impact on the field of linguistics. By gathering and annotating large-scale collections of texts, researchers have gained new possibilities for analyzing language. Corpora can be used to, for example, investigate characteristics associated with the use of a language feature, examine the realizations of a particular function of language, characterize a variety of languages, and map occurrences of a feature through entire texts (Conrad, 2002).

The focus area of corpora has mainly been on lexical, syntactic and semantic characteristics of language. Existing corpora often lack annotations on the discourse level (Carlson, Marcu & Okurowski, 2003; Versley & Gastel, 2012). However, the notion of "discourse", and more specifically the coherence relations between parts of discourse such as *cause-consequence* and *claim-argument,* has become increasingly important in linguistics. This has led to the international tendency in the last decade to create discourse-annotated corpora. Leading examples are the Penn Discourse Treebank (Prasad et al., 2008), the Rhetorical Structure Theory (RST) Treebank (Carlson et al., 2003), the Segmented Discourse Representation Theory (SDRT; Asher & Lascarides, 2003) and the Potsdam Commentary Corpus (Stede, 2004).

While discourse annotation guidelines generally agree on the idea of relations between discourse segments, they differ in other important aspects, such as which features of a relation are analyzed and the types of relations that are distinguished. Some proposals present sets of approximately 20 relations, such as the one developed by Mann and Thompson (1988) and the set of core discourse relations in the ISO project (Prasad & Bunt, 2015), others of only two relations (Grosz & Sidner, 1986). The PDTB contains a three-tiered hierarchical classification of 43 sense tags (Prasad et al., 2008), and the annotation scheme used for the RST Treebank distinguishes 78 relations that can be partitioned in 16 classes (Carlson et al., 2003). The Relational Discourse Analysis (RDA) corpus (Moser, Moore & Glendening, 1996), which is based on RST and the theory proposed by Grosz and Sidner (1986), distinguishes 29 relations on an intentional and an informational level. Hence, it is not clear which and how many categories or classes of relations (for example, *contingency*, *causal,* or *informational*) and end labels (for example, *result, volitional cause*, and *cause-consequence* are all labels for causal relations) are needed to adequately describe and distinguish coherence relations. One thing that is clear is that annotation has proven to be a difficult task, which is reflected in low inter-annotator agreement scores (Poesio & Artstein, 2005).

In current proposals, the developers often make use of two solutions to strive for sufficient agreement scores: (1) employing 'experts', namely professional linguists or annotators who have received extensive training, or (2) providing the annotators with large manuals. For example, for the RST Treebank, professional language analysts with prior experience in other types of data annotation were employed. They also underwent extensive hands-on training (Carlson et al., 2003). Similarly, two linguistics graduate students were employed for the creation of the RDA corpus. These annotators were required to do readings on discourse structure, and they received multiple training sessions (Moser & Moore, 1996). Linguists have more knowledge about language and linguistic phenomena, and are therefore more sensitive to certain linguistic structures. Likewise, annotators who have received extensive training have detailed knowledge about the phenomena that they are annotating. Often, trained annotators have had the opportunity to discuss their annotations and check them with those of other annotators, which benefits the annotation quality. Another solution while striving for sufficient agreement scores is to provide annotators with large manuals that describe the annotation process in great detail. For example, the manual for the PDTB corpus consists of 97 pages (PDTB Research Group, 2007), and the manual for the RST Treebank consists of 87 pages (Carlson & Marcu, 2001), the latter including more detailed information about segmentation. These manuals contain necessary information for the annotators to be able to analyze texts reliably, but considering the length it can be assumed that annotators need time to work through them.

In order to expand the field of discourse annotation and annotate discourse relations on a larger scale, it would be easier if non-trained, non-expert annotators, such as undergraduate students in the Humanities, could be employed, and if smaller manuals could be used. Working with non-trained, non-expert annotators has the practical advantage that they are easier to come by, and it is therefore also easier to employ a larger number of annotators. Using non-trained, non-expert (also referred to as naive) annotators is not new; naive annotators have for example already been employed in anaphoric annotation (Poesio & Artstein, 2005). The benefit of employing naive annotators has also been recognized in other fields. Alonso and Mizzaro (2012) describe the advantages of using crowdsourcing for conducting different kinds of relevance evaluations: the outsourcing of tasks to a large group of people makes it possible to conduct information retrieval experiments extremely fast, with good results, and at a low cost. Moreover, Nowak and Rüger (2010) found in a multi-label image annotation experiment that the annotations of non-expert annotators were of a comparable quality to the annotations of experts. Although these studies investigated annotations in different fields than discourse analysis, their results do provide insight into the usability and reliability of non-expert annotators for discourse annotation. However, working with less-trained annotators should not affect the quality of the annotations. It therefore needs to be investigated what type of instructions naive annotators need in order to annotate reliably. The annotators in the current study are not as naive as the annotators in crowdsourcing tasks, since the annotators in this study are freshman and senior undergraduate students in the Humanities. These students usually have affinity with language, and they were at least trained in some sort of analysis of language, varying from grammatical analyses to literary analyses. We chose this type of annotators because the discourse annotation task is likely to be too complex for people who are not used to work with languages in such a conscious manner. However, the annotators in this study can be considered significantly less expert than linguistics graduate students and professional linguists.

The current study sets out to investigate whether non-trained, non-expert annotators can be employed to annotate discourse relations reliably. These types of annotators might benefit from a different annotation process. More specifically, the annotation task could be less complex for them if they could make use of a step-wise approach, in which they annotate characteristics of coherence relations one at a time (for example, deciding whether the relation is causal or additive and whether the relation is subjective or objective). In many of the current annotation proposals, annotators are required to define the coherence relation at hand by assigning an end label to it. This end label is the type of coherence relation, such as a *result*, *claim-argument*, *contrast*, or *exception*. We believe that the annotation task might become less complex if the process of defining a relation is broken up into several steps. This is explained in more detail in the next section. Additionally, it is hypothesized that several (text-)linguistic tests could help non-trained, non-expert annotators during the annotation process. Instructions containing tests that make use of connective properties and paraphrase tests could guide annotators during the interpretation of the coherence relation at hand. This is further explained in Section 3.

2 A step-wise approach to coherence relations

The discourse annotation task might become less complex if annotators can make use of a step-wise annotation approach. In many of the current proposals, annotators are required to define the relations in terms of end labels. For example, in RST annotators can choose the end label 'cause', which is used to describe a causal relation such as (1).

(1) (In addition,) its machines are typically easier to operate, so customers require less assistance from software. (Penn Discourse Treebank, fragment 1887)

Although it is not explicitly acknowledged by RST, the classification process can be broken up into several smaller steps: the coherence relation is a causal relation (rather than a temporal or additive relation), the polarity of the relation is positive (rather than negative, such as in contrastive relations), and it is an objective relation (rather than a subjective relation). These types of relations are explained in more detail in Section 4. For this relation, the fact that it is causal is quite clear, and it is therefore not expected that its classification leads to many disagreements.

However, other types of relations are more difficult. Especially for these types of relations, breaking up the classification into several smaller types might be beneficial. This is illustrated with example (2), which is an 'anti-thesis' relation according to the RST manual. The end label 'anti-thesis' is described as a specific kind of contrast in which one cannot have a positive regard for both of the situations described (Carlson & Marcu, 2001: 45).

(2) Although the legality of these sales is still an open question, the disclosure couldn't be better timed to support the position of export-control hawks in the Pentagon and the intelligence community. (Penn Discourse Treebank, fragment 2326)

The classification process of this relation can also be broken up into several smaller steps: the coherence relation is causal (rather than temporal or additive), it is negative (rather than positive), and it involves the speaker's reasoning and is therefore subjective (rather than objective). An annotation scheme that breaks up the classification of such a coherence relation into more and smaller steps might help the annotators during the process. Rather than deciding on the end label of the relation at hand, they can decide on separate aspects of relations, which will eventually lead to an end label. A step-wise approach therefore might reduce the need for intensive training and still lead to high agreement between annotators.

Another advantage of a step-wise approach is that it makes use of similarities as well as differences between coherence relations, and therefore shows links between conceptually related relations. End labels carry the risk of dividing these related relations into separate classes. This is illustrated with examples (3) and (4).

(3) Operating revenue rose 69% to $8.48 billion from $5.01 billion. But the net interest has jumped 85% to $687.7 million from $371.1 million. (PDTB Research Group, 2007: 33)
(4) (The biotechnology concern said) Spanish authorities must still clear the price for the treatment but that it expects to receive such approval by year-end. (Pitler & Nenkova, 2009: 16)

Both relations in (3) and (4) are expressed by the connective *but*, but they fall in different classes according to the PDTB tagset. Fragment (3), taken from the PDTB manual, is an example of a typical contrastive relation belonging to the class comparison. The relation in (4) is coded in the PDTB as belonging to the class expansion (Pitler & Nenkova, 2009), even though it is actually also a contrastive relation. Although the PDTB does justice to the fact that these relations differ from each other (for example, (3) is additive and (4) is causal), it disregards the fact that the relations are both negative and contrastive, and that they are therefore conceptually related. By assigning the relations end labels, the conceptual relationship between the two coherence relations is not acknowledged. In contrast, an approach that classifies relations based on a combination of characteristics does account for this: such an approach does not only show differences between relations, it also shows similarities between different relations.

In order to create an annotation scheme in which coherence relations are broken up into several characteristics, a classification of coherence relations is necessary that supports this step-wise process. The cognitive approach to coherence relations (CCR), proposed by Sanders, Spooren and Noordman (1992, 1993) is exactly such a theory in which the coherence relations are defined by their characteristics. The theory is built on the assumption that coherence relations are cognitive, psychological constructs that language users make use of when interpreting text, and not just descriptive constructs that are created by linguists. Sanders et al. (1992, 1993) believe that understanding discourse means constructing a coherent representation of that discourse. Since coherence relations play a crucial role in this representation, different relations over the same discourse will result in different representations. In line with Hobbs' (1979, 1985) and Kehler's (2002) work on coherence relations as cognitive elements of the discourse representation, Sanders et al. (1992, 1993) set out to describe the link between the structure of a discourse as a linguistic object and its cognitive representation.

Sanders et al. (1992, 1993) distinguish four cognitive primitives that they claim to be relevant for every coherence relation. What distinguishes these primitives from other, possibly relevant characteristics or primitives is that they all concern the additional meaning provided by the relations, namely they concern the informational surplus that the coherence relation adds to the interpretation of the discourse segments in isolation. The four cognitive primitives are: *polarity* (relations are positive or negative), *basic operation* (causal or additive), *source of coherence* (objective or subjective), and *order of the segments* (basic or non-basic order).[1] A detailed explanation of the primitives is given in Section 5.

Besides the fact that CCR allows for a step-wise approach, there is another argument for the applicability of CCR for discourse annotation: several studies have shown that these basic primitives and the categories they define are cognitively relevant. For example, acquisition studies have shown that positive relations are acquired before negative relations (Bloom et al., 1980, Spooren & Sanders, 2008), and that additive relations are acquired before causal relations (Bloom et al., 1980; Evers-Vermeul & Sanders, 2009). Processing studies show that once causal relations are acquired, they are processed faster and generate better recall compared to additive and temporal relations (Noordman & Vonk, 1998; Sanders & Noordman, 2000). Furthermore, objective causal relations are processed faster than subjective causal relations (Canestrelli, Mak & Sanders, 2013; Traxler, Bybee & Pickering, 1997; Traxler, Sanford, Aked & Moxey, 1997). And finally, studies have shown that coherence relations with a basic order of the segments are easier to process than coherence relations with a non-basic order (Noordman & De Blijzer, 2000; Noordman & Vonk, 1998). These studies indicate that the primitives and their categories affect language acquisition and processing, and are therefore cognitively relevant.

The four primitives are hypothesized to be useful for discourse annotation because they allow for a step-wise annotation process. They can be visualized in a flowchart, leading to a compressed annotation scheme that can be used to make systematical decisions. This can be beneficial to trained annotators, but perhaps non-trained, non-expert annotators are also capable of applying the cognitive categories method in discourse annotation. Although there is evidence for the relevance of the basic primitives and their categories, it has not been investigated how reliably they can be used to annotate coherence relations in everyday corpora of language use. The present study aims to explore this in an annotation experiment for which a large number of naive annotators analyze a sample corpus.

3 Instructions guiding the annotation process

The aim of the current study is to investigate whether non-trained, non-expert annotators can annotate coherence relations reliably. It also investigates whether the reliability increases when these annotators can make use of linguistic tests during the annotation process. There is a lot of variation in the types of instructions that manuals of different proposals contain. For example, the manual for the RDA corpus contains an instruction for a diagnostic test, for which the annotator has to imagine the context in which the relation occurs. The manual also explicitly mentions that annotators should *not* use discourse cues as a basis for deciding what relation occurs between the two segments (Moser, Moore & Glendening, 1996). In contrast, the PDTB manual encourages annotators to take the discourse cue into account, and supplies the annotators with information on which relations a certain connective can signal (PDTB Research Group, 2007). The RST manual

---

[1] Originally, the terms *objective* and *subjective* were defined in the literature as *semantic* and *pragmatic*, respectively (see Pander Maat & Sanders, 2000 for a discussion of this transition).

also mentions several typical discourse cues that often occur in certain types of relations (Carlson & Marcu, 2001). However, the PDTB and RST manuals do not explicitly provide the annotators with systematic tests that can be used as a diagnostic tool during the process. In one of the conditions in the current study, two types of tests are used to guide the annotator during the annotation process, namely a substitution test and a paraphrase test. Both tests will be explained consecutively.

The substitution test is based on characteristics of connectives. According to CCR, the cognitive primitives and their categories can be distinguished by the connectives they co-occur with. In other words, certain connectives signal certain types of relations, and readers or listeners can therefore use these connectives as processing instructions on how to relate the incoming information to the previous discourse segment. The idea of connectives as processing instructions was already suggested several decades ago by Ducrot (1980) and Lang (1984). Ever since Halliday's and Hasan's (1976) seminal work, it has been argued that connectives differ in the type of relation they signal. For instance, *because* signals a positive causal relation; *meanwhile* signals a positive temporal relation; and *but* signals a negative relation (Knott & Dale, 1994). Restrictions on the use of connectives can also be more subtle, because they can also hold within the same class of relations (Pander Maat & Sanders, 2000). For example, within the class of negative relations, the connectives *although* and *whereas* signal different types, namely negative causal and negative non-causal relations, respectively.

Given that connectives indicate how two segments are related to each other, they can be used by annotators to guide them while analyzing the relation at hand. This can be done by employing substitution tests, which is a method for testing semantic intuitions (Knott & Dale, 1994; Knott & Sanders, 1998, Pander Maat & Sanders, 2000). In a substitution test, the original connective is (mentally) substituted by another connective known for signaling a certain type of relation, while the meaning of the original relation is preserved. If there is no original connective present, the proposed connective is merely mentally inserted. For example, an annotator can ask himself for any given relation: can these two segments be connected by *but*? Or by *because*? Substitution tests therefore rely on the properties of the connectives, such as the polarity and degree of subjectivity they signal (Pander Maat & Sanders, 2000). If two connectives are inter-substitutable in a coherence relation, they should be classified in the same category of coherence relations (Knott & Dale, 1994).

Substitution tests are not the only type of tests that annotators can apply; paraphrase tests can also facilitate the interpretation process (Sanders, 1997; Knott & Sanders, 1998). In a paraphrase test, the annotator is instructed to choose one of two given paraphrases that best suits the coherence relation expressed in the text. The paraphrases both restate the two segments of the relation to give the meaning of the relation in another form. For example, in order to determine the order of the segments, the annotator can ask himself for a given objective causal relation: can the two segments be paraphrased as 'segment 1 presents the cause, and segment 2 presents the consequence' or 'segment 1 present the consequence, and segment 2 presents the cause'?

Substitution tests and paraphrase tests have been used widely in studies on connectives in language use in various languages and across genres and media (see among others, Degand, 2001; Degand & Pander Maat, 2003; Evers-Vermeul, 2005; Knott & Dale, 1994; Knott & Sanders, 1998; Li, Evers-Vermeul & Sanders, 2013; Pander Maat & Degand, 2001; Pit, 2007; Sanders, 1997; Sanders & Spooren, 2015; Stukker & Sanders, 2012; Stukker, Sanders & Verhagen, 2008; Zufferey, 2012), as well as in studies of connective acquisition (Evers-Vermeul & Sanders, 2009, 2011; Spooren & Sanders, 2008). In all these studies, the tests have successfully been applied by expert annotators. Whether such tests will also guide non-expert, non-trained annotators while analyzing real-life texts, is not clear yet. In the remainder of this paper, an annotation experiment is presented that set out to investigate this.

4 Method

In this experiment, 40 non-expert, non-trained subjects were asked to annotate a sample corpus making use of a step-wise approach based on CCR. The CCR approach allows for paraphrase and substitution tests to be used to determine the correct value for a primitive. These tests facilitate the decision making process, and are therefore expected to benefit the reliability of the method. In order to test whether this is true, two versions of the instruction were created: an implicit instruction and an explicit instruction. The implicit instruction relies only on the annotator's knowledge of the categories obtained from the manual. The explicit instruction relies on this knowledge, as well as on paraphrase and substitution tests. This is explained in more detail below, but first the four primitives and their categories are explained.

4.1 Material

The material for the experiment consisted of a manual, a flowchart, two versions of an instruction and a sample corpus of 36 coherence relations.

4.1.1 Manual and flowchart

Each subject received a nine-page manual for the cognitive approach to coherence relations, and a flowchart presenting all annotation choices. Participants received no additional training besides this manual. In the manual, discourse annotation and segmentation is explained, followed by an explanation of every value of each primitive. After this explanation, examples are given for every possible combination, thereby illustrating the categories. A description of the cognitive primitives and their categories, similar to the description given in the manual, can be found below.

Polarity

The first primitive in the taxonomy is polarity. This refers to the positive or negative character of a segment. A relation is positive if the propositions P and Q, expressed in the two discourse segments S1 and S2, are linked directly, without a negation of one of these propositions. A relation with a positive polarity is typically connected by connectives such as *and* or *because*. (5) is an example of a relation with a positive polarity.[2]

(5) [The stocks can decrease tremendously in value]$_{S1}$ and [thereby result in a loss for the investor.]$_{S2}$

In example (5), the second segment has a direct link to the first segment. The second segment is an expected consequence and there is no negation of the entire segment present.

A relation is negative if the negative counterpart of either P or Q functions in the relation. A relation with a negative polarity is typically connected by connectives such as *but* and *although*, as is illustrated in (6).

(6) [The biofuel is more expensive to produce,]$_{S1}$ but [by reducing the excise-tax the government makes it possible to sell the fuel for the same price.]$_{S2}$

In (6), a logical positive second segment would be that the biofuel costs more, as a consequence of the higher production costs. However, the second segment presents a denial of this expectation: the fuel is not sold at a higher price due to a reduced excise-tax. The second segment expresses not-Q, that is, the negation of the consequent of the relation. This negation causes the relation to have a negative polarity.

---

[2] All examples are (translations of fragments) taken from the Dutch DiscAn corpus.

Basic operation

The second primitive that Sanders et al. (1992, 1993) distinguish is the basic operation. This primitive concerns the operation that has to be carried out on the two discourse segments. Three types of basic operation underlie coherence relations: the causal, additive and temporal basic operation.[3] These operations were proposed because they justify the basic intuition that discourse segments are either strongly connected (causality) or weakly connected (addition and temporality). For negative relations, the additive and temporal relations have been taken together as 'non-causal' relations.

A relation is causal if an implicit relation (P → Q) can be deduced between the two discourse segments, as in (7). The brackets indicate where the first segment (S1) and the second segment (S2) start and end.

(7) [The athletics union was forced to emigrate to Belgium,]$_{S1}$ because [there was no accommodation available in the Netherlands.]$_{S2}$

In (7), the consequence is presented in S1, and the cause in S2: a lack of accommodation has led to the emigration of the athletics union.

The class of causal relations can be further divided in non-conditional (causal) and conditional relations. An example of a conditional causal relation can be seen in (8).

(8) If [you don't answer,]$_{S1}$ [I will arrest you.]$_{S2}$

In (8), the speaker confronts the listener with a condition. If the listener does not answer, there will be a consequence: he will be arrested.

A relation is additive if the segments are connected by a logical conjunction (P & Q), as in (9).

(9) [The quality of this fuel with bio component is completely similar to Shell's regular Euro 95]$_{S1}$ and [the price at the pump is the same as well.]$_{S2}$

The relation in (9) consists of two segments that both describe a fact about fuel with a bio component. The segments are in an equal relation to each other: there is no cause, consequence, condition or contrast present.

A relation is temporal if the two segments are linked by their occurrence in the real world. Temporal relations have an additive nature, but differ in that the segments contain two events that are ordered in time. (10) is an example of a temporal relation.

(10) [Next Thursday a second meeting will follow.]$_{S1}$ [The unsatisfied RET-employees will decide after this meeting if they deem it necessary to continue protesting.]$_{S2}$

Example (10) consists of two sequential (future) events. The events have an order in time: S2 follows S1.

Source of coherence

The third primitive is the source of coherence, which can be divided into two categories: objective and subjective. A relation is objective if the discourse segments are connected by their

---

[3] The original proposal did not distinguish temporality as a basic operation, but included temporal relations in the category of positive additive relations. This value was now added at the basic level to improve descriptive adequacy, and because temporal relations have been shown to be relevant in the order of acquisition (Evers-Vermeul & Sanders, 2009). Still, there is some discussion on how basic temporality is.

propositional content. In other words, both segments describe situations in the real world, as in (11). The speaker merely reports these facts, and is not actively involved in the construction of the relation.

(11) [The plaintiff received his car,]$_{S1}$ because [the advertisement was formulated ambiguously.]$_{S2}$

Relations are subjective if speakers or authors are actively engaged in the construction of these relations, either because they are reasoning, or because they perform a speech act in one or both segments. Subjective relations, such as (12), usually express the speaker's opinion, argument, claim or conclusion.

(12) [Drugs destroy people's lives,]$_{S1}$ so [drugs have to be battled judicially.]$_{S2}$

In (12), the statement in the first segment is not the cause for the second segment, but a reason that is given to support the claim in the second segment.

Order

The fourth primitive is the order of the segments. Two segments in a causal relation can be connected in a basic or a non-basic order. The order of the segments is not applicable for additive relations, as they are logically symmetrical.

A relation with a basic order has an antecedent as S1, followed by a consequent in S2, as in (13). The antecedent is the cause or the argument, and the consequent is the consequence or the claim. In a relation with a non-basic order, such as (14), the consequent precedes the antecedent.

(13) Sometimes children tease me. [But I don't reply,]$_{S1}$ that's why [they don't do it anymore.]$_{S2}$

(14) [Universities supposedly cancel subscriptions to scientific journals more often]$_{S1}$ because [there is more information available through the internet.]$_{S2}$

Flowchart

The four primitives can be represented in a flowchart, which can be used for annotating discourse and allows for a systematical, step-wise decision-making process. The entire flowchart can be seen in Figure 1. This flowchart will be explained step by step.

Starting with a discourse relation, the first step in the annotation process is determining the polarity. The category of negatives differs greatly from that of positives; therefore this step is the first one in the flowchart.

Second, the basic operation has to be decided upon. For positive relations, this can be causal, causal-conditional, additive or temporal. For negative relations, this basic operation can be divided into the categories causal and non-causal (any negative relation that is not causal). This step is taken as the second step because the remaining two steps are not applicable to every relation.

The third step is determining the source of coherence, which consists of the same two categories for all relations (objective and subjective), except for temporal and non-causal relations. Because temporal relations are made up of a description of two events that are ordered in time, this type of relation is always objective.

The final step concerns the order, which can be basic or non-basic. The order is not applicable for additive and non-causal relations, since the two segments in such relations are logically symmetric, and for temporal relations in which the segments describe events that occur simultaneously.

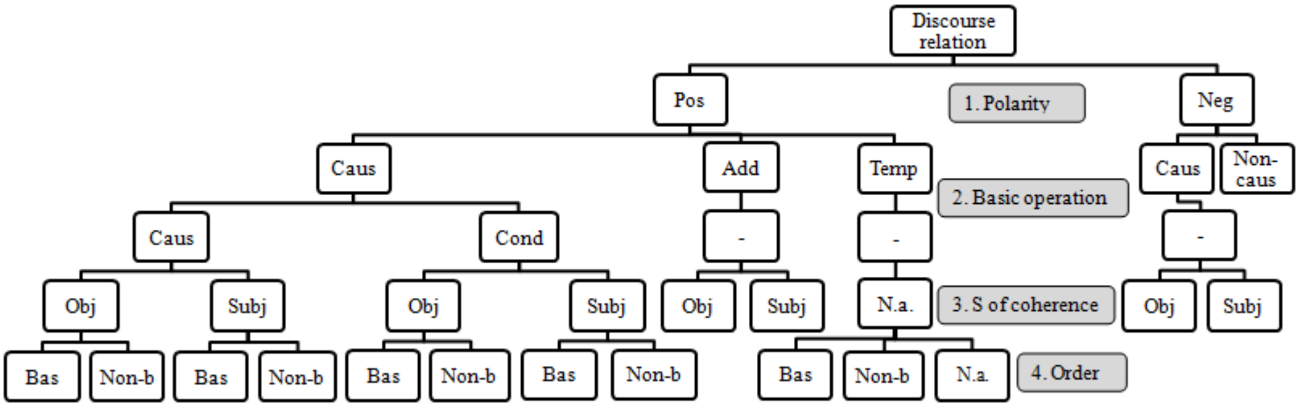


Figure 1. Flowchart of the step-wise annotation instruction.

4.1.2 Instructions

Annotators analyzed fragments using an instruction. Two experimental conditions were created in this study: one group annotated according to an implicit instruction (see Appendix A), and one according to an explicit instruction (see Appendix B), which included text-linguistic tests. Each possible answer for the steps in the instruction was preceded by a box, which participants could tick if they thought that this value was the correct answer.

The implicit instruction consists of four steps; one step for each cognitive primitive. The instruction is straightforward and relies on the annotator's knowledge of the categories. The annotator is instructed to determine the value and is reminded of any anomalies. Take, for example, step 3 of the implicit instruction (originally, this instruction is in Dutch):

3. Determine the source of coherence: is the relation **objective** or **subjective**? This does not apply to temporal or non-causal negative relations, because they do not differ in source of coherence. Therefore, for these relations tick **not applicable**.

☐ Objective

☐ Subjective

☐ Not applicable

Box 1. Fragment of the implicit instruction

The explicit instruction consists of five multileveled steps and contains two types of tests: paraphrase and substitution tests (see Section 3). Decisions for source of coherence and order are based on knowledge of the categories and paraphrase tests (Sanders, 1997; Knott & Sanders, 1998). An example of paraphrase tests for order can be seen in Box 2.

2a Can you paraphrase the relation between S1 and S2 as in option A or rather option B below?

A. The situation / fact / event in one segment causes the situation / fact / event in the other segment.
OR
B. One segment describes the reason for the claim or conclusion given in the other segment.

☐ Paraphrase A, then the source of coherence is OBJECTIVE. **Proceed to question 2b.**

☐ Paraphrase B, then the source of coherence is SUBJECTIVE. **Proceed to question 2c.**

Box 2. Fragment of the explicit instruction

In step 2a in Box 2, the annotator is given two paraphrases that can be used to determine the source of coherence of a relation. A paraphrase test was also used to determine the order of subjective relations; in that case *claim* and *reason* were used instead of *cause* and *consequence*.

In a substitution test, the annotator is first instructed to mentally take out the connective (if present in the relation), and then to replace it with different connectives. Substitution tests are used because they rely on the connective properties. In the current study, substitution tests are used in the explicit instruction for determining the polarity and the basic operation. Box 3 provides an example of a substitution test.

1. Can you use *but* to connect the segments?

☐ Yes, then the polarity is NEGATIVE. **Proceed to 1a.**

☐ No, then the polarity is POSITIVE. **Proceed to 2.**

Box 3. Fragment of the explicit instruction

In step 1, the explicit instruction guides the annotator in his choice for polarity. In this case, the annotator is instructed to substitute the original connective with the connective *but*. This type of substitution test was also used for causal relations (“Can you use *because* to connect the segments?”), conditional relations (“Can you use *if* to connect the segments?”), additive relations (“Can you use *and* to connect the segments?”) and temporal relations (“Can you use *then* to connect the segments?”).

4.1.3 Sample corpus

The sample corpus consists of 36 Dutch coherence relations with context, taken from the DiscAn corpus. The DiscAn corpus is a Dutch corpus with annotated discourse relations, which was developed using an annotation scheme based on CCR (Sanders, Vis & Broeder, 2012). This corpus currently consists of approximately 1500 fragments and includes seven subcorpora used in previous corpus-based research (see, for example, Degand, 2001; Sanders & Spooren, 2009; Stukker, 2010). These subcorpora mainly consist of newspaper articles, but also contain fragments from novels, spoken discourse, and chat fragments. The annotations that are included in the DiscAn corpus were taken from the original annotations of the seven subcorpora and supplemented if any primitives were missing. Currently, DiscAn only contains explicit relations, although several additional subcorpora containing implicit relations have been prepared for inclusion in the DiscAn corpus.

For the current experiment, both spoken and written texts are incorporated in the corpus, as well as chat fragments. The fragments were included in their original formulation, to ensure that the task resembles a real-life annotation task. The fragments were presented with the segment boundaries indicated. This was done to limit effects of segmentation.

4.2 Annotators

40 non-trained, non-expert subjects took part in this experiment and were paid for their participation. 20 subjects were freshman students and 20 subjects were senior students. All participants were students of the Faculty of Humanities at Utrecht University. None had experience with discourse analysis. To ensure that participants in this experiment had an affinity with language and text, participants were recruited from undergraduate studies in Modern Languages, Linguistics and Communication Sciences. These participants were expected to have basic meta-linguistic skills. A comparison is made between freshman and senior undergraduate students in order to investigate whether the amount of formal education in a field in Humanities had an influence on the extent to which annotators can apply a classification scheme to coherence relations.

4.3 Procedure

All materials were presented on paper. The annotators were asked to meticulously read the manual and ask questions if anything was unclear. They were also ensured that they could consult the manual and ask questions throughout the entire experiment. Questions could only concern the interpretation of a value; not the interpretation of a fragment. After the participants had read the manual and instruction, they could start annotating the sample corpus. Each fragment of the sample corpus was followed by the instructions, in which annotators could tick their choices. They were instructed to follow the steps presented in the instruction. They were allowed to annotate at their own pace, take breaks and divide the workload into two sessions. All coders annotated independently. Participants took approximately an hour and a half to read the manual and annotate all fragments.

4.4 Processing the data

Consistency is a challenge for each discourse annotation project. Although inter-annotator agreement is an important issue in the field of discourse analysis, the reliability and validity of coding is still a concern (Spooren & Degand, 2010). To deal with this problem, different statistics were calculated in the current study, namely the percentages of agreement, kappa (κ) scores, and recall, precision and F-scores. Percentages of agreement are often reported in similar studies (Artstein & Poesio, 2008). It is the simplest measure of agreement, but it does not correct for chance agreement. This measure is therefore biased in favor of dimensions with a small number of categories (Scott, 1955). Kappa scores do correct for chance agreement, and therefore show a less biased picture of the data (Carletta, 1996). When there is total agreement, κ is one. When there is no agreement besides chance agreement, κ is zero. Concerning the acceptability of a kappa score, there is no clear definition of what passes as an acceptable agreement score (Artstein & Poesio, 2008). For the current study, it was decided to follow the conventions proposed by Krippendorf (1980: 147), as reported by Carletta (1996: 252): a category with almost perfect agreement ($\kappa > 0.81$) indicates a reliable method; a category with substantial agreement ($0.61 < \kappa < 0.81$) allows for tentative conclusions to be drawn; and everything below substantial agreement ($\kappa < 0.61$) indicates that the method is not reliable enough.

Finally, recall, precision, and F-scores are included to calculate the agreement with the original annotations per value of each primitive. These measures calculate the number of true and false positives and negatives. To illustrate this, consider Table 1 (which is based on Ting, 2010).

Dialogue and Discourse 1 (2010) 1-100
doi: 10.5087/dad.2010.001

Received 2/09; First letter 7/09; Accepted 7/09;
Final revision 12/09; Published online 2/10

		Assigned class by non-trained annotators	
		Positive	*Negative*
Assigned class by expert annotator	*Positive*	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Table 1. The outcomes of classification into positive and negative classes

In Table 1, the values positive and negative can be considered to represent the actual values *positive* and *negative* for the primitive polarity, for example. True positives and true negatives are correct answers; namely when the subject agrees with the expert annotator. A false positive occurs when the subject assigns a positive polarity to an item that actually has a negative polarity. Similarly, a false negative occurs when a subject assigns a negative polarity to a coherence relation that actually has a positive polarity. Based on these outcomes, precision and recall can be calculated as follows:

Recall = True positives / Total number of actual positives assigned by the expert annotator
= TP / (TP + FN)

Precision = True positives / Total number of positives assigned by the subject
= TP / (TP + FP)

In other words, recall represents the number of times the annotators assigned a value correctly, out of all the times that the expert annotators had assigned the value. Precision shows the number of times the annotators assigned a value correctly, divided by all the times they assigned the value. Instead of two measures, these scores are often combined to provide a single, harmonic measure of agreement called the F-measure (Brants, 2000):

F-measure = (2 * Recall * Precision) / (Recall + Precision)

All three scores are reported in the current study. It is not determined what score is acceptable or unacceptable; rather, these scores are used to identify problems with specific categories of primitives, which are further discussed in Section 5.4.

5 Results

Agreement statistics were calculated for each primitive separately. First, the kappa statistics for agreement between annotators are presented. Then the agreement with the original annotations in the DiscAn corpus is shown in kappa statistics, followed by the agreement per type of instruction in percentages. The section is concluded with a more detailed analysis of agreement on separate categories in recall, precision and F-scores.

5.1 Agreement between annotators

Table 2 shows agreement between annotators for each condition separately.

Primitive	**Overall**	**First year**		**Third year**	
		Implicit instruction	*Explicit instruction*	*Implicit instruction*	*Explicit instruction*
Polarity	.73	.68	.84	.64	.78
Basic operation	.42	.33	.47	.50	.47
Source of coherence	.31	.28	.27	.39	.32
Order	.47	.34	.49	.49	.66

Table 2. Kappa statistics for each primitive in general and per condition.

Table 2 shows that the non-expert, non-trained annotators agree substantially on the categories of polarity ($\kappa = .73$). Agreement is moderate for the primitives basic operation ($\kappa = .42$) and order ($\kappa = .47$). Agreement on source of coherence is fair ($\kappa = .31$). Hence, of the four primitives, polarity yields the highest agreement and source of coherence is least agreed on. These results are in line with earlier results (Sanders et al., 1992, 1993).

When analyzed per year, most conditions show a kappa similar to the overall kappa scores. Agreement on polarity is substantial in most conditions ($.64 < \kappa < .78$), but almost perfect in the first year explicit condition ($\kappa = .84$). Agreement on basic operation is moderate in most conditions ($.45 < \kappa < .50$), but it is fair in the first year implicit condition ($\kappa = .33$). For source of coherence, agreement is fair in all conditions ($.27 < \kappa < .39$). Agreement for the primitive order is fair for the first year implicit condition ($\kappa = .34$) and substantial in the third year explicit condition ($\kappa = .66$), whereas it’s moderate in the other conditions ($\kappa = .49$).

Note that it is possible that annotators show agreement on categories that are not the correct ones according to the original annotations. In other words, they can agree on the wrong categories. The kappa statistics for agreement with original annotations in Section 5.2 will show whether participants annotated the correct categories. This will provide more insight into the quality of the instructions: how well do the instructions convey the information that annotators are supposed to know?

5.2 Agreement with original annotations

Table 3 shows the agreement with the original annotations for each condition separately.

Primitive	**Overall**	**First year**		**Third year**	
		Implicit instruction	*Explicit instruction*	*Implicit instruction*	*Explicit instruction*
Polarity	.86	.84	.91	.79	.88
Basic operation	.49	.41	.52	.48	.52
Source of coherence	.31	.31	.25	.36	.31
Order	.61	.50	.62	.59	.69

Table 3. Agreement with original annotations in kappa statistics overall and per condition.

The annotators showed almost perfect agreement with the original annotations on the primitive polarity ($\kappa = .86$). Agreement with order was substantial ($\kappa = .61$). Agreement of the annotators with the original annotations on basic operation was moderate ($\kappa = .49$) and agreement on source of coherence was fair ($\kappa = .31$). Again, the results show that polarity yields highest agreement with the original annotations and source of coherence the lowest.

When the conditions are analyzed separately, most scores of the primitives remain in the same range. All conditions show moderate agreement for the primitive basic operation ($.41 < \kappa < .52$) and fair agreement for the primitive source of coherence ($.25 < \kappa < .36$). For polarity, most conditions show almost perfect agreement ($.84 < \kappa < 91$), except for the third year implicit condition, which shows substantial agreement ($\kappa = .79$). For the primitive order, the two explicit conditions show substantial agreement ($.62 < \kappa < .69$), but only moderate agreement is found in the first year implicit condition ($\kappa = .50$) and third year implicit condition ($\kappa = .59$).

To determine whether these differences in agreement with original annotations between conditions were significant, a univariate ANOVA was performed. Results indicated a significant main effect of type of instruction on agreement with the original annotations (F (1, 5717) = 12.28; $p < .001$). A significant main effect was also found for primitive (F (3, 5717) = 228.00; $p < .001$). No main effect of undergraduate year was found (F (1, 5717) = 3.76; $p = .052$), nor any interaction effects of undergraduate year and type of instruction or primitive. Therefore, the distinction between first and third year students was not taken into account in further analyses.

5.3 Agreement per type of instruction

Table 4 shows the percentages of agreement with the original annotations per type of instruction.

Primitive	**Implicit instruction**	**Explicit instruction**
Polarity	94 (.24)	96 (.19)
Basic operation	63 (.48)	71 (.46)
Source of coherence	57 (.50)	54 (.50)
Order	70 (.46)	78 (.42)

Table 4. Percentages of agreement (and standard deviations) with the original annotations per type of instruction.

An interaction effect was found for type of instruction and primitive (F (3, 5717) = 5.50; $p = .001$). Participants using the explicit instruction showed more agreement with the original annotations on certain primitives than participants using the implicit instruction. Further analyses showed a significant difference in agreement with original annotations between the implicit and explicit instructions for the primitives polarity (t (1356.83) = -2.19; $p = .03$), basic operation (t (1427.10) = -3.33; $p = .001$) and order (t (1418.45) = 3.32; $p = .001$). The annotators using the explicit instruction showed more agreement with original annotations for these three primitives than annotators using the implicit instruction. There was no significant difference in agreement with original annotations for the primitive source of coherence between participants using the implicit instruction and participants using the explicit instruction (t (1425.57) = 1.10; $p = .27$).

5.4 Agreement on separate values per primitive

In order to examine which values of a primitive were annotated better or worse than others, recall, precision and F-scores were calculated per value, primitive and instruction. As described in Section 4.4, recall represents the number of times the annotators assigned a value correctly, divided by all the times that the expert annotator assigned the value. Precision represents the number of times the annotators assigned a value correctly, divided by all the times they assigned the value (both correctly and incorrectly). The F-score is the harmonic mean of both. Together, these three scores will provide more insight into which categories cause annotation problems, and which categories are often confused with each other. Table 5 shows the recall, precision and F-scores for the primitive polarity. According to the original annotations, there were 28 positive and eight negative relations in the sample corpus.

Value	**Implicit instruction**			**Explicit instruction**		
	Recall	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*
Positive	.95	.97	.96	.99	.97	.98
Negative	.89	.83	.85	.89	.95	.92

Table 5. Recall, precision and F-scores for the primitive polarity.

The F-scores reported in Table 5 show that the value negative was annotated correctly more often in the explicit instruction than in the implicit instruction. Hence, in the implicit condition, subjects annotated positive relations as having a negative polarity more often. This is reflected in the precision and F-scores, and suggests that the substitution test used for the explicit instruction led to higher agreement on the category negative for polarity. Overall, the value polarity was annotated well. This was already indicated by the percentages of agreement in Table 4.

Table 6 shows the recall, precision and F-scores for the primitive basic operation.[4]

Value	**Implicit instruction**			**Explicit instruction**		
	Recall	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*
Causal	.91	.63	.75	.92	.79	.85
Conditional	.29	.75	.42	.52	.75	.61
Additive	.47	.73	.58	.48	.76	.59
Temporal	.61	.43	.51	.55	.24	.34
Non-causal	.25	.65	.36	.41	.83	.55

Table 6. Recall, precision and F-scores for the primitive basic operation.

As was indicated in Section 5.3, the substitution and paraphrase tests for basic operation were helpful for the participants in the explicit condition. This finding is confirmed by the F-scores on the categories causal, conditional, and non-causal, which are higher in the explicit than in the implicit condition. Overall, however, the values temporal and non-causal are problematic. The value temporal is often mistaken for the value additive, especially in the explicit condition. This leads to low precision scores for the value temporal, and low recall scores for the value additive. These results indicate that the substitution test for the temporal relations ("Can you use *then* or *when* to connect the segments?") was misleading: annotators did not think they could use *then*, leading them to the next substitution test: "Can you use *and* to connect the segments?"

Table 6 also shows that the value non-causal was used for relations that were actually not non-causal, especially in the implicit condition. This led to lower recall scores for the value non-causal, and lower precision scores for the value causal. It should be noted here that these outcomes were based on only two coherence relations with a non-causal basic operation.

Finally, it appears that the value conditional was also applied to causal relations when it should not have been, especially in the implicit condition. Again, this should be interpreted with care, since there was only one fragment with a conditional relation in the sample corpus.

Table 7 presents the scores for the primitive source of coherence.[5]

---

[4] The sample contained 22 causal, one conditional, six additive, five temporal, and two non-causal relations.

[5] The sample contained twelve objective and seventeen subjective relations, and seven relations to which source of coherence did not apply.

Value	**Implicit instruction**			**Explicit instruction**		
	Recall	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*
Objective	.47	.67	.55	.41	.57	.48
Subjective	.79	.54	.64	.72	.55	.62
Not applicable	.44	.46	.45	.48	.44	.46

Table 7. Recall, precision and F-scores for the primitive source of coherence.

Table 7 indicates that every value of source of coherence is problematic, but especially the values objective and not applicable. The subjects often annotated subjective relations as objective relations in both the explicit and implicit conditions, as shown by the low precision score for subjective relations, and low recall score for objective relations. Also, the subjects often annotated relations for which the source of coherence does not apply as objective relations. This is reflected in the low precision score for the not applicable relations, and the low recall score for the objective relations. These results may be attributed to the step-wise aspect of the approach, which will be discussed in more detail after the recall, precision and F-scores for the primitive order are presented.[6]

Value	**Implicit instruction**			**Explicit instruction**		
	Recall	*Precision*	*F-score*	*Recall*	*Precision*	*F-score*
Basic	.55	.66	.60	.71	.55	.62
Non-basic	.80	.61	.69	.91	.78	.84
Not applicable	.76	.80	.78	.74	.93	.82

Table 8. Recall, precision and F-scores for the primitive order of the segments.

For the primitive order, the subjects in the implicit condition often coded relations with a non-basic order as relations with a basic order. This is reflected in the low recall score for the basic order, and the lower precision score for the non-basic order. Apparently, the substitution and paraphrase tests were helpful in this area, as the participants in the explicit condition obtained higher recall scores for the basic as well as the non-basic order, and higher precision scores for the non-basic category. However, in the explicit condition, the subjects still coded basic relations as having an order that is not applicable, which is reflected in their lower precision score for the basic order.

Similar to the source of coherence, the low scores for the values of order could be due to the step-wise aspect of the taxonomy. Recall that the annotators were instructed to first annotate the basic operation, and then the source of coherence and the order. If they made a mistake in the basic operation, for example they annotated additive relations as temporal relations, or vice versa, they would automatically annotate the wrong category of source of coherence and order. This is because temporal relations do not differ in their source of coherence, whereas additive relations do; and temporal relations can have different segment orders, whereas additive relations cannot. Since the results indicated that the subjects did indeed often annotate temporal relations as additive relations, it is likely that this influenced the results. In order to determine the influence of the step-wise approach on the results, the percentages of correct annotations based on the correct annotations of the previous step was calculated. In other words, the percentages of correct annotations were calculated only for those relations in which the previous step was also annotated correctly. Table 9 shows the results.

---

[6] According to the original annotations, there were ten basic and eleven non-basic relations, and fifteen fragments for which order of the segments was not applicable.

Primitive	**Implicit instruction**	**Explicit instruction**
1. Polarity	94 (N=720)	96 (N=720)
2. Basic operation (based on correct annotation of step 1)	65 (N=673)	72 (N=693)
3. Source of coherence (based on correct annotation of steps 1 and 2)	73 (N=440)	66 (N=500)
4. Order (based on correct annotation of steps 1-3)	80 (N=322)	91 (N=331)
Correct annotation of all steps	36 (N=720)	42 (N=720)

Table 9. Percentages (and actual numbers) of correct annotations for each step, based on correct annotations of previous steps (maximum N = 20 annotators per type of instruction × 36 relations = 720 annotations).

The results in Table 9 indicate that the step-wise nature of this approach has a large influence. First, it results in a relatively low number of relations that were annotated correctly for all four primitives: 36% of the implicit annotations and 42% of the explicit annotations were entirely correct. Second, the step-wise approach had a negative impact on the reliability of certain primitives. More specifically, the results indicate that the primitive source of coherence might not be as problematic as the previous results suggested. Looking at the annotations of source of coherence irrespective of the correctness of previous annotation steps, the subjects annotate this primitive correct in 57% of the relations in the implicit condition, and 54% of the relations in the explicit condition, as shown in Table 4. But when the relations in which the basic operation was incorrectly annotated are excluded, the percentages of correct annotations rise to 73% in the implicit condition and 66% in the explicit condition. In other words, 73% of the relations that were annotated correctly for their basic operation were also annotated correctly for their source of coherence in the implicit condition. These results indicate that the annotations for the different primitives are related: if the subjects annotate the basic operation incorrectly, they also annotate the source of coherence incorrectly more often than when they annotate the basic operation correctly.

A similar conclusion can be drawn for the primitive order: without taking the step-wise process into account, the subjects annotated the order correctly in 70% of the relation in the implicit condition, and 78% of the relations in the explicit condition, as shown in Table 4. However, when the step-wise approach is taken into account, the percentages of agreement rise to 80% in the implicit condition, and 91% in the explicit condition.

6 Discussion and conclusion

The research question that was formulated for this study was: are non-expert, non-trained annotators capable of annotating coherence relations by using a step-wise approach that is based on cognitively plausible primitives, and do substitution and paraphrase tests improve the quality of their annotations? In the following subsections we will address the merits and drawbacks of a step-wise approach (Section 6.1), the usefulness of substitution and paraphrase tests (6.2), and the generalizability of our approach to other annotation systems (6.3).

6.1 Step-wise approach

At a first glance, when looking at the percentages of correct annotations of all steps taken together, the step-wise approach may not seem very promising. If the naive annotators in our study would have had to come up with an end label on the basis of their choices on all four primitives, the participants in the implicit condition would have chosen a correct end-label in 36% of the relations, and the participants in the explicit condition in 42% of the relations (see Table 9). These scores are lower than the results reported in previous studies with expert annotators, who received intensive training before and during the annotation process. A study for agreement using RST showed a kappa ranging from .6 – 1.0 (Carlson et al., 2003) and a study using the Penn Discourse Treebank annotation scheme resulted in percentages of agreement ranging from 59.6% – 95.7% (Miltsakaki et al., 2004). Al-Saif and Markert (2010) report a kappa value of .57 for their PDTB-inspired scheme for Arabic and a study on the Dutch RST corpus resulted in a kappa score of .57 as well (Van der Vliet et al., 2011).

However, if we look at the outcomes of the individual primitives, the step-wise approach does show potential, as these results are comparable to the scores in the aforementioned studies with expert annotators. In our study, percentages of agreement ranged from 54% (for source of coherence) to 96% (for polarity), and the kappa statistic for the explicit condition averaged over the four primitives is .59. Given that the annotators were not trained in discourse annotation and only received a nine-page manual and instructions varying from one to three pages, this amount of agreement is promising.

For polarity, the reliability was satisfactory: the annotators frequently agreed on this value, with each other as well as with the original annotations. On the basis of the agreement with the original annotations, we can also draw tentative conclusions for the primitive order of the segments, although the agreement among the forty annotators was moderate. For the other two primitives, basic operation and source of coherence, there is room for improvement, as we did not find adequate agreement among annotators nor between the naive annotators and the original annotations. We will discuss these two primitives in turn.

The primitive basic operation only yielded moderate agreement. In particular, the results showed low agreement with original annotations on the categories *temporal* and *non-causal*. The category *temporal* was often mistaken for the category *additive*, especially in the explicit condition. This indicates that the substitution test for temporal relations ("Can you use *then/when* to connect the segments?") was misleading (see Section 6.2 for a more extensive discussion of this issue). However, since the agreement on the category *temporal* in the implicit condition was also not acceptable, it can be concluded that the manual did not provide enough information to clarify this concept. After completing the experiment, several subjects declared that the distinction between *additive* and *temporal* was not entirely clear. If more annotators experienced this, they might have employed different definitions for basic operation and the categories *additive* and *temporal*, leading to different annotations. A similar study with clearer instructions and different substitution tests could shed light on the specific issue of temporal relations.

Regarding non-causal relations, the results showed that the annotators frequently analyzed *causal* relations as *non-causal* or *conditional*, especially in the implicit condition. The confusion with non-causal relations implies that annotators especially ran into problems with negative causal relations, since the non-causal category only occurs in relations with a negative polarity. Negative causal relations are known to be more complex than, for example, positive causal and negative additive relations (Evers-Vermeul & Sanders, 2009). This suggests that naive annotators might need some additional instruction on the interpretation of causal relations with a negative polarity. As the discussion in Section 6.2 will show, substitution tests can be part of this additional instruction, as these reduce the number of mistakes with the (negative) causal category.

The second primitive for which agreement was not high enough was the source of coherence: the recall, precision, and F-scores showed that every category of this primitive seems to be problematic. However, an investigation of the influence of the step-wise nature of the approach indicated that the relatively low reliability of the primitive source of coherence is at least partially related to problems with the primitive basic operation. When subjects annotated the basic operation correctly, they also showed greater reliability for the source of coherence, with percentages of correct annotations rising to 73% and 66% for the implicit respectively explicit condition (see Table 9). It is therefore likely that the reliability of source of coherence will increase if the annotators have a better understanding of the basic operation.

Our findings suggest that a step-wise approach can be applied by naive annotators, but that the reliability of this approach still can and should be improved. In the current study, we implemented a hierarchical version of the step-wise approach: participants had to first code polarity, and then basic operation, source of coherence and order of the segments respectively. We thought this would help these naive annotators: as the flow chart in Figure 1 illustrates, specific options are ruled out once annotators have made certain decisions. For example, if annotators selected the value *negative* for the primitive polarity, they would only have a choice between *causal* and *non-causal* relations, and could not select the categories *additive* and *temporal* anymore, as these were grouped together in the non-causal category. Similarly, if they wrongly marked a relation as temporal, they did not have the option anymore to indicate whether the relation was objective or subjective. However, results showed that wrong choices on earlier primitives (also) negatively influenced choices on the following primitive(s), as reliability scores went up if only relations were taken into account that were annotated correctly during previous steps. It is worthwhile to explore whether the step-wise approach can be applied presenting the primitives independently of each other, that is without organizing the steps in a hierarchical way.

6.2 Substitution and paraphrase tests

The current experiment also tested the potential benefits of substitution and paraphrase tests. It was expected that annotators using the explicit instruction – with such tests – would show more agreement than annotators using the implicit instruction without such tests. The results confirmed this hypothesis for the primitives polarity, basic operation and order. No significant differences between the two conditions were found for the primitive source of coherence. These results indicate that the paraphrase and substitution tests indeed guide the annotators in interpreting the relation, except for the paraphrase tests used for source of coherence.

The substitution test used for polarity ("Can you use *but* to connect the segments?") increased the kappa score from .81 to .89, and resulted in higher precision and F-scores for the negative relations.

The substitution tests for basic operation ("Can you use *because / although / whereas / if / then / and* to connect the segments?") increased its kappa score from .45 to .53. More precisely, the substitution tests resulted in higher F-scores for causal, conditional, and non-causal relations: the results in Section 5.4 indicated that participants in the explicit condition less frequently classified causal relations as non-causal than participants in the implicit condition. This indicates that the substitution test for this distinction ("Can you use *although* or *whereas* to connect the segments?") led to higher agreement. A similar result was found for conditional relations: there was more agreement on this value in the explicit condition than in the implicit condition, although it should be noted that both conditional and non-causal relations were relatively infrequent in the sample corpus.

The recall, precision and F-scores showed that the substitution test for temporal relations led to more disagreement. In hindsight, this test ("Can you use *then/when* to connect the segments?") might indeed have been problematic when applied to specific relations participants had to annotate. Several of the temporal relations already included another temporal marker, which made it harder to use *then* or *when* for signalling the temporal relation between the segments. For example, the coherence relation in (15) contains the temporal markers *het komend jaar* 'next year' in S1, and *vervolgens* 'subsequently' and *tot 2010* 'till 2010' in S2. Here, the annotators should have removed *vervolgens* 'subsequently' in order to be able to apply the substitution test. This indeed opens up the possibility to insert *dan* 'then', but if the naive annotators in the current study did not recognize *vervolgens* as a connective, they might have failed to do so.

(15) De intercity zoals we die nu kennen wordt afgeschaft; de intercity nieuwe stijl stopt op meer stations en lijkt op de huidige sneltrein. Daardoor kan de reistijd langer worden. [Er zullen het komend jaar zeven stations bijkomen.]$_{S1}$ [Vervolgens worden tot 2010 in totaal 15 nieuwe stations geopend.]$_{S2}$
‘The intercity as we know it now will be abolished; the intercity new style stops at more stations and looks like the current express train. As a result travel time will increase. [Next year seven stations will be added.]$_{S1}$ [Subsequently till 2010 a total of 15 new stations will be opened.’]$_{S2}$

The paraphrase test used for order of the segments (“Are S1 and S2 ordered as ‘S1 is the cause, S2 is the consequence’ OR ‘S2 is the cause, S1 is the consequence’?”) worked better: it increased the kappa score from .54 to .65, and especially improved the recall, precision, and F-scores of non-basic relations. Only the paraphrase test used for source of coherence (“Can you paraphrase the relation between S1 and S2 as ‘the fact in one segment causes the fact in the other segment’ OR ‘one segment expresses the reason for claiming something in the other segment’?”) did not significantly improve the amount of agreement. Taken together, these results indicate that substitution and paraphrase tests can be beneficial, especially to help annotators identify negative from positive relations and causal from additive relations. It is likely that a step-wise approach will yield more agreement if the explanation of certain concepts, such as temporality and subjectivity, when the manual is adapted, and the substitution and paraphrase tests for these concepts are adjusted.

6.3 Generalizability of the approach

At this point we would like to emphasize again that this study was a first investigation into the viability of a step-wise approach and employing naive annotators for discourse annotation. Many participants were not even acquainted with the notions of discourse and coherence, although all of them were undergraduate students in the Humanities. Their coding of the primitives polarity and order yielded considerable amounts of agreement, but source of coherence and basic operation were shown to be problematic.

The step-wise approach was designed to test relatively naive annotators’ potential for performing a discourse annotation task. This does not mean, however, that this approach is restricted to this type of annotators or to the CCR. The question arises whether a step-wise approach might be useful for other types of annotators and applicable to other annotation systems as well. Our answer to this question is ‘yes’, given that the step-wise approach yielded satisfactory amounts of agreement for polarity and order, and given that the problematic scores for the primitives source of coherence and basic operation were at least partially due to the hierarchical implementation of the step-wise approach and to problems with specific substitution tests. If naive annotators achieve agreement scores on individual primitives that are similar to results from expert annotations of end labels, this makes us wonder what expert annotators like linguists would do if they were provided with the same materials. A follow-up experiment with expert annotators using a step-wise approach might give insight into the specific facets of discourse annotation that give rise to low interrater reliability scores. Additionally, it could be tested what a step-wise approach yields if it is applied to other annotation systems.

Future research might also reveal how much training exactly is needed for various aspects of discourse annotation. Note that more experience in the field of Humanities was not helpful for the annotators in our study, given that there were no significant differences in the agreement scores between first and third year undergraduate students. The current study showed relatively low agreement scores for source of coherence, a primitive that is known for being difficult to determine in everyday texts, even for trained annotators. Previous studies have discussed this already (see Stukker & Sanders, 2012, for a recent overview). Hence, it is possible that this primitive is too complicated to be annotated reliably by non-expert, non-trained annotators.

However, it should be noted that the participants used in this study still managed to reach fair agreement on this primitive without any form of training. The only source of information that was available to the non-trained, non-expert annotators employed in this experiment was two paragraphs in the manual and a paraphrase test in the explicit condition. It needs to be investigated whether a slightly more extensive training of the annotators would improve the reliability of source of coherence, or whether this primitive is better left to experts in the field. This future study could follow suggestions by Spooren and Degand (2010) by investigating whether non-expert annotators can reach higher agreement if they are able to see and possibly discuss the correct annotations of – for example – the first fifteen fragments after they have annotated them. Alternatively, naive annotators might be given just one of the four primitives, which they would have to apply to more relations. This might make these annotators more experienced as they continue to code more coherence relations.

Similarly, the use of substitution and paraphrase tests need not be restricted to the cognitive approach to coherence relations advocated by Sanders et al. (1992, 1993), and applied here. Other annotation systems, such as PDTB and RST, might be supplemented with substitution and paraphrase tests as well. Given that current interrater reliabilities still leave room for improvement, it seems an attractive option to streamline the annotation schemas for these other approaches, and test whether this leads to similar conclusions.

For now, a first investigation into the usability of non-trained, non-expert annotators in discourse annotation has shown that they can yield considerable amounts of agreement in discourse annotation tasks. Analyzing coherence relations is a difficult task, even with extensive training and experience. Yet non-trained, non-expert annotators using a step-wise approach based on cognitively plausible primitives manage to reach moderate to substantial agreement with little instructions. This indicates that a systematic, step-wise annotation process can decrease the complexity of the annotation task. Moreover, it has been shown that an explicit instruction that includes substitution and paraphrase tests benefits annotator agreement. More extensive studies should be conducted to be able to further investigate the extent to which various annotators are able to reliably annotate coherence relations, but the results from the current study can be taken as a clue to the viability of such an approach.

Acknowledgements

This research was partly funded by a CLARIN-NL grant awarded to Ted Sanders for the DiscAn-project, and is also based on Merel Scholman's BA-thesis. We are grateful to Kirsten Vis, Daan Broeder and Sandrine Zufferey for their valuable input. Earlier versions of this paper were presented at the TextLink meeting in Louvain-la-Neuve (2015), Clarin-NL meetings in Utrecht and Soesterberg (2013, 2014), the VIOT conference in Leuven (2014), and the IPrA conference in Antwerp (2015). We thank the colleagues present at these meetings for the inspiring discussions we had. Finally, we thank three anonymous reviewers and the associate editor of this journal for valuable feedback.

Appendix A : Implicit instruction (Translated to English, originally in Dutch)

Fragment 1:
The amount of biocomponent in Shell's Euro 95 is in accordance with the guidelines of Secretary of State Van Geel as announced on Prinsjesdag. (VROM) For logistical reasons the biocomponent is mixed on one of Shell's depositories. This means that the percentage of biocomponent in the Euro 95 per district can differ. [The Euro 95 with biocomponent has the same high quality as the regular Euro 95 from Shell] [and clients can alternate between the two without doubt.] Worldwide Shell is active on multiple fronts in the area of biofuels.

1. Determine the polarity: is the relation **positive** or **negative**?
   ☐ Positive
   ☐ Negative

2. Determine the basic operation: is the relation **causal**, **additive**, **temporal** or, in the case of negative relations, **non-causal**? If the relation is causal, is it formulated **conditionally**?
   ☐ Causal
   ☐ Additive
   ☐ Temporal
   ☐ Non-causal
   ☐ Causal-conditional

3. Determine the source of coherence: is the relation **objective** or **subjective**? This does not apply to temporal or non-causal negative relations, because they do not differ in source of coherence. Therefore, for these relations tick **not applicable**.
   ☐ Objective
   ☐ Subjective
   ☐ Not applicable

4. Determine the order: is the order of the segments **basic** of **non-basic**? This does not apply for additive and negative relations, because they do not differ in order. Therefore, for these relations tick **not applicable**.
   ☐ Basic
   ☐ Non-basic
   ☐ Not applicable

Appendix B: Explicit instruction (Translated to English, originally in Dutch)

Fragment 1:
The amount of biocomponent in Shell's Euro 95 is in accordance with the guidelines of Secretary of State Van Geel as announced on Prinsjesdag. (VROM) For logistical reasons the biocomponent is mixed on one of Shell's depositories. This means that the percentage of biocomponent in the Euro 95 per district can differ. [The Euro 95 with biocomponent has the same high quality as the regular Euro 95 from Shell] [and clients can alternate between the two without doubt.] Worldwide Shell is active on multiple fronts in the area of biofuels.

---

0 If the relation contains a connective, take this out of the relation (mentally). Do take the original connective into account during your interpretation, so that the meaning of the relation does not change. If a relation contains multiple connectives, such as '[I am tired, *and therefore* I am going to bed early]', then take both connectives out of the relation.

---

1 Can you use *but* to connect the segments?

☐ Yes, then the polarity is NEGATIVE and the relation belongs to the class of negatives. **Proceed to question 1a.**
☐ No, then the polarity is POSITIVE. **Continue to question 2.**

1a Which of the two connectives best expresses the relation: *although* or *whereas*?

☐ *Although*, then the basic operation is CAUSAL. This relation does not have an order. **Proceed to question 1b.**
☐ *Whereas*, then the basic operation is NON-CAUSAL. This relation does not have a source of coherence or an order. **You've finished analyzing this relation.**

1b Can you paraphrase the relation between S1 and S2 as option A or option B?

A. One segment describes a situation / fact / event which occurred despite the situation / fact / event in the other segment.
OR
B. One segment describes a conclusion / claim, despite the situation / fact / event that is described in the other segment.

☐ Paraphrase A, then the source of coherence is OBJECTIVE. **You've finished analyzing this relation.**
☐ Paraphrase B, then the source of coherence is SUBJECTIVE. **You've finished analyzing this relation.**

---

2 Can you use *because* or *if* to connect the segments?

☐ *Because,* then the basic operation is CAUSAL. **Proceed to question 2a.**
☐ *If,* then the basic operation is CONDITIONAL. **Proceed to question 2a.**
☐ If neither can be used, NOT APPLICABLE, **proceed to question 3.**

2a Can you paraphrase the relation between S1 and S2 as in option A or rather option B below?

A. The situation / fact / event in one segment causes the situation / fact / event in the other segment.
OR
B. One segment describes the reason for the claim or conclusion given in the other segment.

☐ Paraphrase A, then the source of coherence is OBJECTIVE. **Proceed to question 2b.**
☐ Paraphrase B, then the source of coherence is SUBJECTIVE. **Proceed to question 2c.**

2b Can the order of the segments be described as option A or option B?

A. S1 is the cause, S2 is the consequence.

OR

B. S1 is the consequence, S2 is the cause.

☐ Paraphrase A, then the relation has a BASIC order. **You've finished analyzing this relation.**

☐ Paraphrase B, then the relation has a NON-BASIC order. **You've finished analyzing this relation.**

2c Can the order of the segments be described as option A or option B?

A. S1 describes the reason / argument, S2 describes the claim / conclusion.

OR

B. S1 describes the claim / conclusion, S2 describes the reason / argument.

☐ Paraphrase A, then the relation has a BASIC order. **You've finished analyzing this relation.**

☐ Paraphrase B, then the relation has a NON-BASIC order. **You've finished analyzing this relation.**

---

3 Can you use *then* or *when* to connect the segments?

☐ Yes, then the basic operation is TEMPORAL. These relations do not have a source of coherence. **Proceed to question 3a.**

☐ No**,** then **proceed to question 4.**

3a Are S1 and S2 chronologically order in time, anti-chronologically or do they happen simultaneously?

☐ Chronologically, then the order is BASIC. **You've finished analyzing this relation.**

☐ Anti-chronologically, then the order is NON-BASIC. **You've finished analyzing this relation.**

☐ Simultaneously, then the order is NOT APPLICABLE. **You've finished analyzing this relation.**

---

4 Can you use *and* to connect the segments?

☐ Yes, then the basic operation is ADDITIVE. These relations do not differ in order. **Proceed to question 4a.**

☐ No, then **start again from question 1** and choose the most fitting connective.

4a Can you paraphrase the relation between S1 and S2 as option A or option B?

A. Both segments describe a situation / fact / event.

OR

B. One or both segments describe an opinion / claim / conclusion.

☐ Paraphrase A, then the source of coherence is OBJECTIVE. **You've finished analyzing this relation.**

☐ Paraphrase B, then the source of coherence is SUBJECTIVE. **You've finished analyzing this relation.**

References

Omar Alonso and Stefano Mizzaro (2012). Using crowdsourcing for TREC relevance assessment. *Information Processing and Management, 48*(6): 1053-1066.

Amal Al-Saif and Katja Markert (2010). The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC 2010)*: 2046-2053, Malta.Ron Artstein and Massimo Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4): 555-596.

Nicholas Asher and Alex Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Lois Bloom, Margaret Lahey, Lois Hood, Karin Lifter and Kathleen Fiess (1980). Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language,* 7(2): 235-261.

Thorsten Brants (2000). Inter-annotator agreement for a German newspaper corpus. *Proceedings of the Sixth Conference on Applied Natural Language Processing (LREC)*. Seattle, WA.

Anneloes R. Canestrelli, Willem M. Mak and Ted J.M. Sanders (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye-movements. *Language and Cognitive Processes*, 28(9): 1394-1413.

Jean Carletta (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics,* 22(2): 249-254.

Lynn Carlson and Daniel Marcu (2001). *Discourse Tagging Reference Manual.* Available online via http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

Lynn Carlson, Daniel Marcu and Mary E. Okurowski (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*: 85-112. Dordrecht: Kluwer Academic Publishers.

Susan Conrad (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22: 75-95.

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. van de Weijer (eds.), *Usage-based Approaches to Dutch*: 175-199. Utrecht: LOT.

Liesbeth Degand (2001). *Form and Function of Causation: A Theoretical and Empirical Investigation of Causal Constructions in Dutch.* Leuven: Peeters.

Oswald Ducrot (1980). Essai d'application: MAIS - les allusions à l'énonciation – délocutifs, performatifs, discours indirect. In: H. Parret (ed.), *Le langage en context: Etudes philosophiques et linguistiques de pragmatique :* 487-575. Amsterdam: John Benjamins.

Jacqueline Evers-Vermeul (2005). *The development of Dutch connectives; change and acquisition as windows on form-function relations*. Ph.D. dissertation. Utrecht: LOT. Available online via http://www.lotpublications.nl/Documents/110_fulltext.pdf.

Jacqueline Evers-Vermeul and Ted J.M. Sanders (2009). The emergence of Dutch connectives: how cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language, 36*(4): 829-854.

Barbara J. Grosz and Candace L. Sidner (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3): 175-204.

Michael A.K. Halliday and Ruqaiya Hasan (1976). *Cohesion in English.* London: Longman.

Jerry R. Hobbs (1979). Coherence and coreference. *Cognitive Science,* 3(1): 67-90.

Jerry R. Hobbs (1985). *On the Coherence and Structure of Discourse*. CSLI Center for the Study of Language and Information, Stanford University.

Andrew Kehler (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, CA: CSLI Publications.

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

Alistair Knott and Ted J.M. Sanders (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30: 135-175.

Klaus Krippendorff (1980). *Content Analysis: An Introduction to Its Methodology.* Beverly Hills, CA: Sage.

Ewald Lang (1984). *The semantics of coordination*. Amsterdam: John Benjamins.

Fang Li, Jacqueline Evers-Vermeul and Ted J.M. Sanders (2013). Subjectivity and result marking in Mandarin: A corpus-based investigation. *Chinese Language and Discourse*, 4(1): 74-119.

William C. Mann and Sandra A. Thompson (1988). Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3): 243-281.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2004). Annotating discourse connectives and their arguments. *Proceedings of the Frontiers in Corpus Annotation 2004 NAACL/HLT Conference Workshop,* Boston.

Megan Moser and Johanna D. Moore (1996). *On the Correlation of Cues with Discourse Structure: Results from a Corpus Study.* University of Pittsburgh, Learning Research and Development Center. Available online via homepages.inf.ed.ac.uk/jmoore/papers/rda.ps.

Megan Moser, Johanna D. Moore and Erin Glendening (1996). *Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units.* University of Pittsburgh, Learning Research and Development Center. Available online via http://homepages.inf.ed.ac.uk/jmoore/papers/rda-instr.ps.

Leo G.M. Noordman and Femke de Blijzer (2000). On the processing of causal relations. In: E. Couper Kuhlen & B. Kortmann (eds.), *Cause, Condition, Concession, Contrast: Cognitive and Discourse Perspectives.* Berlin, New York: Mouton de Gruyter.

Leo G.M. Noordman and Wietske Vonk (1998): Memory-based processing in understanding causal information*,* *Discourse Processes*, 26(2-3): 191-212.

Stefanie Nowak and Stefan Rüger (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. *Proceedings of the International Conference on Multimedia Information Retrieval (MIR),* Philadelphia, USA.

Henk Pander Maat and Ted J.M. Sanders (2000). Domains of use or subjectivity? The distribution of three Dutch causal connectives explained. *Topics in English Linguistics*, 33: 57-82.

PDTB Research Group (2007). *The Penn Discourse Treebank 2.0 Annotation Manual.* Available online via http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf.

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

Emily Pitler and Ani Nenkova (2009). Using syntax to disambiguate explicit discourse connectives in text. *Proceedings of the ACL-IJCNLP 2009 Conference*: 13-16, Singapore.

Massimo Poesio and Ron Artstein (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. *Proceedings of the Workshop on Frontiers in Corpus Annotations ii: Pie in the sky: 76-83.*Rashmi Prasad and Harry Bunt (2015). Semantic relations in discourse: The current state of ISO 24617-8. In H. Bunt (ed.), *Proceedings of the 11th Joint ACL - ISO Workshop on Interoperable Semantic Annotation (isa-11)*: 80-91. Tilburg: TiCC, Tilburg University.

Rashmi Prasad, 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 of Language Resources and Evaluation (LREC 2008)*, Marrakech.

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

Ted J.M. Sanders and Leo G.M. Noordman (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes,* 29: 37-60.

Ted J.M. Sanders and Wilbert P.M.S. Spooren (2009). Causal categories in discourse: Converging evidence from language use. In: T.J.M. Sanders and E. Sweetser (eds.), *Causal categories in discourse and cognition*. Berlin: Walter de Gruyter.

Ted J.M. Sanders and Wilbert P.M.S. Spooren (2015). Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. *Linguistics, 53*(1): 53-92.

Ted J.M. Sanders, Wilbert P.M.S. Spooren and Leo G.M. Noordman (1992). Toward a taxonomy of coherence relations. *Discourse Processes,* 15: 1-35.

Ted J.M. Sanders, Wilbert P.M.S. Spooren and Leo G.M. Noordman (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics,* 4(2): 93-133.

Ted J.M. Sanders, Kirsten Vis and Daan Broeder (2012). *Project notes of CLARIN project DiscAn: Towards a Discourse Annotation system for Dutch language corpora.* Project notes. Utrecht: Utrecht University.

William A. Scott (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*(3): 321-325.

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

Wilbert P.M.S. Spooren and Ted J.M. Sanders (2008). The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics,* 40: 2003-2026.

Manfred Stede (2004). The Potsdam Commentary Corpus. *Proceedings ACL Workshop on Discourse Annotation.* Pennsylvania: ACL.

Ninke M. Stukker and Ted J.M. Sanders (2012). Subjectivity and prototype structure in causal connectives. A cross-linguistic perspective. *Journal of Pragmatics*, 44(2): 169-190.

Ninke Stukker, Ted J.M. Sanders and Arie Verhagen (2008). Causality in verbs and in discourse connectives: Converging evidence of cross-level parallels in Dutch linguistic categorization. *Journal of Pragmatics*, 40:1296-1322.

Kai-Ming Ting (2010). Precision and Recall. In: C. Sammut and G.I. Webb (eds.), *Encyclopedia of Machine Learning*. New York: Springer US.

Matthew J. Traxler, Michael D. Bybee and Martin J. Pickering (1997). Influences of connectives on language comprehension: Eye-tracking evidence for incremental interpretation. *Quarterly Journal of Experimental Psychology*, 50(3): 481-497.

Matthew J. Traxler, Anthony J. Sanford, Joy P. Aked and Linda M. Moxey (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1): 88-101.

Nynke van der Vliet, Ildikó Berzlanovich, Gosse Bouma, Markus Egg and Gisela Redeker (2011). Building a discourse-annotated Dutch text corpus. In: S. Dipper and H. Zinsmeister (eds.), *Proceedings Beyond Semantics (DGfS workshop*). Bochumer Linguistische Arbeitsberichte 3: 157–171.

Yannick Versley and Anna Gastel (2012). Linguistic tests for discourse relations in the TüBa-D/Z corpus of written German. *Dialogue and Discourse*, 4(2): 142-173.

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

Taming our wild data: On intercoder reliability in discourse research

Renske van Enschot, Wilbert Spooren, Antal van den Bosch, Christian Burgers, Liesbeth Degand, Jacqueline Evers-Vermeul, Florian Kunneman, Christine Liebrecht, Yvette Linders & Alfons Maes

PRELIMINARY VERSION, DO NOT QUOTE

Version: 5-1-2016

ABSTRACT

Many research questions in the field of linguistics, communication and cognition are answered by manually analyzing data collections or corpora: collections of (transcribed) spoken, written or visual communicative messages. In this kind of quantitative content analysis of discourse the coding of subjective language data often leads to disagreement among raters. In this paper we discuss causes of and solutions for disagreement problems in the analysis of discourse. We discuss the effects of three sources of difficulty in coding discourse variables. We discuss the sometimes tense relation between reliability and validity. We describe the advantages and disadvantages of using a popular formal assessment of intercoder reliability, namely Cohen's Kappa, and some of its alternatives. We suggest a number of ways to improve the reliability, such as the precise specification and carving up the coding process into smaller substeps. The paper ends with a reflection on challenges for future work in discourse analysis, with a special attention to big data and multimodal discourse.

KEYWORDS intercoder reliability, discourse, quantitative content analysis

1. Introduction

Many research questions in the field of communication, linguistics and cognition are answered by manually analyzing data collections or corpora: collections of (transcribed) spoken, written or visual communicative messages. Although many different forms of corpus analysis are used (Krippendorff, 2013), the generic base may be defined as assigning interpretative levels to particular variables in the corpus. For example, particular words or expressions can be classified as having an intensifying meaning or as being ironic or metaphoric; gestures or pictures can be classified as representational or decorative; pitch patterns can be categorized as either expressing or lacking a feeling of knowing on the part of the producer; a particular interpretation of a metaphoric poster or advertisement can be classified as ‘matching the intended meaning’ or not; a relationship between two utterances in a discourse can be labeled as semantic or pragmatic, or with one label out of a list of 25; et cetera.

In this kind of quantitative content analysis of discourse the coding of subjective language data often leads to disagreement among raters. This is partly due to coding errors and partly due to the inherent ambiguity of the language phenomena (Spooren & Degand, 2010). Disagreement can occur even when there has been an extensive training phase, even when an explicit code book is used that has been tested and adapted, even when the number of coding categories is limited, and even when experts are used instead of naive and untrained coders (Spooren & Degand, 2010). Often, several rounds of coding are necessary to reach a sufficiently high intercoder reliability statistic such as Cohen's kappa. Problems increase for the analysis of static and dynamic visuals in discourse. At the same time our publication outlets require that such a kappa is reached after only one round of coding and that only naive coders are employed. Allegedly only then the variables would be sufficiently concrete and the categorizations could be considered replicable and valid.

One may prevent low intercoder agreement results by suggesting researchers to study only clear-cut variables. They will be less stubborn than explorative variables to be detected in randomly selected data produced in uncontrolled conditions. Interrater agreement tests are superfluous when variable levels can be assigned without any interpretation noise. However, too many interesting questions in the field of human communication are not ready for such types of controlled research. Examples of such intriguing questions are: What makes a visual message metaphorical? Which linguistic or audiovisual cues can we consider to be deceptive? The second question is addressed by Hancock, Curry, Goorha and Woodworth (2007) using elicited data.

In this paper we first sketch the scope of the problem by describing different degrees of messiness in discourse data. We will then address how validity is at stake as well and describe the tension between reliability and validity. We will then move on to an overview of shortcomings of using Cohen's kappa scores as a measure of intercoder reliability -- ICR from now on --, and suggest alternative statistical ICR metrics. The paper continues with hands-on advice on how to improve ICR and concludes with a reflection on where to go from here, discussing big data and multimodality in particular.

2. Different degrees of messiness in discourse data

The quality and outcome of corpus analysis, as well as the success of interrater agreement tests, is determined by a large number of conditions. In this section, we will discuss the effects of three factors: different ways of collecting data, the ways in which coding categories are established, and the number of discourse levels that are taken into account.

First, data collections can be elicited experimentally or selected randomly in the field, for example by collecting the materials by random or representative sampling. The latter tends to be more difficult to code reliably. For instance, Arts et al. (2011) asked their participants to produce referring expressions describing entities on the screen in a controlled setting. The result was an easy to code dataset of referential expressions containing only attributes visible on the screen.

This differs sharply from the problems encountered while encoding the stylistic elements in naturally occurring newspaper and web texts, such as the ones reported in Liebrecht (2015).

Second, coding variables can be established in different ways. On the one hand, they can have a predefined theoretical position and definition. Examples are using an enchiridion - a short handbook - to categorize intensifiers on a lexical base (van Mulken & Schellens, 2012), or using an unambiguous and theory-based definition of verbal irony (Burgers, van Mulken & Schellens, 2011). On the other hand, coding variables can start from an explorative intuition and emerge gradually as the analysis proceeds (van Enschot & Donné, 2013). The coding of the latter type of data tends to be much more difficult than when the coding variables and their levels have been defined in a precise way. Agreement tests can be useful in both kinds of studies. Controlled studies require a high degree of precision and validity, and consequently only high levels of ICR outcomes are acceptable. In explorative studies, ICR scores can be used as a heuristic tool to objectify or specify individual intuitions, to try out coding level definitions or segmentation options in data collections. In this case, lower ICR rates are acceptable, although one may want to perform an additional more controlled analysis to validate the new coding system.

Third, coding categories variables can consist of a more or less closed set of levels. On the one extreme, the variable levels are a dyadic, mutually exclusive, closed set (e.g., yes-no, high-low, figurative-literal, etc.). In case of such closed variables, it is relatively easy to determine conditions with near-guaranteed agreement success: a small number of levels attached to one variable, clearly defined in terms of objective characteristics. An ICR score is hardly relevant in these cases. On the other extreme, the amount of levels is not fixed: There are different ways to intensify an utterance: from typographical elements and word parts to multiple words and syntactic constructions, or the levels are not mutually exclusive. Such variables leave room for an exploratory analysis but at the same time pose problems for ICR. Extra complicating are the cases in which the unit of analysis is underspecified, e.g., when both a lexical item and the sentence in which it is included can be categorized as intensifiers.

The above suggests that analytical data can be less or more messy, and that the degree to which our data are messy depends on the choices made by the researcher. The researcher makes decisions about the type of questions asked (e.g., does she focus on the use of the conjunction ‘because’ or on ‘epistemic cues’ in discourse), about the collection method used (e.g., does she focus on a corpus of all tweets produced in a one hour slot in The Netherlands or on the one word production results of a word recognition task), about the theoretical framework and position taken (e.g., does she start from the position that all gestures can be divided in two or three major categories or not), and about the way in which theory is translated into definitions of coding levels (e.g., does she define epistemic cues as a closed set of identifiable elements in discourse, or as an open ended class).

These choices in turn determine to a large extent the success of exercises in which different individuals are asked to do the same coding job in order to obtain a satisfactory level of interrater agreement. Data can vary. Once we agree that data can vary to a very large extent, the question is which measures can be taken to make interrater agreement exercises doable, not trivial, and successful or at least informative. We will address this issue in section 5.

3. Reliability versus validity

Quality of data is not only a matter of the reliability of the data but also of their validity. Increasing reliability means reducing the level of random (coding) errors. Increasing validity refers to a reduction at the level of systematic errors, and hence to a more accurate reflection of reality. Validity and reliability can be at odds with each other. Aiming for high intercoder reliability scores is not a guarantee for good validity and may even yield serious problems for the construct, external and internal validity of the research.

Construct validity. Construct validity refers to the question whether the coded data accurately reflect the theoretical constructs they are supposed to measure (Elmes, Kantowitz, & Roediger III, 2011, pp. 185-187). In a recent paper, Wallace (2015) contests the way in which various computational linguists have operationalized irony and sarcasm. Many operationalizations aimed to automatically code sarcasm are based on looking for specific words or word combinations, such as variants of the word *sarcasm* (*sarcastic, sarcastically*, etc.), or words that are often used to mark sarcasm (e.g., *yeah right*). Wallace (2015) argues that such automatic identification procedures based on word usage are “shallow”, because they do not take into account semantic and pragmatic information about the speaker or situation, and are likely to miss many instances of sarcasm. For instance, an utterance such as “Barack Obama is a great president” is likely to be literal when said by a supporter of the Democratic Party, and sarcastic when said by a supporter of the Tea Party. Wallace thus calls upon computational linguists to develop more advanced computational models that take into account not only syntactic aspects, but also semantic and pragmatic aspects. Thus, even when identification procedures achieve satisfactory or high levels of reliability – coding instances of “Yeah, right” in a corpus can easily be done very reliably– it is important to critically analyze whether specific examples of the variable of interest are not systematically excluded or overrepresented.

External and internal validity. External validity refers to the way in which observations can be generalized to other situations outside of the specific data investigated.The internal validity criterion invites the researcher to search for confounding factors, and is particularly relevant for corpus-based studies in which textual features in two or more (sub)corpora are compared. Both kinds of validity can be at odds with reliability. An example comes from research on the quality of the spelling in students’ writing. It makes an enormous difference whether the researcher analyzes the spelling errors in dictations, or in texts that are composed by students themselves. As van den Bergh, van Es and Spijker (2011, p. 6) point out, analyzing these text types can be done in a very reliable way (e.g., they report 100% intercoder agreement for dictations).

However, there are issues of external validity. Although dictations can give a systematic picture of what children are capable of in terms of specific spelling difficulties, children's numbers of spelling errors in dictations are not predictive of the number of spelling errors in their own writing: van den Bergh et al. report correlations between .11 and .17 (2011, p. 12). The internal validity is at stake in this analysis as well. First, the number of errors in students' own writing may (also) be determined by their proficiency in coming up with an alternative formulation, allowing them to avoid words that are difficult to spell. Second, variation in numbers of spelling errors in dictations and students' own texts may also be attributed to differences in task. If students take a dictation, their main focus is on form, not on content. However, if students write their own texts, their focus is on content, and less so on correctness of form. This confounding of factors makes it hard to compare the outcomes of studies in which different tasks are used.

The above shows that reliability is not a sufficient condition for validity. A traditional viewpoint is that reliability is at least a *necessary* condition for validity (Moss, 1994). An interesting issue is whether this viewpoint is tenable, i.e., whether we can imagine research that is valid but unreliable (Moss, 1994). The issue is even more pressing given that we often find it difficult to establish the reliability of our codings. If our reliability scores are lagging behind, can we still establish validity (cf. van Enschot & Hoeken, 2015)? Should the answer to this question be negative, we anticipate insurmountable problems for our discipline. A possible viewpoint is that the theory or the coding procedure yielding the analysis of such unreliable data is underdeveloped to such a degree that the researchers should go back to the drawing board. Alternatively, we could restrict the generalizability of our results to the limited set of our data that we *can* code reliably. Such a solution is chosen by Liebrecht (2015) for the analysis of intensified language.She reports analyses on the subset of the data on which both coders agreed. Of course, this limits the generalizability of the results. To accommodate that problem she also reports findings for the intensifiers identified by each coder separately. She only draws firm conclusions of all three sets of results point in the same direction.

4. How to deal with Cohen's kappa

In this section, we describe a selection of settings in which the most widespread metric, Cohen's Kappa, can be misleading: when the selection of annotators is alternated, the levels are in an ordering relation, or the annotation of levels is highly imbalanced. We provide an overview of metrics that can be applied alternatively to Cohen's Kappa. We will focus on the main characteristics and advantages of alternative metrics, without providing raw formulas. To apply these metrics, we recommend the package written by Andrew Hayes[1] in the framework of SPSS or SAS, or the NLTK toolkit[2] in the framework of the Python programming language.

Cohen's Kappa was proposed by Cohen (1960), and takes into account the prior chance that two annotators agree on the annotation of any level. This makes Cohen's Kappa a more realistic metric for interrater agreement than percentage agreement, which can easily give misleading insights. For instance, one study using two coding levels has a fifty percent chance that coders agree whilst another study using four coding levels yields a 25 percent agreement chance. As a result, the first study will probably yield higher ICR scores than the second study simply due to chance (Artstein & Poesio, 2008, pp. 558-559). Still, the prior chance of any two annotators to agree is not the only factor that might influence agreement. In this sense, Cohen's Kappa has its own biased focus on reliability. Artstein and Poesio (2008) describe two biases of Cohen's Kappa: the annotator bias -- the case where annotators prefer using different levels of a variable to be coded -- and the prevalence bias -- the case where one level of a variable is used much more than the other. Both lead to different and invalid estimates of the 'true' reliability.

Perreault and Leigh (1989, p. 146) state that "...different indices reflect different aspects of reliability". To give a complete insight of any interrater agreement, it is therefore valuable to show experimental outcomes with other reliability metrics in addition to, or in some cases as replacement of, Cohen's Kappa.

Cohen's Kappa presumes all items in a set to be coded by the same two annotators. The metric does not take into account settings in which the items are annotated by more than two annotators and/or different constellations of annotator sets. A better option in such a context is to use Fleiss' Kappa (Fleiss, 1971). Unlike Cohen's Kappa, which takes into account the answers of any specific *coder*, Fleiss' K is calculated based on the proportion of times that each *category* is chosen by annotators, as well as the agreement per single *item*. These two components are used to calculate the chance of agreement. By focusing on single items rather than the whole of items per annotator, Fleiss' K allows the items to be annotated by any combination of annotators, as long as the number of annotators per item remains the same.

An important property of the coding task is the scale of the variable(s). In standard form, the Cohen's Kappa metric presumes a nominal scale, in which no ordering exists between the levels. It will give wrong insights when applied to other than nominal variables, with ordinal, interval or ratio levels. With these kinds of variables, it should be penalized when two coders annotate levels that have a larger distance to one another. Metrics that do apply such a penalization are the Weighted Cohen's Kappa (Gwet, 2010, pp. 34-36) and Krippendorff's Alpha (Hayes & Krippendorff, 2007; Krippendorff, 2013). In these metrics, the agreement (or disagreement in the case of Krippendorff's Alpha) for any pair of levels is weighted by taking into account the distance between the levels, such that more distanced levels add a lot less to the

---

[1] Available at http://www.afhayes.com

[2] http://www.nltk.org/_modules/nltk/metrics/agreement.html

agreement score (or a lot more to the disagreement score). These two metrics also allow for missing data points, by taking into account the total number of annotations that were made.

Another factor that needs to be taken into account when assessing interrater agreement is data skew: the degree to which data are annotated as belonging to the same levels. Jeni, Cohn and De La Torre (2013) show that Cohen's Kappa and Krippendorff's Alpha are highly sensitive to imbalanced variables: the agreement score will drastically decrease with a bigger data skew. The reason is that the prior chance of annotators to make similar annotations is high when there is a dominant class. Consequently, the percentage agreement is subtracted by a higher number and any disagreement has a large effect. A solution is to calculate the Kappa max (Umesh, Peterson & Sauber, 1989), which returns a kappa value that is relative to the upper bound of the kappa that follows the strict chance agreement.

The Mutual F-score is another metric that can be used to conceal the influence of class imbalance. It applies to the agreement about *specific* levels rather than the overall agreement. Mutual F-score is based on F1, which is often used in Information Retrieval and Machine Learning for system evaluation (van Rijsbergen, 1979). When focusing on the annotations of one level, we can regard the annotations of a first annotator as ground truth and the annotations of a second annotator as output of the system. The F1 score evaluates the agreement of the second annotator with the first annotator in terms of recall and precision. The mutual F-score first regards annotator 1 and annotator 2 as ground truth and system output (recall), and then the other way around (precision). It can be typically calculated for the predominant level, and is especially useful for comparing the agreement for multiple datasets.

A number of factors can influence the outcome of any metric to assess interrater agreement. In section 2, we suggested that the ICR depends on the nature of the data and the research question. This suggestion would imply that an interpretation of, for example, Cohen's Kappa should be used relative to the research question. Instead of following Landis and Koch's (1977) proposal to consider kappa's .41< kappa < .60 as moderate and kappa's .61 < kappa < .80 as substantial, interpretation may vary per type of study: for hypothesis testing kappa's >.61 are required, whereas for explorative studies lower kappa's down to .41 can be sufficient. Similar suggestions can be found in Grove et al. (1981) and Spooren and Degand (2010).

In light of these examples, it is important to fully understand the mechanisms of a metric when interpreting its outcomes. Furthermore, existing heuristics to interpret the outcomes, that do not take into account the context of annotation, seem too simplistic. We advise to assess interrater agreement with several metrics, so as to achieve a more complete interpretation of the factors that are in play.

5. How to improve reliability

Researchers need hands-on advice and practical solutions to ICR problems. In this section we meet this need and suggest a number of concrete ways to improve reliability.

5.1. Specify the unit of analysis

A coder can be asked to code predefined units (words, sentences, pictures, audio fragments, etc.), consider them as indivisible, and generate one code for each unit. Again, this situation makes agreement tests easily successful. For example, Pasma (2011) uses the word as a unit of analysis that is coded as being metaphorical or not, with impressively high ICRs as a result.

However, many research topics are embedded in larger contexts (e.g., words and sentences in discourse, turns in conversations, objects in visual scenes, etc.), and the units of analysis can differ (e.g., the valence of a conversation contribution can be defined based on a complete conversation turn, on clauses in one turn, or on words in one clause). A case in point is the study by van Enschot and Hoeken (2015) in which the unit of analysis is the entire TV commercial, without any further specification; unsurprisingly, the ICRs started off low, and increased only after a second round of coding.

In case of hypothesis testing specific units of analysis are preferable. But during an explorative phase, it may well be useful to leave it to the coders to determine which unit of analysis is most appropriate. Such a first coding and agreement exercise may provide more insight into the way in which a more controlled agreement test should deal with the presentation of and instruction about units of analysis.

5.2. Make your categories independent

Agreement success is highly determined by the relation between coding levels. Two major conditions are relevant here: one is whether the coding levels are mutually exclusive or not, the other is whether coding levels are hierarchically ordered.

The same unit of analysis can have different functions or interpretations, which may result in units belonging to more than one of the coding levels (e.g., clauses can have different relations with other clauses) or in scalar levels. Obviously, agreement success is easier when coding levels are mutually exclusive.

For example, in a study of the subjectivity of adjectives preceding or following causal connectives Hendrickx and Spooren (in preparation) used the subjectivity scores on a continuous scale from 0 to 1 that were available in the so-called gold1000 lexicon of subjective adjectives (De Smedt & Daelemans, 2012). For the sake of the analysis explicit boundaries were used to create subsets of objective adjectives (subjectivity score <. 20) and subjective adjectives (subjectivity score > .70). The other adjectives were considered ambiguous and therefore excluded from the analysis.

The same holds for hierarchical variables, with which coders are first asked to determine major classes and then to subclassify units within the assigned level. An example is coding discourse relations first in terms of semantic vs. pragmatic relations, and then within the assigned level the exact relation type. Again, agreement success is endangered when coders have to apply such embedded coding tasks. Splitting up the agreement tasks is an easy solution in such situations. For instance, Zufferey and Degand (in press) report percentage agreements of three types of multilingual discourse relation annotation differing in the amount of specificity. For the least specific type of annotation, i.e., distinguishing between four types of differentiated discourse relations (*temporal, comparison, contingency, expansion,* cf. PDTB Research Group, 2007), agreement is highest, above 90%. For the second type, which subdivides, for example, contingency into conditional and causal, agreement drops to 60-72%. The third, most specific, type yields agreement percentages between 39% and 53%. Part of the disagreement concerning the second and third type is caused by disagreements concerning the first type, because decisions regarding this first type directly impact decisions that have to be made for the second and third type. An example is (1), in which the relation conveyed by *when* could arguably be either temporal or conditional. Disagreement regarding this first type automatically induces disagreement regarding the second and third type because the available decision features will be different.

(1) 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.

A way of circumventing the problem of combined variables is by asking the coders to decide for each option or level in the coding system whether it applies or not. By doing so, chances become very small that they code a case with the first level that comes to mind while ignoring other relevant levels. In the above case, disagreement regarding the more specific types 2 and 3 would not appear because some of the options have become non-applicable.

5.3 Reduce the number of coding levels

Reducing the number of levels in a coding system might improve intercoder agreement, but usually is undesirable because a reduced coding system yields less information. An obvious first check is whether all levels are really needed. For example, van Enschot and Hoeken (2015) originally had two levels in their analysis of tropes - a subcategory of rhetorical figures - in TV commercials: one in which the verbal part of the TV commercial explicitly mentioned the trope, as in *this woman is as beautiful as a rose*, and one in which the verbal part did not address this link explicitly, as in *this woman is beautiful*. Both were regarded as explicit explanations of the trope, and were therefore combined in the final phase of the analysis, resulting in higher ICRs.

An advantage of reducing the number of levels per variable is that the levels occur more frequently, which often avoids the statistical bias of unequal distribution. A case in point is the coding of the syntactic class of discourse markers (Bolly, Crible, Degand & Uygur-Distexhe, forthcoming). This class of linguistic expressions is very heterogeneous, consisting mostly of coordinate and subordinate conjunctions such as *but* and *because* and adverbials such as *well* and *actually*, but also of less frequent members such as parentheticals (*I mean, I think*) or adjectives (*first, good*). This results in a variable with many levels some of which occur infrequently. Depending on the general theory and the research question at hand, coders can question whether it is useful to keep all possible syntactic categories or whether they should group some of them. Should they maintain fine-grained distinctions such as the one between coordinate and subordinate conjunctions, or between prepositions and prepositional phrases, or should they, for instance, choose to distinguish only the most probable syntactic classes (e.g., adverbials, conjunctions and prepositional phrases) and group all other possibilities in one encompassing "other" class, or even retain only two coding choices (e.g., between 'conjunctive' and 'non-conjunctive'). Some of these options are illustrated in Table 1.

Table 1. Coding options for the variable "syntactic class" of discourse markers

Syntactic class 1	**Syntactic class 2**	**Syntactic class 3**
clause	adverbial	conjunctive
verbal phrase	conjunction	non-conjunctive
adverb	prepositional	
coord. conj.	other	
subord. conj.		
adjective		
preposition		
prep. phrase		
noun		
interjection		

Let us assume that the coders have a data set of 50 occurrences to annotate. If they choose to code according to option 1 in Table 1 (ten levels), an equal distribution of all levels would lead to a maximum of five occurrences per level. Now, knowing that adjectives or nouns used as discourse markers are very rare in English, it is highly probable that these levels will receive zero counts. This may lead to biases in the statistical analysis. Therefore, either the sample has to be increased to account for rare events, or the number of levels has to be reduced, or a statistical measure such as Kappa Max should be used that is sensitive to uneven distributions (see section 4). A simpler coding schema such as that in option 3 of Table 1, with only two levels, simplifies both the coding decisions and the statistical analysis.

5.4 Decompose the process of analysis in smaller steps

If reducing the number of levels is not possible without losing too much information, an interesting alternative is to decompose the analytical process into smaller, simpler steps, by dividing the coding system into several steps. Thus, instead of reducing the number of levels to be coded, one can simplify the coding decisions by increasing the number of coding steps, while at the same time reducing the number of levels that need to be considered during each step. The net result is that the same number of coding levels will be considered. The main advantage of this procedure is that the decision process is split up into smaller decision trees.

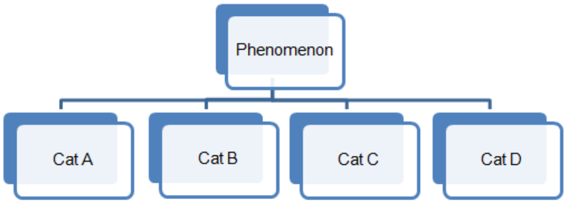


Figure 1a. Schematic process of analysis of a complex category.

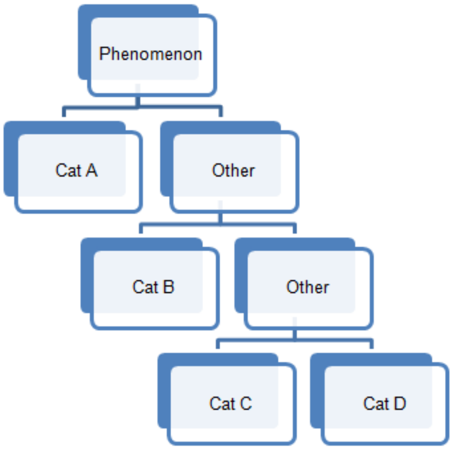


Figure 1b. Simplified version of the analysis in Figure 1a.

For example, in a research project on literary criticism, coders had to indicate what aspects (such as *style* and *structure*) and characteristics (*efficiency* or *clarity*) of novels were being evaluated by critics. For each evaluative statement they had to choose which of the fifteen listed types of characteristics applied. Characteristics varied from *efficiency*, to *emotiveness* to *religious value* (Linders, 2014). Instead of making coders choose one of these fifteen options, they could have been confronted with a number of decisions: the first step might have been deciding whether the statement was about the book itself, the effect the book had on the reader, or the book in relation to the world. The next step would then be to decide which specific subcategory within this larger category applied. If they had coded the evaluation as being a statement about the effect the novel had on the reader, they would have been asked to then choose from a limited number of characteristics that belong to the category *characteristics about the effect on the reader*, i.e. *humor*, *emotiveness* and *didactic value*. This would have narrowed down the number of options and structures they were coding.

Decomposing looks like a promising strategy. However, it also has some disadvantages. One is that splitting up analyses into smaller steps may be more time-consuming than a more straight-forward coding procedure. Another problem is that decomposing an analysis into smaller steps may lead to an inaccurate estimate of reliability. Suppose that reliability scores are calculated for the most specific levels (such as whether an evaluation in a book review is about *humor*, *emotiveness,* or *didactic value*). To obtain these scores only the cases are used in which the coders already agreed at more generally (i.e., they agreed that evaluation was about the effect the novel had on the reader). The result is that, while agreement for more specific levels is high, the picture is incomplete because this high agreement score does not reflect the difficulties at the more general coding levels. An obvious recommendation is to report agreement scores for all steps.

The conclusion is that decomposing an analysis into several smaller steps should only be used to guide coders through the analysis and, by doing this, to improve reliability, not to enhance the intercoder agreement scores without actually improving the reliability itself.

5.5 Procedural measures

A number of elements have to be taken into account to facilitate the coding the procedure, among these are, at least, the number of coders involved, the instructions given to the coders, and their training.

Consider the number of coders. Two or three coders are standard in most studies in the field of communication, linguistics and cognition (e.g. Kunneman, Liebrecht, van Mulken, & van den Bosch, 2014; Phillips & McQuarrie, 2002; Renkema, 1997; van Enschot & Donné, 2013; van Mulken & Schellens, 2006; 2012). When coding is relatively simple (a few coding levels to be assigned in well-defined units) one additional coder who recodes part of the data is considered sufficient (e.g., Mol, Krahmer, Maes, & Swerts, 2012). When data are more messy or levels more diffuse, coding by two or more coders may be useful, not only to obtain a reliable analysis, but also to gradually develop a sufficient understanding of the research phenomenon. In general, including more coders seems to be more reliable (Potter & Levine-Donnerstein, 1999), but practical considerations make two or three coders reasonable.

Optimize the coding instruction. The coding instruction also plays an important role. How specific is the instruction? In the majority of cases, a written instruction is used. The instructions differ in specificity: some only present a description of the task, others contain various examples of the phenomenon under investigation with the risk that coders are biased by those examples.

Frequently, coders get the opportunity to ask the researcher for further explanation after they have read the instruction. Then, the coders analyze the materials with the instruction in mind. A risk of this procedure is that coders gradually start leaving out certain analytical steps because of tiredness or subconscious automatic behavior. Possibly, a more clear and ordered way of instructing the coders and guiding them through the analytical process, is to not only let them read a written instruction, but also to present the analytical procedure step by step in a decisional flowchart, like Burgers et al. (2011) did for example. With the decisional flowchart at hand, coders can follow the analytical process step by step, which prevents tiredness and automaticity.

Train the coders. The final factor involving the coding procedure, is the degree to which the coders are trained. Coders can be not trained at all - if they only read the instruction by themselves -, they can practise the instruction with a single text, or they can be trained with multiple texts and feedback rounds with the researcher. The more trained coders are, the more likely it is that they are doing 'the same' during the actual analysis. However, intensive training of coders includes the risk of a coding bias: they code the studied phenomenon the same way because they learned to do this which results in a high internal validity. It is questionable, though, to what degree this affects the external validity of the study: is the research phenomenon still studied in the analysis, or did the coders learn some kind of superficial 'trick'? This relates to Potter and Levine-Donnerstein's (1999) remark that there should be projective content, i.e. some room for the coders' own interpretations. In such an approach, the goal of the coder training should be to recognize the phenomenon and to analyze the materials based on their own interpretation. Of course, room for own interpretations may have a negative influence on the interrater agreement.

6. Conclusions: where do we go from here

For scholars of discourse studies using quantitative content analysis, issues of intercoder reliability are of the highest importance, both for practical reasons (how do we convince our peers that our studies are worthwhile despite low ICRs?) and for methodological reasons. In this paper, we have shown that intercoder reliability issues are not insurmountable. Nevertheless, we see important challenges for future work.

Big data. A first issue is how to maintain insightful analyses when confronted with big data. In present-day corpus-based analyses the availability of large quantities of discourse data raises all sorts of interesting opportunities compared to small-scale analyses, but also many problems. On the plus side we have the possibility to look at our phenomena of interest in large groups of texts, consisting of a wide range of genres, which increases the richness of our analyses and the generalizability of the results. At the same time, the sheer amount of data forces us to complement our manual analyses with automatic procedures, which can lead to ill-informed decisions in comparison to human annotations.

A good example of the problems that automatic analyses can yield is provided by Vis (2011). She wanted to distinguish between words from the journalist and words from quoted sources in a wide variety of news texts from the 1950s and the 2000s. To automatize this identification she used the strategy of searching for quotation marks as the indicator of quoted sources. Although efficient, it is also a very coarse measure for quoted discourse. It neglects all forms of indirect and free indirect speech and writing, and it relies on the systematicity with which the journalists made use of quotation marks. Unsurprisingly, such an automated procedure forces the researcher to build in manual checks on the quality of the resulting analysis.

A possible improvement is the use of machine learning. Automatic classification by machine learning can be helpful for some coding tasks. For example, van den Bosch, Schuurman and Vandeghinste (2006) describe the word class or part-of-speech annotation of 50 million words. Rather than manually annotating all words, which would take a very long time, an automatic tagger was applied as a first filtering step. The tagger combined the classification of a word with a certainty score for each possible part-of-speech tag, and only the words that might be assigned to different part-of-speech tags and surpass a selected certainty threshold were extracted for manual annotation. The other words were labeled with the automatically assigned tag. This way, the number of units to be annotated manually decreases drastically. In addition, the certainty scores for different categories, along with information about common mistakes, guides the human annotators in their decision, which increases the ICR.

Automated analysis can thus help when a coding task encompasses a large dataset. Because an automated system, such as a machine learning classifier, often lacks the analytic skills of a human expert, part of the data will still need to be corrected on the basis of manual annotation. The automated system can, however, provide certainty scores for its decisions, including the certainty for categories that were not chosen. These certainty scores both help to select the data for manual annotation, typically the uncertain and ambiguous ones, and to provide the human annotators with additional context to make their decision. It should be stressed that such a procedure is especially feasible for tasks that do not require a lot of world knowledge.

Multimodal discourse. Another challenge for the near future is the annotation of multimodal discourse. Present-day discourse frequently combines different modes of communication: verbal and visual, static and dynamic. Consider a TV commercial, consisting of text as seen on screen, combined with a voice-over describing the quality of the product, a clip showing a sequence of events, plus a static depiction of the logo and the product at the end of the commercial. How do we analyze this combination of written language plus visuals plus spoken language plus the interactions between all of these features? We are dealing with the combination of codes that differ fundamentally in that the verbal code is basically non-iconic as opposed to the iconic nature of the visuals. Although the study of multimodal discourse is booming (e.g., Bateman, 2011; Bateman & Wildfeuer, 2014; Forceville & Urios-Aparisi, 2009; Jewitt, 2009; Kress, 2010;

Royce & Bowcher, 2007), attention to the ICR of this multifaceted type of discourse is scarce. An interesting initiative to use the existing knowledge of metaphor in verbal language to analyze visual metaphor are the Metaphor Lab Amsterdam's subprojects VisMet (Visual Metaphor, vismet.org) and CogVim (Cognitive Grounding of Visual Metaphor, cogvim.org). The VisMIP (Visual Metaphor Identification Procedure) seeks to identify the metaphorical elements and their relationships in a reliable way. Other initiatives are the work by Taboada et al. (2013) on rhetorical relations in multimodal documents and by Brone et al. on gesture annotation. Other than that, there is to our knowledge no methodological work that particularly addresses the reliability of coding dynamic visuals, let alone the interaction between visuals and verbals.

Naturally occurring discourse data are messy. It is no wonder researchers engaged in the quantitative corpus analysis of natural discourse sometimes feel they are in one of Augeias' stables, not cleaned in over thirty years. We hope that the suggestions made in this paper contribute to dealing with that messiness and help discourse analysts to tame their wild data.

References

ARTS, A., MAES, A., NOORDMAN, L.G.M., JANSEN, C. 2011. Overspecification in written instruction. *Linguistics*, *49*(3), 555-574.

BATEMAN, J. (2011). *Multimodality and genre. A foundation for the systematic analysis of multimodal documents.* London, New York: Palgrave Macmillan.

BATEMAN, J. & WILDFEUER, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics, 74*, 180-208.

BIBER, D., CONRAD, S., & REPPEN, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

BOLLY, C., CRIBLE, L., DEGAND, L., & UYGUR-DISTEXHE, D. (forthc.). Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers. In C. Fedriani & A. Sansó (Eds.), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*. Amsterdam: Benjamins.

BURGERS, C., VAN MULKEN, M., & SCHELLENS, P.J. (2011). Finding irony: An introduction of the Verbal Irony Procedure (VIP). *Metaphor and Symbol*, *26*(3), 186-205.

CLARIDGE, C., & WILSON, A. (2002). Style evolution in the English sermon. In T. Fanego, B. Mendez-Naya, & E. Seoane (Eds.), *Sounds, words, texts, and change: Selected* *papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000. Volume 2* (pp. 25-44). Amsterdam/Philadelphia: John Benjamins.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), pp. 37–46.

DE SMEDT, T. & DAELEMANS, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*, 2063-2067.

ELMES, D.G., KANTOWITZ, B.H., & ROEDIGER III, H.L. (2011). *Research methods in psychology* (9th Ed.). Belmont, USA: Wadsworth.

EVERS-VERMEUL, J., DEGAND, L., FAGARD, B., & MORTIER, L. (2011). Historical and comparative perspectives on subjectification: A corpus-based analysis of Dutch and French causal connectives. *Linguistics, 49*(2), 445-478.

FLEISS, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

FORCEVILLE, C. & URIOS-APARISI, E. (Eds.) (2009). *Multimodal Metaphor*. Berlin, Boston: Mouton de Gruyter.

FRASER, B. (1999). What are discourse markers? *Journal of Pragmatics, 31*(7), 931-952.

GRIES, S.Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora, 1*(2), 109-151.

GROVE, W.M., ANDREASEN, N.C., MCDONALD-SCOTT, P., KELLER, M.B., & SHAPIRO, R.W. (1981). Reliability studies of psychiatric diagnosis. Theory and practice. *Archives of General Psychiatry, 38*, pp. 408–413.

GWET, K.L. (2010). *Handbook of Inter-rater Reliability* (2nd Ed.). Gaithersburg, MD: Advanced Analytics.

HAYES, A. & KRIPPENDORFF, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89, DOI: 10.1080/19312450709336664.

HANCOCK, J. T., CURRY, L. E., GOORHA, S., & WOODWORTH, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 4*5(1), 1-23.

HENDRICKX, I. & SPOOREN, W. (In Preparation). Beyond manual analyses of discourse coherence.

HERRING, S.C., VAN REENEN, P.Th., & SCHOSLER, L. (2000). On textual parameters and older languages. In S.C. Herring, P.Th. van Reenen, & L.Schøsler (Eds.), *Textual parameters and older languages* (pp. 1-31). Amsterdam/Philadelphia: Benjamins.

JENI, L.A., COHN, J.F., & DE LA TORRE, F. (2013, September). Facing imbalanced data-- Recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference* (pp. 245-251). IEEE.

JEWITT, C. (Ed.) (2009). *The Routledge handbook of multimodal analysis*. London: Routledge.

LANDIS, J. R., & KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

KRESS, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.

KRIPPENDORFF, K. (2013). *Content analysis: An introduction to its methodology* (3rd Ed.). Thousand Oaks, CA (USA): Sage.

KUNNEMAN, F., LIEBRECHT, C., VAN MULKEN, M., & VAN DEN BOSCH, A. (2014). Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management*, dx.doi.org/10.1016/j.ipm.2014.07.006.

LEWIS, D. (2006). Discourse markers in English: A discourse-pragmatic view. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 43-59). Amsterdam: Elsevier.

LIEBRECHT, C. (2015). *Intens krachtig. Stilistische intensiveerders in evaluatieve teksten [Intensely powerful. Stylistic intensifiers in evaluative texts.]*. PhD Dissertation, Radboud University Nijmegen.

LINDERS, Y. (2014). *Met waardering gelezen. Een nieuw analyse-instrument en een kwantitatieve analyse van evaluaties in Nederlandse literaire dagbladkritiek, 1955-2005 [Read with appreciation. A new instrument of analysis and a quantitative analysis of evaluations in literary reviewsin Dutch daily newspapers]*. PhD Dissertation, Radboud University Nijmegen.

MOL, L., KRAHMER, E., MAES, A., & SWERTS, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language, 66*, 249-264.

PADILLA-WALKER, L.M., COYNE, S.M., FRASER, A.M., & STOCKDALE, L.A. (2013). Is Disney the nicest place on earth? A content analysis of prosocial behavior in animated Disney films. *Journal of Communication*, *63*(2), 393-412.

THE PDTB RESEARCH GROUP (2007). The Penn Discourse Treebank 2.0 Annotation Manual. IRCS Technical Reports Series, 99p.

PASMA, T. (2011). *Metaphor and Register Variation: The Personalization of Dutch News Discourse*. PhD Dissertation, VU University Amsterdam.

PERREAULT, W.D., Jr., & LEIGH, L.E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*, 135-148.

PHILLIPS, B.J., & MCQUARRIE, E.F. (2002). The development, change, and transformation of rhetorical style in magazine advertisements 1954-1999. *Journal of Advertising, 31*(4), 1-13.

ROYCE, T.D. & BOWCHER, W.L. (Eds.) (2007). *New directions in the analysis of multimodal discourse*. Mahwah, NJ: Lawrence Erlbaum.

SCHOLMAN, M.C.J., EVERS-VERMEUL, J., & SANDERS, T.J.M. (submitted), A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. Submitted to Dialogue & Discourse.

SPOOREN, W. & DEGAND, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, *6*(2), 241-266. DOI 10.1515/cllt.2010.009

POTTER, W.J., & LEVINE-DONNERSTEIN, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.

RENKEMA, J. (1997). Geïntensiveerd taalgebruik: een analyseschema [Intensified language: A scheme of analysis]. In H. van den Bergh, D. Janssen, N. Bertens, & M. Damen (Eds.), *Taalgebruik Ontrafeld* (pp. 495-505). Dordrecht: Foris.

TABOADA, M. & HABEL, C. (2013). Rhetorical relations in multimodal documents. *Discourse Studies, 15*(1), 59-85.

UMESH, U.N., PETERSON, R.A., & SAUBER, M.H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49*(4), 835-850.

VAN DEN BERGH, H., VAN ES, A., & SPIJKER, S. (2011). Spelling op verschillende niveaus: werkwoordspelling aan het eind van de basisschool en het einde van het voortgezet onderwijs [Spelling at different levels: Verb spelling at the end of primary education and at the end of secondary education]. *Levende Talen Tijdschrift*, *12*(1), 3-14.

VAN DEN BOSCH, A., SCHUURMAN, I., & VANDEGHINSTE, V. (2006). Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development. In *Proc. of the Fifth International Conference on Language Resources and Evaluation, LREC-2006*.

VAN ENSCHOT, R. & DONNE, L. (2013). Retorische vormen in gezondheidsvoorlichting [Rhetorical figures in health communication]. In R. Boogaert & H. Jansen (Eds.), *Studies in Taalbeheersing 4* (pp. 91-101). Assen: Van Gorcum.

VAN ENSCHOT, R., & HOEKEN, H. (2015). The occurrence and effects of verbal and visual anchoring of tropes on the perceived comprehensibility and liking of TV commercials. *Journal of Advertising*, *24*(1), 25-36. doi: 10.1080/00913367.2014.933688

VAN MULKEN, M., & SCHELLENS, P.J. (2006). Overtuigend? Een stilistische analyse van persuasieve teksten [Persuasive? A stylistic analysis of persuasive texts]. In H. Hoeken, B. Hendriks & P. J. Schellens (Red.), *Studies in Taalbeheersing 2* (Vol. 2). Assen: Van Gorcum.

VAN MULKEN, M., & SCHELLENS, P.J. (2012). Over loodzware bassen en wapperende broekspijpen. Gebruik en perceptie van taalintensiverende stijlmiddelen [On weighty basses and fluttering pant legs. Use and perception of intensifying stylistic devices]. *Tijdschrift voor Taalbeheersing*, *34*(1), 28-55.

VAN RIJSBERGEN, C.J. (1979). *Information retrieval* (2nd Ed.). London: Butterworths.

VIS, K. (2011). *Subjectivity in news discourse: A corpus linguistic analysis of informalization*. PhD Dissertation VU University Amsterdam.

WALLACE, B.C. *(*2015*)*. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, *43*(4), 467-483.

ZUFFEREY, S. & DEGAND, L. (In Press). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 1-24. Available online at http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2013-0022/cllt-2013-0022.xml.

The RST Basque TreeBank: an online search interface to check rhetorical relations

Mikel Iruskieta[1], María Jesús Aranzabe[2], Arantza Diaz de Ilarraza[3], Itziar Gonzalez-Dios[3], Mikel Lersundi[2], Oier Lopez de Lacalle[3]

[1]Department of Didactics of Language and Literature
University of the Basque Country (UPV/EHU)
Postcode 48940 – 0034.94601.7569 – Leioa – Basque Country

`mikel.iruskieta@ehu.es`

[2]Department of Basque Language and Communication (UPV/EHU)

[3]Department of Computer Science (UPV/EHU)

Abstract. *This paper introduces the first Basque discourse TreeBank annotated with rhetorical relations following Rhetorical Structure Theory. We report the main features of the corpus, such as the annotation criteria, inter-annotator agreement and harmonization procedure. We describe an online search system to check the annotation of discourse relations.*

1. Introduction

In computational linguistics discourse analysis covers a wide range of structural phenomena, such as identification of referential and relational structures. The main task when studying referential structures is correference resolution [Mitkov 2002, Recasens et al. 2010] while relational structures are related to coherence relation assignment [Asher and Lascarides 2003, Mann and Thompson 1988].

Annotated corpus are necessary in order to build advanced applications such as automatic text generation systems [Bouayad-Agha 2000], automatic summarizers [Marcu 2000b] or machine translation systems [Marcu et al. 2000]. These systems rely on different linguistic information, including the discourse level. Consequently, it is important to have a corpus which is annotated at different linguistic levels. Aforementioned systems could take advantage of the available *automatic discourse analyzers* [Marcu 2000b, Pardo et al. 2004], in order to improve their output.

There are a few works that deal with the annotation of referential structures for corpus written in languages such as English [Carlson et al. 2002, Taboada and Renkema 2011], German [Stede 2004], Dutch [van der Vliet et al. 2011], Portuguese [Pardo and Seno 2005] and Spanish [da Cunha et al. 2011a].

In the case of corpus annotation for Basque, we can find studies on referential structure [Goenaga et al. 2012, Ceberio et al. 2009] and relational structure [Iruskieta et al. 2013, Iruskieta et al. 2011]. From the linguistic point of view it is interesting to study languages with a different typology as Basque and to offer annotated corpus to the scientific community.

This work is the first RST corpus for Basque created to serve as a reference for several NLP applications for this language. The annotations follow the RST theory introduced by [Mann and Thompson 1988]. From our point of view: *i*) RST facilitates the

Anais do IV Workshop "A RST e os Estudos do Texto", páginas 40–49, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

representation of coherence in real texts, establishing relations among all the units in a tree-like structure; *ii*) RST has been applied to different languages and used for advanced applications and, *iii*) there are tools which facilitate working with RST annotated corpora: RSTTool [O'Donnell 2000] and Rhetorical DataBase [Pardo 2005]. We present the annotated corpus and we describe an online search interface to check the annotated discourse structure.

The remainder of this paper is structured as follows. Section 2 lays out the theoretical framework and Section 3 the methodology utilized to annotate the corpus. Section 4 sets out the results of the annotation and presents the online search interface. Finally, Section 5 presents the discussion and establishes directions for future work.

2. Annotation in Rhetorical Structured Theory

Rhetorical Structured Theory is a language-independent theory describing coherence between text fragments. It combines the idea of nuclearity, i.e. the importance of an individual fragment from within the discourse, with the presence of rhetorical relations (R) (hypotactic and paratactic relations) between these fragments. Hypotactic and paratactic relations connect discourse units, either a single unit (EDU) or groups of units (span). According to the theory, these relations can be paratactic (N-N) —when they establish relations between fragments that are equally important to the author (LIST, CONTRAST, DISJUNCTION, etc.)— or hypotactic (N-S) —when they connect a less-important unit with a unit the author views to be more important (ELABORATION, MEANS, PREPARATION, CONCESSION, CAUSE, RESULT, etc.). Relations are defined in light of the restrictions established between the nucleus and satellite and by describing the effect they have on the reader. A more detailed explanation of RST can be found in [Mann and Thompson 1988] and in [Mann and Taboada 2010].

Refering to the annotation process, it is well known that agreement is higher when there is training among coders. Works in which annotators did not have a training phase present a similar agreement [van der Vliet et al. 2011]. This fact is reported in the work carried out on the English language [Carlson et al. 2003]; a total of six professional annotators tagged the corpus measuring inter-annotator agreement in different texts (53 to be precise) in a pairwise manner (and in a few cases three-wise manner). There are methods for improving inter-annotator agreement: in [Carlson et al. 2003], for example, it is reported that at the beginning of the project the highest level of agreement attained between the three annotators in a small sample was a Kappa score of 0.602, while at the end of the project, after training, it was 0.755. In this project, in addition to the professional annotators, the authors also measured the agreement between two non-profesional annotators, with very different results: Kappa scores of between 0.597 and 0.792 (1918 EDUs, 30 texts).

The size of the corpus is another aspect to take into acount. We can say that, while the size of our corpus is smaller than that of the corpora found in the bibliography, the fragment tagged in a pairs was comparable as regards both size and number of annotators.

Although the delivery phase is important in annotation [Hovy 2010], it is usually forgotten. This is not the case in the RST Spanish Treebank [da Cunha et al. 2011b]. Relation extraction from a corpus is very helpful for a better understanding of the relation itself or for the study of patterns (this information will be useful to be on the design of automatic rules or as features in machine learning algorithms). In the RST Basque TreeBank the delivery phase is of great importance as we will see in the Section 4.

3. Methodological principles

Our corpus is composed by abstracts, short but well structured texts, written in Basque.[1]

Regarding coherence relations, abstracts function as independent discourse and summarize the main idea of the paper. The percentages of each relation —which are available on the web— are similar to the ones of [Pardo and Nunes 2004].

As regards relational structure, agreement between annotators was measured manually, using the evaluation system based on rhetorical relations presented in [da Cunha and Iruskieta 2010]. We decided not to use the evaluation system that assesses the tree structure [Marcu 2000a], mainly in order to avoid the shortfalls described in [Iruskieta et al. 2013]. According to these authors, span and nuclearity factors are not independent phenomena in the tree structure evaluation proposed in [Marcu 2000a], since they influence the evaluated factor of rhetorical relations. In contrast, [da Cunha and Iruskieta 2010] propose an evaluation method based on rhetorical relations where three factors are assessed: satellite unit or composition span (C), nuclear unit or attachment span (A),[2] and rhetorical relations (R).

3.1. Annotated corpus

The corpus utilized in this study is composed of abstracts from three specialized domains: medicine, terminology and science. Medical texts include the abstracts of all medical articles written in Basque in the Medical Journal of Bilbao (GMB) between 2000 and 2008. Texts related to terminology were extracted from the proceedings of the International Conference on Terminology (TERM) organized in 1997 by UZEI, while scientific articles are papers from the University of the Basque Country's Faculty of Science and Technology (ZTF) Research Conference, which took place in 2008. We have collected 60 documents that contain 15566 words (803 sentences). The created gold standard contains 1355 EDUs and 1292 Rs.

3.2. Annotators

The corpus was annotated by two linguists. The two annotators had previously annotated other linguistic levels (morphosyntax, syntax and semantics), and were familiar with RST and its annotation interface, RSTTool, but no training was provided.

3.3. Annotation phases

The process of tagging the rhetorical structure was divided into four phases. Each phase was evaluated and harmonized by a judge, in order to ensure that all annotators started each new phase from the same basic criteria. The four phases were as follows:

i) **Segmentation:** annotators were asked to divide the text into EDUs; in general, each EDU is either a subordinate clause containing a verb or an independent clause (more details in [da Cunha and Iruskieta 2010]).

---

[1] In the same sense as [Swales 1990] mentions that abstracts follows an IMRaD (*Introduction*, *Method*, *Results* and *Discussion*) structure.

[2] In multinuclear relations any of the nucleus can be considered as composition or attachment span.

ii) **Identifying the macrostructure:** before identifying the rhetorical relations, annotators were asked to identify most important part of the text or central unit (CU).

iii) **Representing the relational structure:** bearing in mind the CU, rhetorical structure was annotated in a modular and incremental way as proposed in the work by [Pardo 2005] and with the extended classification of rhetorical relations [Mann and Taboada 2010].

iv) **Annotating the signals of relations:** one annotator has tagged the signals of rhetorical relations, as proposed in [Taboada and Das Forthcoming]. The cause subset (CAUSE, RESULT and PURPOSE) was annotated by two annotators and evaluated.

The method mainly used in RST to increase annotator agreement on rhetorical relations is to establish a training phase. From our point of view this could carry a circular process between relations and their signals [Spenader and Lobanova 2009]. To provide a more reliable annotated corpus and do not fall in this circular problem, we analyzed the problems arising amongst annotators, and, in order to achieve our aim (a reference corpus annotated with relational structure), we established the criteria for annotation and we designed a manual for a judge to decide the cases of disagreement.

3.4. Results

We carried out an evaluation to assess each of the annotation steps by means of different agreement measures. This way, we calculated the agreements of segmentation (EDU), the agreement on CU identification, the agreement on rhetorical structure and the agreement on signals of the cause subset. At the rhetorical structure level we provide an analysis of the source of the disagreement, categorizing them in different types.

Segmentation (EDU). Inter-annotator agreement between annotators is 81.35%.

CUs identification. The overall mean agreement between annotators is 81.67%.[3]

Relational structure level. Based on the factors we defined —composition span (C), attachment span (A) and rhetorical relations (R)— the following types of agreements: *i*) **CAR**: agreement in composition span, attachment span and relation, *ii*) **CR**: agreement in composition span and relation, *iii*) **AR**: agreement in attachment span and relation and *iv*) **R**: agreement only in relation. Table 1 shows the agreement level obtained on the four types of measurements.

Agree	K. α	%	Gain
CAR	0.394	47.76%	-
CR	0.458	54.03%	6.27%
AR	0.431	51.17%	3.41%
R	0.561	61.47%	13.71%

Table 1. Types of agreement

Disagree	%	Disagree	%
No-Match	0.23%	Different R	13.62%
Nuclearity	6.73%	Similar R	5.88%
N/N-N/S	8.90%	MissMatch R	2.01%
Attachment	0.08%	Specificy	0.93%
Composition	0.15%	Segmentation	0.15%

Table 2. Types of disagreement

The results show how the agreement increases as the relaxation of the agreement increases too, being CAR the most demanding agreement, and R the more relaxed one.

---

[3] Agreement related to CU has been different in the three domains. The agreement is related to the number of candidates (text size) and to the enough explicit linguistic evidence which highlights the CU.

The inter-annotator agreement level [Krippendorff 2012] is moderate for relations. It must be noted that we are in the initial phase of the annotation project. Nevertheless, the results obtained are comparable to those achieved in the initial phases of the main work of rhetorical relation annotation carried out for English [Carlson et al. 2003].

On the other hand, we defined different types of disagreement, taking into account the following phenomena: *i*) **No-match**: The composition of the tree results in relations that cannot be compared. *ii*) **Nuclearity**: Different choices in nuclearity entailed discrepancy in hypotactic relations. *iii*) **N/N vs N/S**: Different choices in nuclearity entailed a paratactic/hypotactic mix-up. *iv*) **Attachment span**: Different choices in attachment span entailed a different relation. *v*) **Different R**: A relation has the same composition and attachment span, but not the same relation. *vi*) **Similar R**: Relations chosen are similar in nature. *vii*) **Mismatch R**: Relations with mismatched RST trees. *viii*) **Specificity**: The relation chosen is more specific in one annotation than in the other. *ix*) **Segmentation**: Segmentation does not match.

As shown in Table 2, although the Different R label is the main source of disagreement (13.62% of the times), one of the main disagreement comes from the choice of nuclearity: in total, 15.63% of the annotation disagree on Nuclearity or the N/N-N/S factors. The other types of disagreement (the 8.82% of the annotations) can easily be resolved explaining how the annotator understand the relations involved in Similar R, Mismatch R and Specificity labels.

Signals for rhetorical relations. Finally, a judge resolved the disagreements between annotators, establishing the relational structure model and specifying the signals for rhetorical relations. The average agreement between annotators of the cause subset —which is often signalled— was 78.11% (PURPOSE 90%, CAUSE 76.79% and RESULT 59.7%).

4. The RST Basque TreeBank

When entering in the website,[4] you can find information of the general characteristics of the RST Basque TreeBank and facilities to consult the contents of the tagged corpus, as for example: *i*) discourse units, the central unit and relations linked to the central unit (4.1 subsection); *ii*) all instances of a selected rhetorical relation in the corpus (4.2 subsection); *iii*) the rhetorical structure of a desired text (4.3 subsection); *iv*) all the signals of relations (4.4 subsection) and, *v*) searching facilities for further studies about typical patterns about combination of word-forms, lemma and POS present in the corpus (4.5 subsection).

4.1. Consulting EDUs and CU of a tree

The application offers the possibility to check the linear segmentation (EDUs) of a document as well as its CU. Table 3 shows the segmentation for the GMB0301 document. The text has seven EDUs[5] and the last one, EDU_7, has a button called *See* in the CU column. If you click on this button, you will see all the relations linked to the CU of this text.

4.2. Dealing with rhetorical relations

The web application allows you to look up all the occurrences of a specific relation, or restrict your search to a particular sub-corpus (GMB, TERM or ZTF). If the segments are

---

[4] `http://ixa2.si.ehu.es/diskurtsoa/en/`

[5] Translations thereof are found underneath these.

	GMB0301-GS.rs3 (7)		
EDU	**Segment**	**Annotator**	**CU**
1	Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.	GS	
	Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.		
2	“Estomatitis aftosa recurrente” deritzon patologia, ahoan agertzen den ugarienetako bat da.	GS	
	“Recurrent aphthous stomatitis” is one of the most frequent oral pathologies.		
3	tamainu, kokapena eta iraunkortasuna aldakorra izanik.	GS	
	having a variable size, location and duration.		
4	Honen etiologia eztabaidagarria da.	GS	
	It has a controversial etiology.		
5	Ultzera mingarri batzu bezela agertzen da,	GS	
	It is characterized by the apparition of painful ulcers,		
6	Hauek periodiki beragertzen dira.	GS	
	These ulcers appear recurrently.		
7	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.	GS	See
	In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.		

Table 3. Example of the EDUs section, GMB0301

very long and you are only interested in the beginning of each, you can also limit the size.

Table 4 shows a fragment of a search conducted in the relation database. Since the search was limited to the TERM corpus, there are only 27 CAUSE relations, rather than the 56 shown in corpus. The first 3 columns of Table 4 describe the order and direction of the discourse units. Since the segments —left span and right span— follow the order in where they appear in the text, the second column specifies the nuclearity of the relations: if the relation is NS (nucleus on the left and satellite on the right), then the arrow points left (<–), towards the nucleus. If it is SN, then the arrow points right (–>). The fourth column specifies the relation and relation type: in this case, a single nucleus relation (N/S) CAUSE; when there are multiple nuclei, this is indicated by the letters (N/N). Finally, the source of the example (Ref.) and annotator (Annot.) is specified.[6]

		Relation: Cause (27)			
Left span	**NS**	**Right span**	**Relation**	**Ref.**	**Annot.**
Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unita[. . .]	<–	Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingelesari eman diote [. . .]	Cause	TERM18	GS
In recent decades, many Serbian researchers working in different scientific fields have noticed a tendency and this is outlined here: the English unit [. . .]		Indeed, Serbian scientists from different schools of thought have reached a consensus and have given English [. . .]			
Terminologiak berak ere, uztartu egin behar ditu joera orokor horiek, eransten zaizkien beste batzuekin batera, hala nola: teknologien [. . .]	<–	gizartearekin lotuta dagoen jarduera denez,	Cause	TERM19	GS
Terminology itself must seek to unite these general trends, along with others related to them, for example: technology [. . .]		since it is an activity linked to society,			

Table 4. Example of a CAUSE relation search

---

[6] Note: due to space limitations we only mention here the most important information contained in the database. The signals for rhetorical relations are underlined in Table 4.

4.3. Checking all relations of a RST tree

You can also consult the database file by file: viewing the rhetorical relations of the chosen file or its image in JPG format. The rhetorical structure can be consulted in different formats (XML and RS3). Other information can be consulted here: text file in TXT format, morphosyntactic information annotated automatically in KAF format [Bosma et al. 2009], and the signals for relations annotated in RHETDB format.

4.4. Signals of rhetorical relations

You can check if a signal is in more that one relation. We show as an example a query based on the adversative conjunction *baina* 'but' in Table 5, which signals two similar relations (CONTRAST and CONCESSION).[7]

Signal: *baina* 'but'			
Gainerakoan, prokasu adierazle egokiak daude,	Kontzesioa	baina altan dagoen gaixoaren ahalmen funtzionalaren erregistro urria antzematen da,	GMB0504
With respect to the other aspects, the indicators of process are good	Concession	but there is poor recording of the patient's functional capacity on discharge,	
Bestalde, Euskaltzaindiak hitz elkartuen bidea (1995eko urtarrilaren 27an onartutako araua) proposatzen du adjektibo erreferentzialak itzultzeko,	Kontrastea	baina arauan bertan esaten denez, "... ahal den guztian...",	TERM22
Euskaltzaindia proposed a mechanism of compound words (in a standard approved on January 27th 1995) for the translation of referential adjectives.	Contrast	However the academy also confirmed, ... "whenever possible",	

Table 5. Example of the SIGNALS section, the discourse marker *baina* 'but'

4.5. Word form, lemma and POS search interface

Searches combining word-form, lemma and POS features can be done in the application due to the fact that all the words in the texts have associated morphological and syntactical information in KAF format.

	Doc.	**Sent Id**	**Word**	**CU**	**Sentence**
1	TERM50	sent2	taldeek / helburua	BAI	[. . .] Hitzaldi honek azken hiru urteotan lau unibertsitate hauen *talde*ek egindako ikerkuntzaren ondorioetako batzuk azaltzeko *helburua* izango luke.
			groups / aim	YES	"[. . .] The aim of this talk is to present some of the results of the research carried out by groups from these four universities over the last three years."
2	ZTF13	sent1	taldearen / helburu	BAI	[. . .] Gure *ikerkuntza talde*aren *helburu* nagusia, [. . .]
			group's / aim	YES	[. . .] Our research group's principal aim, [. . .]
3	ZTF13	sent17	taldearen / helburu	EZ	Alor honetan, gure *ikerkuntza talde*aren *helburu* nagusiak bi dira.
			group's / aim	NO	In this field, our research group has two main aims.
1	ZTF15	sent7	helburu / talde	EZ	[. . .] bestelako galdera zailagoei ere erantzutea dute *helburu*, hala nola, espezieen biogeografia, *talde*aren filogenia, eta abar.
			aim / group	NO	[. . .] the aim is to answer other such difficult questions, such as species biogeography, group phylogeny, etc.

Table 6. Example of the SEARCH section

These searches provide the option of searching patterns. For example, in a two-word search, you can specify to show the sentences which contain words starting with the forms *talde* 'group' or 'team' and *helburu* 'goal' or 'aim'. You can also define whether or not other words can be located between the target terms. Table 6 shows a search for the

---

[7] More information about ambiguity in this corpus can be read in [Iruskieta and da Cunha 2010] and in [Iruskieta et al. 2009].

terms *talde* 'group' and *helburu* 'aim' results in two YES responses for CU, but another search with the terms the other way round (aim and group) would only give one NO response for CU.

5. Discussion and Future Work

This paper presents the first RST Basque TreeBank, where the gold standard files that have been used to compile the database are at the disposal of anyone who wishes to use them. Moreover, the study also served to design the harmonization processes for the different annotation phases (segmentation, identification of central units, rhetorical relations and its signals), as well as giving the judge the opportunity of consulting both their annotations and those of the annotators, seeing at a single glance the frequency of each relation and its signals. This in turn enabled the detection of errors and incoherence during the establishment of the gold standards.

The work carried out is useful for certain language processing tasks. Indeed, during the course of the project we established a segmented gold standard for 60 texts, on the road towards automatic segmentation. As regards rhetorical relations, after establishing a gold standard for 60 texts, we marked the signals of those relations, being the size of the work similar to that of others in the literature [Taboada and Das Forthcoming]. In the future, this work will help us define rhetorical relation patterns, and this in turn will help us achieve automatic detection of those most commonly signaled relations.

The authors are currently striving to achieve the following aims: in the short medium term, their goal is to annotate texts from another genre: newspaper articles, texts from the EPEC corpus and to study deeply the signals of relations in the RST Basque TreeBank. With the data provided by the RST Basque TreeBank, they are implementing an automatic discourse segmentation program. Besides, and considering how time consuming the tagging and evaluation processes are, the authors are working on the implementation of a new interface to facilitate the editing of rhetorical relations and programs for automatic evaluation program based on rhetorical relations.

Acknowledgments

This study was carried out within the framework of the following projects: Ber2Tek (IE12-333); Hibrido Sint (TIN2010-20218); NewsReader project (FP7- ICT-2011-8-316404); IXA group, Research Group of type A (IT344-10). We would like to thank Esther Miranda and Kike Fernandez for their help in designing the web page.

References

[Asher and Lascarides 2003] Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge Univ Pr, Cambridge.

[Bosma et al. 2009] Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: a generic semantic annotation format. In *GL2009 Workshop on Semantic Annotation*, Italy.

[Bouayad-Agha 2000] Bouayad-Agha, N. (2000). Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts. In *Annual Meeting-ACL*, volume 38, pages 16–22.

[Carlson et al. 2003] Carlson, L., Marcu, D., and Okurowski, M. E. (2003). *Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory*, pages 85–112. Current and new directions in discourse and dialogue. Springer, Berlin.

[Carlson et al. 2002] Carlson, L., Okurowski, M. E., and Marcu, D. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus].* PA: Linguistic Data Consortium, Philadelphia.

[Ceberio et al. 2009] Ceberio, K., Aduriz, I., Díaz de Ilarraza, A., and Garcıa, I. (2009). Empirical study of the relevance of semantic information for anaphora resolution: the case of adverbial anaphora. In *7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC09)*, pages 56–63, Goa, India.

[da Cunha and Iruskieta 2010] da Cunha, I. and Iruskieta, M. (2010). Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

[da Cunha et al. 2011a] da Cunha, I., Torres-Moreno, J. M., and Sierra, G. (2011a). On the Development of the RST Spanish Treebank. In *5th Linguistic Annotation Workshop (LAW V '11)*, pages 1–10, Portland, USA.

[da Cunha et al. 2011b] da Cunha, I., Torres-Moreno, J. M., Sierra, G., Cabrera-Diego, L.-A., and Castro-Rolón, B.-G. (2011b). The RST Spanish Treebank On-line Interface. In *International Conference Recent Advances in NLP*, Bulgaria.

[Goenaga et al. 2012] Goenaga, I., Arregi, O., Ceberio, K., de Ilarraza, A. D., and Jimeno, A. (2012). Automatic coreference annotation in basque. In *Eleventh International Workshop on Treebanks and Linguistic Theories*, Portugal.

[Hovy 2010] Hovy, E. (2010). Annotation: A Tutorial. In *48th Annual Meeting of the ACL*, Uppsala, Sweden.

[Iruskieta and da Cunha 2010] Iruskieta, M. and da Cunha, I. (2010). Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera. In *XXVIII Congreso Internacional AESLA: Analizar datos > Describir variación*, pages 13–159, Vigo.

[Iruskieta et al. 2009] Iruskieta, M., de Ilarraza, A. D., and Lersundi, M. (2009). Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso. In *Proceedings of 27th AESLA International Conference*, pages 963–971, Ciudad Real, Spain.

[Iruskieta et al. 2011] Iruskieta, M., Díaz de Ilarraza, A., and Lersundi, M. (2011). Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, 47:144.

[Iruskieta et al. 2013] Iruskieta, M., Díaz de Ilarraza, A., and Lersundi, M. (2013). A critical analysis of rhetorical annotation: fundamental principles of discourse segmentation in basque. *Corpus Linguistics and Linguistic Theory*, 0(0):1–32.

[Krippendorff 2012] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. SAGE, London.

[Mann and Taboada 2010] Mann, W. C. and Taboada, M. (2010). RST web-site. *http://www.sfu.ca/rst/*.

[Mann and Thompson 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[Marcu 2000a] Marcu, D. (2000a). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

[Marcu 2000b] Marcu, D. (2000b). *The theory and practice of discourse parsing and summarization*. The MIT press, Cambridge.

[Marcu et al. 2000] Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17, Seattle (USA).

[Mitkov 2002] Mitkov, R. (2002). *Anaphora resolution*, volume 134. Longman London.

[O'Donnell 2000] O'Donnell, M. (2000). Rsttool 2.4: a markup tool for rhetorical structure theory. In *6th European Workshop on Natural Language Generation*, Germany.

[Pardo 2005] Pardo, T. A. S. (2005). Métodos para análise discursiva automática. PhD Thesis. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

[Pardo and Nunes 2004] Pardo, T. A. S. and Nunes, M. G. V. (2004). Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil. Technical Report NILC-TR-04-03.

[Pardo et al. 2004] Pardo, T. A. S., Nunes, M. G. V., and Rino, L. H. M. (2004). DiZer: An automatic discourse analyzer for Brazilian Portuguese. *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234.

[Pardo and Seno 2005] Pardo, T. A. S. and Seno, E. R. M. (2005). Rhetalho: um corpus de referência anotado retoricamente. *Anais do V Encontro de Corpora*, pages 24–25.

[Recasens et al. 2010] Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *5th International Workshop on Semantic Evaluation*, pages 1–8, Sweden. Association for Computational Linguistics.

[Spenader and Lobanova 2009] Spenader, J. and Lobanova, A. (2009). Reliable discourse markers for contrast relations. In *Proceedings of the 8th International Conference on Computational Semantics*, Tilburg, The Netherlands.

[Stede 2004] Stede, M. (2004). The Potsdam Commentary Corpus. In *2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.

[Swales 1990] Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge Univ Pr, Cambridge, UK.

[Taboada and Das Forthcoming] Taboada, M. and Das, D. (Forthcoming). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*.

[Taboada and Renkema 2011] Taboada, M. and Renkema, J. (2011). Discourse Relations Reference Corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

[van der Vliet et al. 2011] van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

Lang Resources & Evaluation (2015) 49:263–309
DOI 10.1007/s10579-014-9271-6

ORIGINAL PAPER

A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora

Mikel Iruskieta · Iria da Cunha · Maite Taboada

Received: 26 June 2013 / Accepted: 8 May 2014 / Published online: 28 May 2014

Abstract Explaining why the same passage may have different rhetorical structures when conveyed in different languages remains an open question. Starting from a trilingual translation corpus, this paper aims to provide a new qualitative method for the comparison of rhetorical structures in different languages and to specify why translated texts may differ in their rhetorical structures. To achieve these aims we have carried out a contrastive analysis, comparing a corpus of parallel English, Spanish and Basque texts, using Rhetorical Structure Theory. We propose a method to describe the main linguistic differences among the rhetorical structures of the three languages in the two annotation stages (segmentation and rhetorical analysis). We show a new type of comparison that has important advantages with regard to the quantitative method usually employed: it provides an accurate measurement of inter-annotator agreement, and it pinpoints sources of disagreement among annotators. With the use of this new method, we show how translation strategies affect discourse structure.

Keywords Annotation evaluation · Discourse analysis · Rhetorical Structure Theory · Translation strategies

---

M. Iruskieta (✉)
Department of Didactics of Language and Literature, University of the Basque Country,
Sarriena auzoa z/g, 48940 Leioa, Spain
e-mail: mikel.iruskieta@ehu.es

I. da Cunha
University Institute for Applied Linguistics, Universitat Pompeu Fabra, C/ Roc Boronat 138,
08018 Barcelona, Spain
e-mail: iria.dacunha@upf.edu

M. Taboada
Department of Linguistics, Simon Fraser University, 8888 University Dr, Burnaby, BC V5A 1S6,
Canada
e-mail: mtaboada@sfu.ca

1 Introduction

Translation or parallel corpora on the one hand and comparable corpora on the other are useful in many tasks, in applied linguistics and in natural language processing. Compiling such corpora can provide insight into translation strategies, can help validate or disprove intuitions about differences across languages, and can be useful in computational applications such as machine translation or terminology extraction.

Translation corpora have been useful in testing hypotheses about language contrasts. Granger (2003), for instance, using translation corpora, put into question the over-generalization that "French favors explicit linking while English tends to leave links implicit". Translation corpora also help identify strategies used in the translation process, such as the strategy that Xiao (2010) found in translated Chinese texts, where there was an increased use of discourse markers, presumably to more clearly identify the rhetorical structure of the text (although introducing discourse markers may lead to subtle changes in rhetorical structure as well, in cases when the translator interprets a different relation than that intended by the original author).

Most contrastive corpus-based studies emphasize surface-level aspects of language, such as differences in terminology in general (Gomez and Simoes 2009; Morin et al. 2007; Fung 1995; Wu and Xia 1994) and specific lexical items in particular (Fetzer and Johansson 2010; Flowerdew 2010); differences in aspects of modality (Kanté 2010; Usoniene and Soliene 2010); or the use of discourse markers (Mortier and Degand 2009). There exists, however, a sizeable body of work on differences in the rhetorical structure of texts across languages, in particular within the framework of Rhetorical Structure Theory (RST), a theory of text structure proposed by Mann and Thompson (1988). The first contrastive RST study comparing one European language and one Asian language was carried out by Cui (1986), who compared English and Chinese expository rhetorical structures. Kong (1998) and Ramsay (2000, 2001) studied the same pair of languages, in both cases examining specific genres (business request letters and news texts). Other pairs of languages studied within RST include Arabic and English (Mohamed and Omer 1999), Japanese and English (Marcu et al. 2000), or a range of European languages, such as Dutch-English (Abelen et al. 1993), Finnish–English (Sarjala 1994), French-English (Delin et al. 1996; Salkie and Oates 1999), Spanish–English (Taboada 2004a, b), and Spanish–Basque (da Cunha and Iruskieta 2010).

Contrastive studies comparing the rhetorical structures of more than two languages are not very common, although we can mention the study in Portuguese–French–English by Scott et al. Scott et al. (1998). They show a methodology to carry out RST contrastive analysis of instructional texts in different languages, and they present the results of an empirical cross-lingual experiment based on this methodology. More information about contrastive RST studies or studies about other languages can be found in Taboada and Mann (2006a, b).

One observation in RST-based work is that the same passage, when conveyed in two different languages, may have different underlying rhetorical structures (Bateman and Rondhuis 1997; Delin et al. 1994). An explanation for such differences is that translation strategies reorganize the structure of the discourse, with the resulting underlying structures being different. Translation literature deals with many aspects of this phenomenon, one being differences in explicitness, which in some cases result in different underlying structures (House 2004).

This proposal (that translation strategies lead to different structures) is often presented on the basis of individual examples, with no unifying principle for the representation of underlying structure. In this paper, we present a new method for the evaluation of discourse structures across multiple languages to analyze which translation strategies affect rhetorical structure.

The first aim of this paper is to provide a new qualitative method to compare rhetorical structures in different languages and/or by different annotators. Existing work comparing different annotations uses a quantitative methodology (Marcu 2000a). The main comparison methodology consists of quantifying the agreement between the rhetorical analyzes done by annotators, in terms of Elementary Discourse Units (EDUs), spans (sets of related EDUs), nuclearity (nucleus or satellite role of a span) and rhetorical relations (set of hypotactic and paratactic relations). To compare rhetorical analyzes, typical precision and recall measures are used. Work by da Cunha and Iruskieta (2010) and van der Vliet (2010) presents some criticisms of Marcu's methods, arguing that this quantitative method amalgamates agreement coming from different sources, because decisions at one level in the tree structure affect decisions and factors at other levels, with the result that the factors are not independent. Disagreement on segmentation or attachment point at lower levels in the tree significantly affects agreement on the upper rhetorical relations in a tree, and should be accounted separately. Mitocariu et al. (2013) have proposed an evaluation method (for RST and Veins Theory Cristea et al. 1998) which checks the inner nodes[1] (attachment point), nuclearity of the relation (nuclearity) and the vein expressions or constitution of the units ("constituent" Marcu 2000a) but excludes the names of relations as a comparison criterion. In our evaluation method we consider Mitocariu et al.'s factors (attachment point, constituent and nuclearity) and the rhetorical relations. We believe that the qualitative method that we present here addresses the deficiencies in previous proposals and provides a qualitative description of dispersion annotation, while at the same time allows the quantitative evaluation.

The second aim of this paper is to test this method. In order to detect differences among rhetorical structures and study the origin of such differences, we analyze a corpus of parallel texts in three different languages: English, a Germanic language; Spanish, a Romance language; and Basque, a non-Indo-European language. We investigate whether differences are motivated by different translation strategies or by the choice of one relation over another in a group of similar relations, as Stede (2008b) proposes. Our corpus, albeit small, is comparable to the only other trilingual comparative corpus (Scott et al. 1998), and it is rich enough to allow the development and evaluation of a qualitative comparison method for rhetorical relations.

Our study is useful from a theoretical point of view, because it will help us understand how the rhetorical structures of texts in different languages are

---

[1] Soricut and Marcu (2003, pg. 152) use the term "attachment point" or "dominance set".

constructed. Moreover, the study provides rhetorical analyzes of a less-commonly studied language,[2] Basque, the only pre-Indo-European language of Western Europe (Trask 1997) and one of the four official languages of Spain (together with Catalan, Galician and Spanish), spoken in the Basque country. From an applied point of view, this work supports the development of computational linguistics systems (such as summarization, information extraction and retrieval systems), where accurate annotation is of paramount importance. In addition, our methodology can be useful in research on automatic compilation of specialized corpora, and can help professional translators and machine translation researchers.

The paper is organized as follows: Section 2 presents the methodology and theoretical background of our study. Section 3 describes our methodological proposal and provides the results of the discourse analysis of our corpus. Section 4 provides conclusions and proposals for future work.

2 Methodology

Our work consisted of three stages. First, we decided on the theoretical framework of our study, RST. Second, we built the corpus. Finally, we carried out the analysis, including a comparison of the three different RST structures for each text, using both a quantitative methodology and our proposed new qualitative methodology.

2.1 Theoretical framework

In this study, we use RST, since it is a language-independent theory. RST is a descriptive theory for textual organization that characterizes text structure using relations among the discourse or rhetorical elements that a text contains. These elements are called spans, and they can be nucleus (if the element is more essential to the speaker's purpose) or satellite (if it provides some rhetorical information about the nucleus). The relations can be: (a) nuclear relations (e.g., ANTITHESIS, CAUSE, CIRCUMSTANCE, CONDITION, ELABORATION, EVIDENCE, JUSTIFICATION, MOTIVATION, PURPOSE), that is, hypotactic relations between nuclei and satellites, and (b) multinuclear relations (e.g., CONTRAST, JOINT, LIST, SEQUENCE), that is, paratactic relations among nuclei, where more than one unit is central with regard to the speaker's purposes. For a more detailed explanation of RST, see Mann and Thompson (1988) and the RST web site by Mann and Taboada (2010).

RST relations are typically represented as trees. Figure 1 shows a fragment of an RST tree,[3] with one multinuclear relation (CONJUNCTION) and two multinuclear

---

[2] Although great efforts have been made to stimulate Machine Translation studies for different language pairs, non-official languages that are typologically different and could be interesting are not considered. For example Koehn (2005) presents a 30 million word corpus translated to the 11 official of the European Union: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish to study different language pairs translations, but less common languages spoken in the EU are not included.

[3] The source of the text (TERM#_original language) is shown in square brackets at the end of the figures, tables or examples.

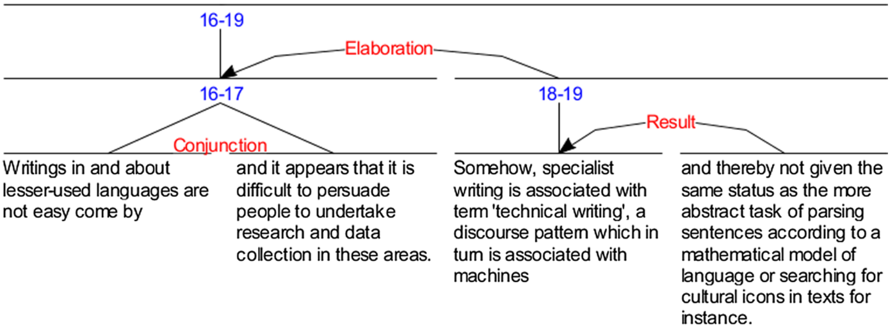

16-19
Elaboration
16-17
18-19
Result
Conjunction
Writings in and about lesser-used languages are not easy come by
and it appears that it is difficult to persuade people to undertake research and data collection in these areas.
Somehow, specialist writing is associated with term 'technical writing', a discourse pattern which in turn is associated with machines
and thereby not given the same status as the more abstract task of parsing sentences according to a mathematical model of language or searching for cultural icons in texts for instance.

Fig. 1 Example of an RST tree, TERM30_ENG

relations (RESULT and ELABORATION). The annotator recognized that spans 16 and 17 are conjoined, forming another span where each item has a comparable role (moreover, each span has a verb *are* and *appears*, and they are linked by the connector *and*). The annotator also found a RESULT relation, since she understood that span 18 could be the cause for the situation explained into the span 19 (again, each unit has a finite verb: *is associated* and *[is] given*, and they are linked by the double connector *and thereby*). It is important to observe that rhetorical relations are applied recursively, i.e., spans that stand in a relation: 18 and 19 in Fig. 1 form a new span (18–19) that can enter into new relations, such as the ELABORATION relation. In this case, the annotator labelled this relation as such because the span made up of units 18–19 (satellite) provides additional information about the previous span (16–17), which constitutes the nucleus of the relation. Following Marcu's (2000b) strong compositionality criteria, the most important units for the 16–19 span are 16 and 17. For the span 18–19 the most important unit is 18.

In the literature on RST, there is agreement that the most important unit of the tree is the "central unit(s)" (Stede 2008b) and the most important unit of a span is the "central subconstituent" (Egg and Redeker 2010). So following this framework we will use the term "Central Unit(s)" (CU) of the text for the most important unit of an rhetorical structure tree (RS-tree) and "Central Subconstituent(s)" (CS) of a relation for the most important unit of the modifier span that is the most important unit of the satellite span. When there is a simple constituent (that is no more than one EDU), we formalized this simple constituent as the CS, and when there is a multinuclear relation, we describe it with all of its constituents.

Table 1 provides a representation of this example.

There are several classifications of RST relations: the classic one by Mann and Thompson of 24 relations (Mann and Thompson 1988), the extended one by Mann and Thompson of 30 relations, available on the RST site (Mann and Taboada 2010), and Marcu's classification of 78 relations (Carlson et al. 2003), among others. We have chosen the extended classification for the annotation of our trilingual corpus. Space constraints preclude an extensive discussion of its merits over other approaches (see Taboada and Mann 2006a, for a discussion).

Table 1 Formalization of Fig. 1, TERM30_ENG

Relation	Left span	Right span	CS	Nuclearity
Result	18	19	19	NS
Conjunction	16	17	16–17	NN
Elaboration	16–17	18–19	18	NS

2.2 Corpus

As Granger (2003) proposes, a multilingual translation corpus is:

> [. . .] the most obvious meeting point between CL (Contrastive Linguistics) and TS (Translation Strategies). Researchers in both fields use the same resource but to different ends: uncovering differences and similarities between two (or more) languages for CL and capturing the distinctive features of the translation process and product for TS.

(Granger 2003, pg. 22)

In translation studies, where the intention is to search for similarities and differences in large corpora, it is difficult to find a balanced corpus in size and similar composition of genres (Baker 2004). Our problem was to find a balanced multidirectional corpus of such size that allowed for a manual comparison of all the rhetorical structures by language pair. One of our aims, as we said, is to propose a methodology to describe when a different RST relation can be attributed to annotator interpretation or to different language forms.

As far as we know, no multilingual corpus with English, Spanish and Basque texts exists. Our corpus was then compiled specifically for this work.[4] It is a multidirectional translation corpus which contains abstracts of research papers published in the proceedings of the International Conference about Terminology that took place in Donostia and Gasteiz in 1997 (UZEI and HAEE-IVAP 1997). In this conference, authors were allowed to send full papers in English, French, Spanish or Basque, but they had to provide titles and abstracts in the four languages. In order to have a multidirectional and trilingual balanced corpus, we have chosen abstracts for which the original paper was written in English (five texts), Spanish (five texts) and Basque (five texts). Thus, we have analyzed 15 abstracts (the same ones for each language), written by different authors, constituting three subcorpora. In sum, our corpus includes 45 texts. Table 2 summarizes the statistics of the subcorpora.

In order to find correlations between translation strategies and rhetorical relations, a methodology that can compare parallel rhetorical structures is needed. We built our corpus in order to develop such a methodology, and consider that the number of texts is sufficient for the design of the qualitative method that we present.

---

[4] A problem with work in the framework of RST is that there is no annotated bilingual or trilingual corpus to study the effects of translation strategies on rhetorical structure. As a consequence, a researcher in such situation first needs to learn RST and perform annotations, as Maxwell (2010) suggests.

Table 2 Corpus statistics

Subcorpus	Annotators	Texts	Words	Sentences	EDUs
ENG	A1	15	5,706	201	318
SPA	A2	15	6,324	193	318
BSQ	A3	15	4,800	197	318

This qualitative method applies to any type of text,[5] since the principles on which it is based are general RST-based principles. We believe that the analysis is general enough and the method applicable across genres. We also discuss some examples detected with the qualitative evaluation in this parallel corpus that show how translation strategies could be related to rhetorical structures (see Sect. 3.2.2).

After the corpus compilation, we carried out the analysis. This analysis had two main phases: discourse segmentation and rhetorical analysis.

2.3 Discourse segmentation

The first step in analyzing texts with RST consists of segmenting the text into spans. Exactly what a span is, in the framework of RST, and more generally in discourse, is a well-debated topic. RST (Mann and Thompson 1988) proposes that spans, the minimal units of discourse—later called Elementary Discourse Units (EDUs) (Marcu 2000a)—are clauses, but that other definitions of units are possible.

From our point of view, adjunct clauses stand in clear rhetorical relations (cause, condition, concession, etc.). Complement clauses, however, have a syntactic, but not discourse, relation to their host clause. Complement clauses include, as Mann and Thompson (1988) point out, subject and object clauses, and restrictive relative clauses, but also embedded report complements, which are, strictly speaking, also object clauses.

Other possibilities for segmentation exist; one of the better-known ones is the proposal by Carlson et al. (2003) for segmentation of the RST Discourse Treebank (Carlson et al. 2002). Carlson et al. (2003) propose a much more fine-grained segmentation, where report complements, relative clauses and appositive elements constitute their own EDUs.

In our work three annotators segmented the EDUs of each subcorpus (A1 segmented English texts, A2 segmented Spanish texts, and A3 segmented Basque texts).[6]

---

[5] It was used also to evaluate the RST Basque TreeBank (Iruskieta et al. 2013a), available at: http://ixa2.si.ehu.es/diskurtsoa/en/.

[6] When a corpus is annotated only with one annotator per language, the results may yield subjective idiosyncrasies. This is not a problem for the aim of this paper, because we do not want to provide a reliable annotated corpus in three languages, but we do provide a qualitative way to compare annotation in different languages. Comparisons have been done manually and by pairs of languages following two different evaluations: (a) Marcu's quantitative method and (b) a new qualitative-quantitative method. So even if the corpus is small, the comparison work is extensive. The aim to provide reliable corpora has been achieved in other papers by the authors [English SFU corpus (Taboada and Renkema 2008), Spanish RST TreeBank (da Cunha et al. 2011a) and Basque RST TreeBank (Iruskieta et al. 2013a)].

These annotators are experts in RST, having carried out research in this field for a number of years, and they have participated in several projects related to the design and elaboration of RST corpora in the three languages under consideration. Annotators performed this segmentation task separately and without contact among them. In our segmentation, we follow the general guidelines proposed by Mann and Thompson (1988) which we have operationalized for this paper. We detail the principles below.

2.3.1 Every EDU should have a verb

In general, EDUs should contain a (finite) verb. The main exception to this rule is the case of titles, which are always EDUs, whether they contain a verb or not. Non-finite verbs form their own EDUs only when introducing an adjunct clause (but not a modifier clause; see "Appendix" for a detailed explanation).

2.3.2 Coordination and ellipsis

Coordinated clauses are separated into two segments, including cases where the subject is elliptical in the second clause. In Spanish and Basque, both pro-drop languages, this is in fact the default for both first and second clause, and therefore we see no reason why a clause with a pro-drop subject cannot be an independent unit. We follow the same principle for English.

Coordinated verb phrases (VPs) or verbs do not constitute their own EDUs. We differentiate coordinated clauses from coordinated VPs because the former can be independent clauses with the repetition of a subject; the latter, in the second part of the coordination, typically contain elliptical verbal forms, most frequently a finite verb or modal auxiliary.

2.3.3 Relative, modifying and appositive clauses

We do not consider that relative clauses (whether restrictive or non-restrictive), clauses modifying a noun or adjective, or appositive clauses constitute their own EDUs. We include them as part of the same segment together with the element that they are modifying. This departs from RST practice, where (restrictive) relative clauses are often independent spans, as seen in many of the examples in the original literature and the analyzes on the RST web site (Mann and Thompson 1988; Mann and Taboada 2010). We found that relative clauses and other modifiers often lead to truncated EDUs, resulting in repeated use of the SAME-UNIT label,[7] and thus decided that it was best not to elevate them to the status of independent segments.

2.3.4 Parentheticals

The same principle applies to parentheticals and other units typographically marked as separate from the main text (with parentheses or dashes). They do not form an

---

[7] See the paragraph on Truncated EDUs in this section.

individual span if they modify a noun or adjective, but they do if they are independent units, with a finite verb.

2.3.5 *Reported speech*

We believe that reported and quoted speech do not stand in rhetorical relations to the reporting units that introduce them, and thus should not constitute separate EDUs, also following clear arguments presented elsewhere (da Cunha and Iruskieta 2010; Stede 2008a). This is in contrast to the approach in the RST Discourse Treebank (Carlson et al. 2003), where reported speech (there named ATTRIBUTION) is considered as a separated EDU. There are, in any case, no examples of reported speech in our corpus.

2.3.6 *Truncated EDUs*

In some cases, a unit contains a parenthetical or inserted unit, breaking it into two separate parts, which do not have any particular rhetorical relation between each other. In those cases, we make use of a non-relation label, Same-unit, proposed for the RST Discourse Treebank (Carlson et al. 2003).

Once our segmentation criteria were established and the three annotators carried out the segmentation, the three segmentations were compared in terms of F-measure and Kappa. In this way, we quantified agreement and disagreement across segmentations. Moreover, we analyzed the main causes of the disagreements. Results are shown in Sect. 3.1. After the segmentation agreement evaluation, we harmonized the segmentation, ensuring that units were comparable across the languages. At this point, we also calculated linguistic distance between the pairs of languages, by calculating which language required the most changes in the harmonization process. This harmonization process was necessary to start out the analysis with similar units, and to avoid confusing analysis disagreement and segmentation agreement. Marcu et al. (2000) and Ghorbel et al. (2001) also align (which we termed harmonize) their texts, decreasing the granularity of their segmentation to avoid complexity. With this decision, we lose some rhetorical information at the most detailed level of the tree. This does not, however, affect higher levels of tree structure. The results of this harmonization are shown in Sect. 3.1.1.

2.4 Rhetorical analysis

Starting from the same discourse segmentation, we carried out the discourse annotation of our corpus. Once again, A1 annotated English texts, A2 annotated Spanish texts and A3 annotated Basque texts, using the mentioned extended discourse relations set and RSTTool (O'Donnell 2000), a graphical interface widely used for RST annotation. We compared the resulting rhetorical trees using two different evaluation methods. One of them, which we characterize as a quantitative evaluation, was proposed by Marcu (2000a), and the other one, which we describe as a qualitative evaluation, was developed by our research team.

A qualitative comparison method for rhetorical structures in multilingual corpora should quantify data, but also (and more importantly) should show linguistic features affecting rhetorical structure. The quantitative/qualitative distinction is due to the fact that the first method only gives us an approximate measure of agreement, whereas the second method provides a qualitative description of annotation dispersion. The qualitative evaluation, in addition to its use as a measure of inter-annotator agreement, can also be deployed to evaluate discourse structures built by a parser.

2.4.1 Quantitative evaluation

In this section we present the quantitative method of Marcu (2000a) and its limitations, already pointed out in other works (van der Vliet 2010; da Cunha and Iruskieta 2010; Iruskieta et al. 2013b). The main limitations are:

1. Two of the factors evaluated, nuclearity and relation, are not independent of each other: factor conflation.
2. The description of comparison and weight given to the agreement in certain rhetorical relations could be improved: deficiencies in the description.

Marcu (2000a) presented a method to evaluate the correctness of discourse trees, comparing automatically-built trees with manually-built ones. This method measures recall and precision according to four factors: Elementary Discourse Units (EDU), units linked with relations (Span), nuclear or satellite position (Nuclearity) and rhetorical meaning of units (Relation). We refer to this method as the quantitative method, because it uses exclusively numerical measures.

1. Factor conflation: nuclearity and relations. When measuring the relation factor, the quantitative method conflates the label SPAN with a relation. Thus, the SPAN label carries the same weight as any other relation. As we can see in Fig. 2, one of the annotators has labelled the relation as ELABORATION, and the other as EVIDENCE.

If we describe such disagreement with the quantitative method, we can see that there is a degree of agreement with respect to the relation in the Fig. 3, when in fact the agreement captured is simply the agreement in nuclearity, that is, in SPAN. Figure 3 shows the results obtained after the comparison of the two rhetorical structures included in Fig. 2 by using the quantitative evaluation. These results have been obtained automatically by using RSTeval, which is an implementation of Marcu's comparison method.[8]

RSTeval does not take into account the language of the rhetorical structures; however, it eliminates the stopwords of each language from the text, which are not used to build the EDUs and Spans. In the first table of Fig. 3, absolute matches between structures can be observed (e.g. Units: Matches = 2 of 2), as well as percentages (e.g. Units: Recall = 1/Precision = 1), for the four mentioned factors.

---

[8] This evaluation method has been automated by Maziero and Pardo (2009) and nowadays it can be used in four languages: English, Spanish, Portuguese and Basque. Available at http://www.nilc.icmc.usp.br/rsteval/.

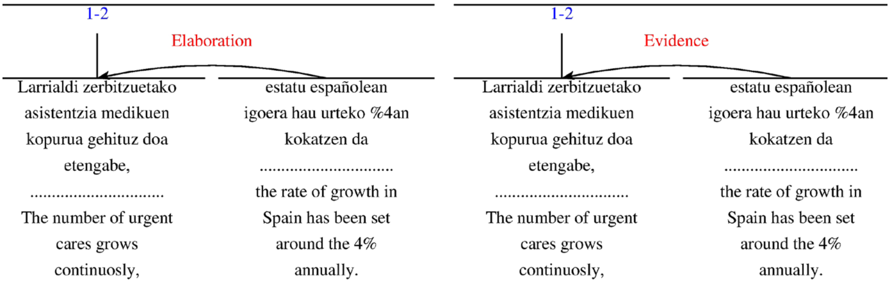
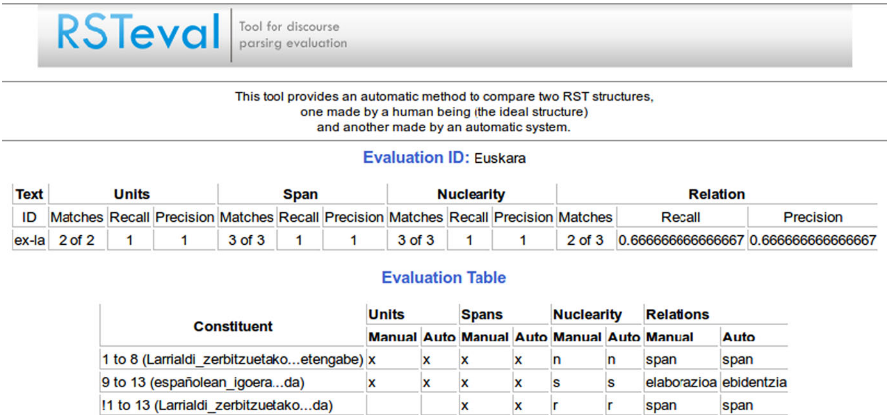

1-2
Elaboration
Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,
....
The number of urgent cares grows continuosly,
estatu españolean igoera hau urteko %4an kokatzen da
....
the rate of growth in Spain has been set around the 4% annually.
1-2
Evidence
Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,
....
The number of urgent cares grows continuosly,
estatu españolean igoera hau urteko %4an kokatzen da
....
the rate of growth in Spain has been set around the 4% annually.

Fig. 2 Quantitative evaluation: factor conflation (Iruskieta et al. 2013a, GMB0401)

The second table of Fig. 3 shows the detailed comparison process, where all the constituents of the structures are included. In this case, the first constituent corresponds to the first EDU, that is, words from "1 to 8" in the text; the second constituent corresponds to the second EDU, that is, words from "9 to 13"; and the third constituent corresponds to the Span formed by the two mentioned EDUs, that is, words from "!1 to 13" (the exclamation point at the beginning means that the constituent is a Span). The symbol "x" indicates that a Unit or Span is included in the corresponding rhetorical structure; "n" means nucleus; "s" means satellite, and "r" refers to the biggest span, that is, the span including the complete text. In the Relations factor, if there is a nucleus, the category "span" is included when a nuclear relation is under consideration or the name of relation when a multinuclear relation is under consideration, while, if there is a satellite, the name of the corresponding rhetorical relation is included.

Figure 4 shows a real example extracted from Iruskieta et al. (2013a).

In Table 3 we can see how *RSTeval* describes the agreement. The agreement levels are shown in Table 4. For ease of reference, we have highlighted the disagreements in italicize.



This tool provides an automatic method to compare two RST structures,
one made by a human being (the ideal structure)
and another made by an automatic system.

Evaluation ID: Euskara

Text	Units			Span			Nuclearity			Relation		
ID	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision
ex-la	2 of 2	1	1	3 of 3	1	1	3 of 3	1	1	2 of 3	0.666666666666667	0.666666666666667

Evaluation Table

Constituent	Units		Spans		Nuclearity		Relations	
	Manual	Auto	Manual	Auto	Manual	Auto	Manual	Auto
1 to 8 (Larrialdi_zerbitzuetako...etengabe)	x	x	x	x	n	n	span	span
9 to 13 (españolean_igoera...da)	x	x	x	x	s	s	elaborazioa	ebidentzia
!1 to 13 (Larrialdi_zerbitzuetako...da)			x	x	r	r	span	span

Fig. 3 Quantitative evaluation of Fig. 2 with RSTeval

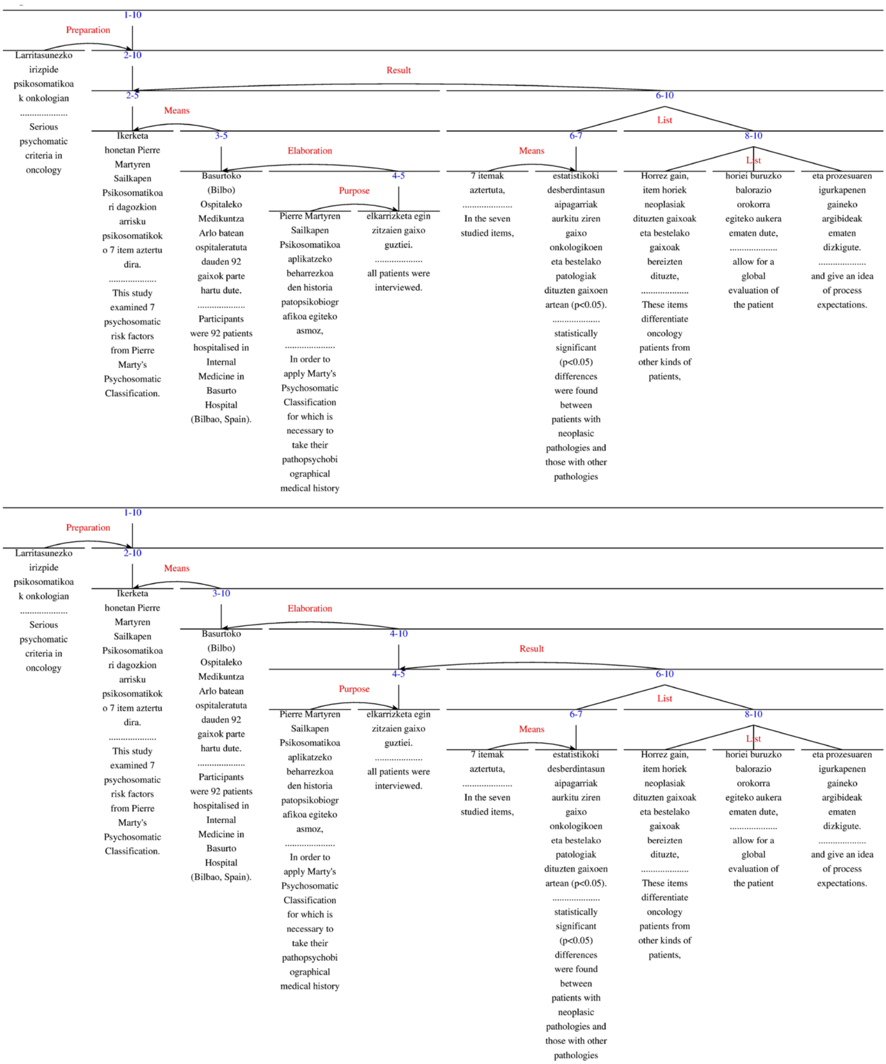

1-10
Preparation
2-10
Result
2-5
6-10
Means
List
3-5
6-7
8-10
Elaboration
Means
List
4-5
Purpose
Larritasunezko irizpide psikosomatikoak onkologian .... Serious psychomatic criteria in oncology
Ikerketa honetan Pierre Martyren Sailkapen Psikosomatikoari dagozkion arrisku psikosomatikoko 7 item aztertu dira. .... This study examined 7 psychosomatic risk factors from Pierre Marty's Psychosomatic Classification.
Basurtoko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute. .... Participants were 92 patients hospitalised in Internal Medicine in Basurto Hospital (Bilbao, Spain).
Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, ... In order to apply Marty's Psychosomatic Classification for which is necessary to take their pathopsychobiographical medical history
elkarrizketa egin zitzaien gaixo guztiei. .... all patients were interviewed.
7 itemak aztertuta, ... In the seven studied items,
estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05). .... statistically significant (p<0.05) differences were found between patients with neoplasic pathologies and those with other pathologies
Horrez gain, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte, ... These items differentiate oncology patients from other kinds of patients,
horiei buruzko balorazio orokorra egiteko aukera ematen dute, ... allow for a global evaluation of the patient
eta prozesuaren igurkapenen gaineko argibideak ematen dizkigute. .... and give an idea of process expectations.
1-10
Preparation
2-10
Means
3-10
Elaboration
4-10
Result
4-5
6-10
Purpose
List
6-7
8-10
Means
List
Larritasunezko irizpide psikosomatikoak onkologian .... Serious psychomatic criteria in oncology
Ikerketa honetan Pierre Martyren Sailkapen Psikosomatikoari dagozkion arrisku psikosomatikoko 7 item aztertu dira. .... This study examined 7 psychosomatic risk factors from Pierre Marty's Psychosomatic Classification.
Basurtoko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute. .... Participants were 92 patients hospitalised in Internal Medicine in Basurto Hospital (Bilbao, Spain).
Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, ... In order to apply Marty's Psychosomatic Classification for which is necessary to take their pathopsychobiographical medical history
elkarrizketa egin zitzaien gaixo guztiei. .... all patients were interviewed.
7 itemak aztertuta, ... In the seven studied items,
estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05). .... statistically significant (p<0.05) differences were found between patients with neoplasic pathologies and those with other pathologies
Horrez gain, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte, ... These items differentiate oncology patients from other kinds of patients,
horiei buruzko balorazio orokorra egiteko aukera ematen dute, ... allow for a global evaluation of the patient
eta prozesuaren igurkapenen gaineko argibideak ematen dizkigute. .... and give an idea of process expectations.

Fig. 4 Annotations of text GMB0701 (Iruskieta et al. 2013a)

When examining the rhetorical relations factor, we can see that the SPAN label plays a role in the description of agreement levels in Table 4: F-measure: 0.842 (16 agreements out of 19). If we describe the agreement without the SPAN label, however, the degree of agreement changes, as we can see in Table 5: F-measure: 0.778 (7 agreements out of 9).[9]

---

[9] Note that, after harmonizing discourse segmentation, accuracy, precision, recall and F-measure obtain the same value. Therefore, although this results in a somewhat artificial level of agreement, we are conscious about this fact, we use the standard measure employed in the RST literature (Marcu 2000a; Maziero and Pardo 2009).

Table 3 Qualitative method for text GMB0701

EDU	Constituent	Units		Spans		N/S		Relations	
		A3	A4	A3	A4	A3	A4	A3	A4
1	1 to 4 (Larritasunezko_irizpide...onkologian)	x	x	x	x	s	s	Preparation	Preparation
2	5 to 15 (Ikerketa_Pierre...aztertu)	x	x	x	x	n	n	Span	Span
3	16 to 22 (Basurtoko_Ospitaleko...gaixok)	x	x	x	x	n	n	Span	Span
4	23 to 31 (Pierre_Martyren...asmoz)	x	x	x	x	s	s	Purpose	Purpose
5	32 to 35 (elkarrizketa_zitzaien...guztiei)	x	x	x	x	n	n	Span	Span
4–5	!23 to 35 (Pierre_Martyren...guztiei)			x	x	*s*	*n*	*Elaboration*	*Span*
6	36 to 38 (7_itemak...aztertuta)	x	x	x	x	s	s	Means	Means
7	39 to 50 (estatistikoki_desberdintasun...05)	x	x	x	x	n	n	Span	Span
6–7	!36 to 50 (7_itemak...05)			x	x	n	n	List	List
8	51 to 57 (Horrez_item...bereizten)	x	x	x	x	n	n	List	List
9	58 to 60 (horiei_balorazio...orokorra)	x	x	x	x	n	n	List	List
8–9	!51 to 60 (Horrez_item...orokorra)			x	x	n	n	List	List
10	61 to 65 (prozesuaren_igurkapenen...dizkigute)	x	x	x	x	n	n	List	List
8–10	!51 to 65 (Horrez_item...dizkigute)			x	x	n	n	List	List
6–10	!36 to 65 (7_itemak...dizkigute)			x	x	s	s	Result	Result
4–10	!23 to 65 (Pierre_Martyren...dizkigute)				*x*		*s*		*Elaboration*
3–10	!16 to 65 (Basurtoko_Ospitaleko...dizkigute)				*x*		*s*		*Means*
2–10	!5 to 65 (Ikerketa_Pierre...dizkigute)			x	x	n	n	Span	Span
1–10	!1 to 65 (Larritasunezko_irizpide...dizkigute)			x	x	r	r	Span	Span
3–5	!16 to 35 (Basurtoko_Ospitaleko...guztiei)			*x*		*s*		*Means*	
2–5	!5 to 35 (Ikerketa_Pierre...guztiei)			*x*		*n*		*Span*	

Table 4 Quantitative method: agreement level for text GMB0701

Units			Spans			N–S			Relations		
Match	R	P	Match	R	P	Match	R	P	Match	R	P
10 of 10	1	1	17 of 19	0.895	0.895	16 of 19	0.842	0.842	16 of 19	0.842	0.842

Table 5 Agreement level according to rhetorical relations in GMB0701

Relations		
Match	R	P
7 of 9	0.778	0.778

2. Deficiencies in the description. When annotators decide that a relation has an attachment point at different levels in the tree structure (da Cunha and Iruskieta 2010), the method proposed by Marcu (2000a) is not able to compare the relations where constituents has changed. Observe the following issues in Fig. 4:

- In Table 3 the agreement in the ELABORATION relation cannot be included, because the relation has different spans: in A3 ‘23 to 31’ and in A4 ‘!23 to 65’ both attachments are referred as the same constituent, ‘23 to 31’.
- The MEANS constituent of A3 ‘!16 to 35’ and in A4 of ‘!16 to 65’, both attach to the same EDU (EDU2 or ‘5 to 15’); but, since the constituents do not coincide, the two MEANS relations cannot be compared.

Following da Cunha and Iruskieta (2010), Iruskieta et al. (2013b) and Mitocariu et al. (2013), we think that a qualitative method should describe the six factors involved in all rhetorical relations independently: EDU and Span (segmentation), nucleus-satellite function (Nuclearity), and attachment point, constituent and rhetorical meaning (Relation). When parallel texts are compared, a qualitative method should take in account whether the language form is parallel, as explained in the next section.

2.4.2 *Qualitative evaluation*

The qualitative evaluation method that we propose considers both type of agreement and source of disagreement, which results in a better explanation of the dispersion in annotator interpretations about text structure. When analyzing rhetorical structures using Marcu’s method, we observed that similar structures at the intermediate level of a tree structure spans could not be compared, because the constituents did not coincide. Such structures had, however, the same rhetorical relation, and the fact that the relation is the same should be reflected in a measure of agreement. If we accept that constituents do not need to coincide in their (span size) entirety to be compared, the issue is whether we can state that there is agreement with respect to the rhetorical relation, but disagreement about the constituents.

In our evaluation method it is not necessary for the constituents to be compared to be identical, like in Marcu's (2000b) method; only the central subconstituent (CS) has to be the same.[10] With such restriction we are able to compare rhetorical relations, using four independent criteria: constituent, attachment point, the direction of the relation (nuclearity) and effect of the relation.

When comparing RST structures with independent factors, we do not use typical nucleus and satellite terms to describe the extension of spans, because our method assesses independently nuclearity and unit size. The comparison in our method is based on rhetorical relations and not in spans of relations as Marcu's (2000b) method does. In our method we have a line for each relation, while in Marcu's (2000b) method there are two lines for each relation. The term constituent (C) refers to the length of the constituents, and the term attachment point (A) refers at the height of the tree where the constituent is linked (in Marcu's (2000b) evaluation method this factor is not considered, because what is compared are spans of relations). Because we are comparing relations and not spans of relations, in our comparison also nuclearity has a different meaning; while in Marcu's (2000b) method nuclearity has two possible values (S or N, where S means satellite and N means nucleus) for each span, in our method nuclearity has three values (SN, NN and NS) for each relation.

First of all, we present the types of agreement, and the two sources of disagreement in the qualitative evaluation by comparing annotators' RST trees. We measure the agreement in rhetorical relations based on the following factors: constituent (C), attachment point (A) and the name of relation (R), checking some agreement types:

1. Agreement in relation, constituent and attachment point (**RCA**).
2. Agreement in relation and constituent (**RC**).
3. Agreement in relation and attachment point (**RA**).
4. Agreement only in relation (**R**).

A decision tree formalizes the method to check the agreement types in rhetorical relations (see Fig. 5). As we mentioned before, to check agreement in rhetorical relation, the constituent of this relation must have the same central subconstituent (CS). If this condition is fulfilled, we check if relation name (R), constituent (C) and attachment point (A) are exactly the same.

We distinguish two sources of disagreement, disagreements of type A and type L, for Annotator and Language disagreements:

Disagreements of type A (Annotator). No significant linguistic differences in the text, but distinct relations labelled by two annotators (marked with an [A] in column Disagree of Table 7, and in corpus results in Table 17 under Annotation Discrepancies). We have found seven sources of such disagreement:

1. Different choice in nuclearity entailed a N/N–N/S mix-up **(N/N–N/S)**.
2. Different choice in nuclearity entailed discrepancy in N/S relations **(N/S)**.

---

[10] If there is more than one CS (because there is a multinuclear relation) at least one of them has to be the same for N/S-N/N mix-up.

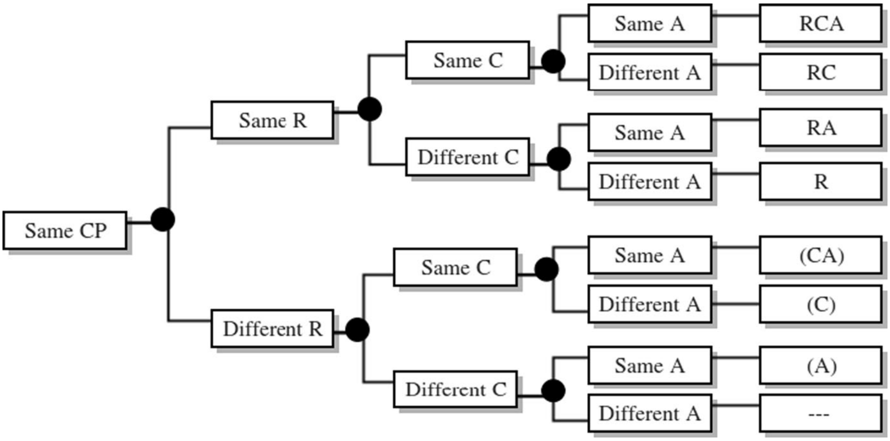


Fig. 5 Decision tree based on CS to establish the agreement types about R

3. A relation has the same constituent and attachment point, but not the same relation label ($\neq$ **R**).
4. Relations chosen are similar in nature **(Similar R)**.
5. Relations with mismatched RST trees **(Mismatch R)**.
6. A relation is more specific than the other **(Specificity)**.
7. Different choice in attachment entailed a different relation **(Attachment)**.

Disagreements of type L (Language). Two annotators labelled distinct relations because there is a significant difference in the linguistic form (marked with an [L] in column Disagree of Table 7 and in corpus results in Table 20 under Translation Strategies). We have found three different sources. These are in fact translation strategies, and are sensitive to corpus and language. Studies in other corpora, genre or languages may reveal different strategies and sources of disagreement:

1. A relation is signaled with a different discourse marker **(Marker Change or MC)**.
2. A different organization of constituent phrases is used, mostly from non-finite verb phrase to finite verb phrase **(Clause Structure Change or CSC)**.
3. A change in unit level (phrase–clause–sentence) is done **(Unit Shift or US)**.

In Table 6 we show an example extracted from the corpus of text TERM38_SPA which was segmented and harmonized in Spanish (A2) and in English (A1) (Fig. 7) to illustrate the qualitative method (Table 7).[11]

---

[11] Basque segments (A3) were also harmonized, but space constraints preclude us to align with Spanish and English. Anyway, the harmonization of TERM38_SPA segmentation in the three languages can be consulted at: http://ixa2.si.ehu.es/rst/segmentuak_multiling.php?bilatzekoa=TERM38%. The English RS-tree can be consulted at: http://ixa2.si.ehu.es/rst/diskurtsoa_jpg/TERM38_A1.jpg. The Spanish RS-tree can be consulted at: http://ixa2.si.ehu.es/rst/diskurtsoa_jpg/TERM38_A2.jpg.

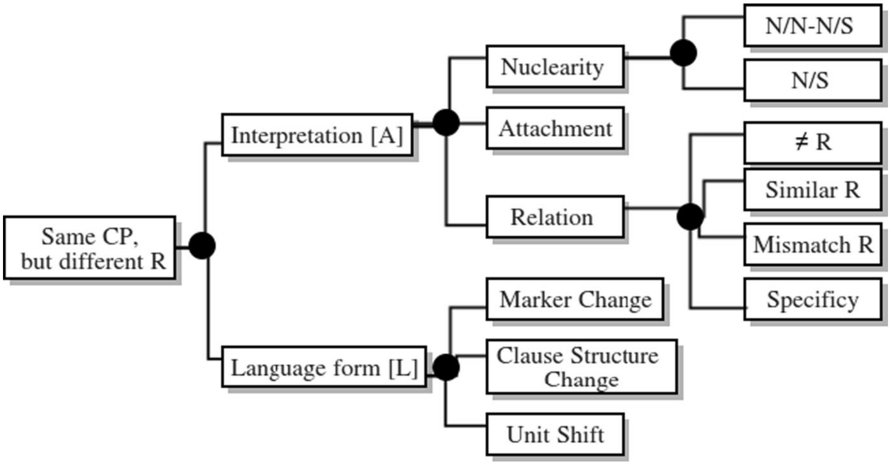


Fig. 6 Decision tree to establish the sources of agreement and disagreement about R

Table 7 includes the analyzed factors for Fig. 7: nuclearity (N), relation (R), constituent (C) and attachment point (A). These factors compare A2 (Spanish) and A1 (English). In the Qualitative Evaluation columns, we mark with a "✓" an instance of agreement, and with an "×" a disagreement. The last two columns summarize the type of agreement (Agree) or the disagreement source (Disagree).

If there is a multinuclear relation inside of a constituent of another relation (see lines 22 and 23 in Table 7) comparing CSs is not trivial, because multinuclear relations have more than one CS. Line 23 is representative of this problem. If we look at this line we can see that the problem is not the relation that we are comparing, but the problem comes from a lower level, since there is full agreement (RCA) between annotators (on R: ELABORATION, on C: 11N and on A: 12–14S). When this is the case there are two choices: (a) do not compare relations and annotate as "no-match"[12] and (b) compare first non-ambiguous CSs and leave problematic comparisons (lines 22 and 23) for the end. Following the last choice there is not any ambiguous CS in Table 7, because the other CS candidate (CS 12 in line 10) was used in other structure. Because of that, when we have to compare relations with more than one CS with another that has only one CS, at least one of the CSs has to be identical. If still there were cases in which we can not compare structures we have used the no-match label. This problem was found also in text summarization by Marcu Marcu (2000b), since the most important unit can be formed by more than one EDU.[13]

In Table 8 we present the results of our evaluation method for the example in Fig. 7.

---

[12] If we follow this decision, we could not compare structures that contain a N/N–N/S mix-up inside the relation.

[13] As the evaluation has been done manually, there have been some problematic cases that have not counted as an agreement. For cases in which some structures cannot be compared, no-match label has been used, which represents not more than 0.06 % of all relations (53 no-match/900 relations), about 1.18 relations per text on average (53 No Match/45 texts).

Table 6 TERM38_SPA segmented and harmonized in Spanish and English

Tables		Languages	
7	9	Spanish	English
1	1 to 6	**La neología contrarreloj: Internet**	**Neology against the clock: the Internet**
2	7 to 22	El propósito de esta comunicación es hacer una reflexión sobre los retos a que se está enfrentando la neología terminológica en la realidad actual	This paper is intended to look at the challenges faced by neology in terminology at the present time
3	23 to 38	para lo cual vamos a abordar diversos aspectos que influyen en la creación neológica en el ámbito de Internet	I will do this by discussing various points which influence neology in the field of the Internet
4	39 to 67	Los términos referidos a Internet nacen y se difunden a una velocidad y con una amplitud tal que constituye una verdadera carrera contrarreloj en las distintas lenguas	Terms referring to the Internet are coined and spread at such speed and to such an extent that they have turned into a race against the clock in different languages
5	68 to 92	Efectivamente, la formación de nuevos términos está sometida a un ritmo trepidante, paralelo al avance e innovación tecnológica en el sector de la informática y, en general, de las telecomunicaciones	The formation of new terms goes on at a dizzy speed, parallel to technological advances and innovations in the field of computer science and telecommunications in general
6	93 to 105	Si bien este aspecto es común al progreso científico y técnico y, por lo tanto, característico de la neología terminológica	This is common in all scientific and technological progress, and therefore characteristic of neology in terminology
7	106 to 123	la especificidad del área tratada confiere a la neología que le es propia unas particularidades que cabe tener en cuenta	but the specific nature of this area confers particular features on neology which must be taken into account
8	124 to 164	En primer lugar, el canal por el que se dan a conocer los términos de Internet, la misma red, no sólo supone una rápida difusión de la terminología—la información en Internet es de acceso (casi) inmediato—, sino también un alcance muy vasto—llega a cualquier parte del mundo—	First of all the channel through which Internet terms are made known is the net itself. This means that they not only spread rapidly (information on the internet can be accessed almost immediately) but also reach vast areas (all over the world)
9	165 to 173	Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados	Furthermore, terms can be compiled, discussed and assessed anywhere

Table 6 continued

Tables		Languages	
7	9	Spanish	English
10	174 to 196	de ahí, por ejemplo, los apartados que encontramos en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar	many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them
11	197 to 203	Esto nos lleva a una cuestión fundamental	This leads us to the fundamental point
12	204 to 224	la terminología de Internet traspasa los límites del área de especialidad (a la que se circunscribe por definición el léxico científico y técnico)	Internet terminology extends beyond the bounds of its specialist field (which by definition is part of the lexicon of science and technology)
13	225 to 229	e irrumpe en la lengua de uso general	and breaks into general language
14	230 to 256	siendo utilizada tanto por los usuarios heterogéneos de la red (de cualquier o ninguna especialidad) como por las personas que leen la prensa o están atentas a los medios de comunicación	It is used both by a wide variety of net users (from any or no specialist fields) and by people who read the press or follow the media
15	257 to 262	¿Qué tipo de terminología se está creando?	What type of terminology is being created?
16	263 to 267	¿Qué sistemas de creación léxica predominan?	What lexical creation systems predominate?
17	268 to 273	Un único denominador común existe para todas las lenguas	There is a common denominator in all languages
18	274 to 278	los términos se generan en inglés	terms are generated in English
19	278 to 281	y penetran como préstamos en aquellas	and come in as loanwords
20	282 to 289	¿Cómo responden las lenguas receptoras?	How do the receiving languages respond to this?
21	290 to 296	¿Cómo tratan la terminología de Internet?	How do they deal with Internet terminology?
22	297 to 307	¿Son términos todos los que lo parecen	Are all those words which seem to be terms actually terms?
23	308 to 314	responden a necesidades reales de denominación	Do they meet actual needs for names
24	315 to 320	o abundan las creaciones léxicas sensacionalistas y efímeras?	or do sensationalist, ephemeral terms abound?

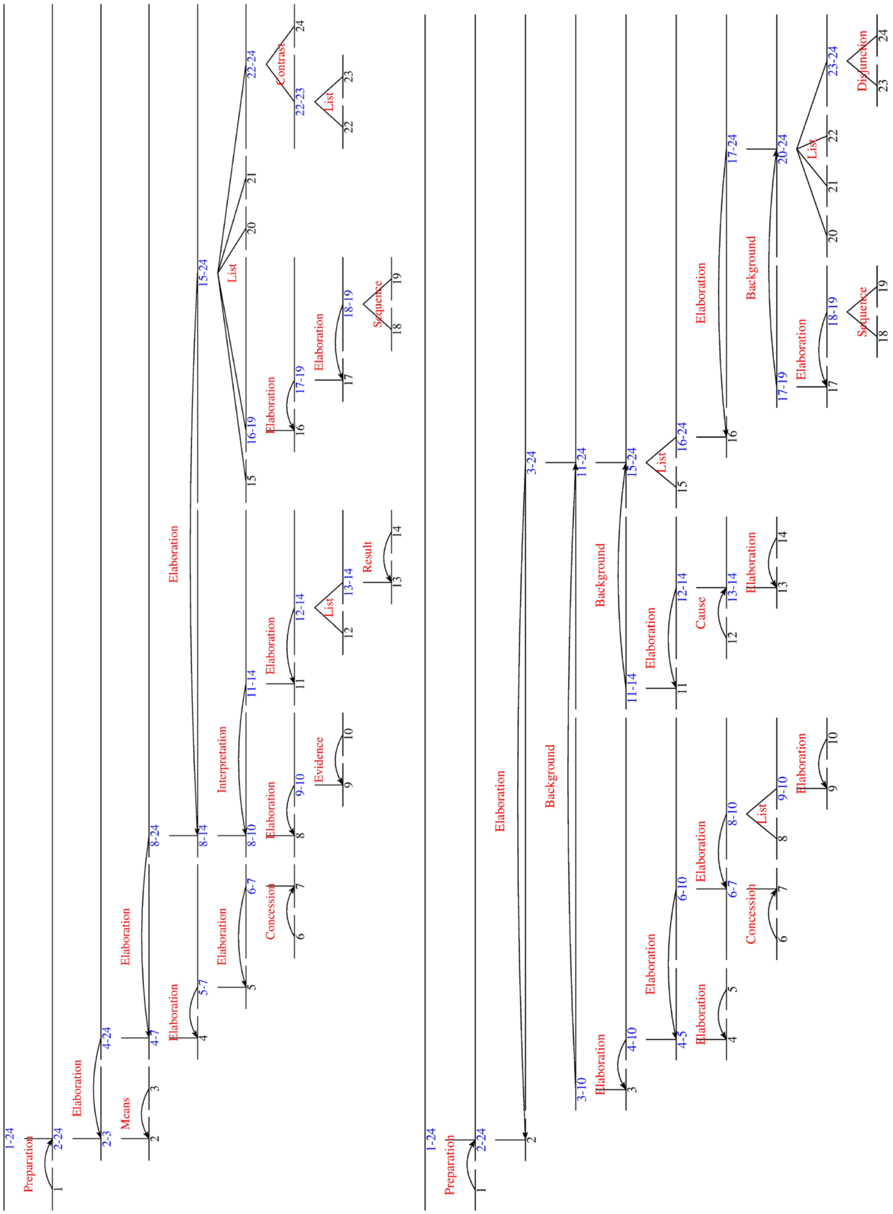


Fig. 7 Rhetorical tree elaborated by A2 (Spanish) and A1 (English), TERM38_SPA

In order to better highlight the differences between the quantitative method and our qualitative proposal, we have kept the rhetorical structure, but have used one of the languages to compare using RSTeval in contingency Table 9.

Table 7 Qualitative evaluation matrix TERM38_SPA

L	ENG				SPA				Qualitative evaluation					
	CS(s)	R	C	A	CS(s)	R	C	A	N	R	C	A	*Agree*	*Disagree*
1	1	Preparation→	1S	2–24N	1	Preparation→	1S	2–24N	✓	✓	✓	✓	RCA	
2	3	Means←	3S	2N	3	Background→	3–10S	11–24N	×	×	×	×		$N/S[A]$
3	4	Elaboration←	4–24S	2–3N	4	Elaboration←	4–10S	3N	✓	✓	×	×	R	
4	5	Elaboration←	5–7S	4N	5	Elaboration←	5S	4N	✓	✓	×	✓	RA	
5	7	Elaboration←	6–7S	5N	7	Elaboration←	6–10S	4–5N	✓	✓	×	×	R	
6	6	Concession→	6S	7N	6	Concesion→	6S	7N	✓	✓	✓	✓	RCA	
7	9	Elaboration←	9–10S	8N	8|9	List↔	9–10N	8N	×	×	✓	✓	(CA)	$N/NversusN/S[A]$
8	10	Evidence←	10S	9N	10	Elaboration←	10S	9N	✓	×	✓	✓	(CA)	$MC[L]$
9	11	Interpretation←	11–14S	8–10N	11	Background→	11–14S	15–24N	×	×	✓	×	(A)	$N/S[A]$
10	12	List↔	12N	13–14N	12	Cause←	12S	13–14N	×	×	✓	✓	(CA)	$N/NversusN/S[A]$
11	14	Result←	14S	13N	14	Elaboration←	14S	13N	✓	×	✓	✓	(CA)	$CSC[L]$
12	15|16–24	List↔	15N	16–24N	15|16–24	List↔	15N	16–24N	✓	✓	✓	✓	RCA	
13	15|16|20–24	Elaboration←	15–24S	8–14N	15|16|20–24	Elaboration←	3–24S	2N	×	✓	×	✓	R	
14	16|20|21|22–24	List↔	20–24N	16–19N	20|21|22–24	Elaboration←	17–24S	16N	×	×	×	×		$N/S[A]$
15	17	Elaboration←	17–19S	16N	17	Background→	17–19S	20–24N	×	×	✓	×	(A)	$N/S[A]$
16	18–19	Elaboration←	18–19S	17N	18–19	Elaboration←	18–19S	17N	✓	✓	✓	✓	RCA	
17	18|19	Sequence↔	18N	19N	18|19	Sequence↔	18N	19N	✓	✓	✓	✓	RCA	
18	20|21–24	List↔	20N	21–24N	20|21–24	List↔	20N	21–24N	✓	✓	✓	✓	RCA	
19	21|22–24	List↔	21N	22–24N	21|22–24	List↔	21N	22–24N	✓	✓	✓	✓	RCA	
20	22|23–24	List↔	22N	23N	22|23–24	List↔	22N	23–24N	✓	✓	✓	✓	RCA	
21	22–23|24	Contrast↔	22–23N	24N	23|24	Disjunction↔	23N	24N	✓	×	×	✓	(C)	$\neq R[A]$
22	8	Elaboration←	8–24S	4–7N	8|9	Elaboration←	8–10S	6–7N	✓	✓	×	×	R	
23	12|13	Elaboration←	12–14S	11N	13	Elaboration←	12–14S	11N	✓	✓	✓	✓	RCA	

Table 8 Qualitative evaluation results for the example in Fig. 7, TERM38_SPA

Nuclearity		Relation		Composition		Attachment	
Matches	F1	Matches	F1	Matches	F1	Matches	F1
16 of 23	0.6957	14 of 23	0.6087	15 of 23	0.6522	16 of 23	0.6957

Both methods measure the similar factors: (1) EDUs and spans (constituent and attachment), (2) nuclearity (of each unit, or direction of the relation) and rhetorical relations (of each unit: relation plus span, or relation as a whole). Thus, in Table 11 we can compare how each method accounts for these factors.

In Table 11 both methods describe total agreement in segmentation. This is due to the fact that segmentation was harmonized before the analysis was undertaken. The span factor of the quantitative method is described using factors C and A, this factor being more positive in the quantitative method. In terms of nuclearity and rhetorical relations, the qualitative method is able to describe more agreements in the evaluation of text TERM38.

In Table 12 we can observe further detail on how both methods describe agreement in relations, and the weight given to each relation in the calculation of agreement. To better understand the table, we have highlighted in italicize the most important differences.

As we can see in Table 12, an important part of the agreement in quantitative evaluation method is captured in the SPAN label (which is not an RST relation). In addition, the contingency table shows that the relation with most agreement is the LIST relation, followed by ELABORATION and SEQUENCE. Thanks to the qualitative evaluation, however, we can see that the ELABORATION relation actually has a higher degree of agreement, followed by LIST. In contrast, SEQUENCE has little importance, the same as CONCESSION and PREPARATION. We would like to point out that the difference is more striking when describing agreement (Match: columns 4 and 8), rather than when describing how often the annotator has used such relation (A1: columns 2 and 6, and A2: columns 3 and 7). For instance, in both methods we can see that A1 has used 10 ELABORATION relations, whereas A2 has used 9 relations. The quantitative method captures an agreement of 4.35 %, while the qualitative method throws a much higher agreement, reaching 26.09 %.

The root of this difference can be found in the fact that the quantitative evaluation does not evaluate nuclearity and rhetorical relations in an independent way. When creating relation pairs, the pairs do not have well-formed members (in particular because of the use of the SPAN label). This is the reason why in the quantitative method, out of 10 ELABORATION relations, only two of them show agreement.

Advantages of the qualitative evaluation method. The formalization of qualitative evaluation (Table 7) describes the annotation agreement (Agree) in a more complete way than quantitative evaluation (Table 9): the relation factor (R) is compared in an isolated manner, that is, nuclearity is not reanalyzed in the relation factor. This fact has methodological implications and some of advantages are shown in contingency Table 7:

Table 9 Contingency table for text TERM38_SPA with quantitative method, using *RSTeval*

Constituent	Units		Spans		Nuclearity		Relation		Constituent	Units		Spans		Nuclearity		Relation	
	A1	A2	A1	A2	A1	A2	A1	A2		A1	A2	A1	A2	A1	A2	A1	A2
1 to 6	x	x	x	x	s	s	Preparation	Preparation	!268 to 281			x	x	s	s	Background	Elaboration
7 to 22	x	x	x	x	n	n	Span	Span	!263 to 281				x		n		List
23 to 38	x	x	x	x	n	s	Span	means	257 to 262	x	x	x	x	n	n	List	List
!7 to 38				x		n		Span	282 to 289	x	x	x	x	n	n	List	List
39 to 67	x	x	x	x	n	n	Span	Span	!257 to 289				x		n		List
68 to 92	x	x	x	x	s	n	Elaboration	Span	290 to 296	x	x	x	x	n	n	List	List
93 to 105	x	x	x	x	s	s	Concession	Concession	!257 to 296				x		n		List
106 to 123	x	x	x	x	n	n	Span	Span	297 to 307	x	x	x	x	n	n	List	List
!93 to 123			x	x	n	s	Span	Elaboration	308 to 314	x	x	x	x	n	n	Disjunction	List
!68 to 123				x		s		Elaboration	!297 to 314				x		n		Contrast
!39 to 123				x		n		Span	315 to 320	x	x	x	x	n	n	Disjunction	Contrast
124 to 164	x	x	x	x	n	n	List	Span	!297 to 320				x		n		List
165 to 173	x	x	x	x	n	n	Span	Span	!257 to 320			x	x	n	n	Span	List
174 to 196	x	x	x	x	s	s	Elaboration	Evidence	!263 to 320			x	x	n	s	List	Elaboration
!165 to 196			x	x	n	s	List	Elaboration	!124 to 320				x		s		Elaboration
!124 to 196			x	x	s	n	Elaboration	Span	!39 to 320				x		s		Elaboration
197 to 203	x	x	x	x	n	n	Span	Span	!7 to 320			x	x	n	n	Span	Span
204 to 224	x	x	x	x	s	n	Cause	List	!1 to 320			x	x	r	r	Span	Span
225 to 229	x	x	x	x	n	n	Span	Span	!39 to 92			x		n		Span	
230 to 256	x	x	x	x	s	s	Elaboration	result	!93 to 196			x		s		Elaboration	
!225 to 256			x	x	n	n	Span	List	!39 to 196			x		s		Elaboration	
!204 to 256			x	x	s	s	Elaboration	Elaboration	!23 to 196			x		s		background	
!197 to 256			x	x	s	s	Background	Interpretation	!282 to 296			x		n		List	

Table 9 continued

Constituent	Units		Spans		Nuclearity		Relation		Constituent	Units		Spans		Nuclearity		Relation	
	A1	A2	A1	A2	A1	A2	A1	A2		A1	A2	A1	A2	A1	A2	A1	A2
!124 to 256				x		n		Span	!282 to 307			x		n		List	
263 to 267	x	x	x	x	n	n	Span	Span	!308 to 320			x		n		List	
268 to 273	x	x	x	x	n	n	Span	Span	!282 to 320			x		n		Span	
274 to 277	x	x	x	x	n	n	Sequence	Sequence	!268 to 320			x		s		Elaboration	
278 to 281	x	x	x	x	n	n	Sequence	Sequence	!197 to 320			x		n		Span	
!274 to 281			x	x	s	s	Elaboration	Elaboration	!23 to 320			x		s		Elaboration	

Table 10 Quantitative method results for text TERM38_SPA

Units		Span		Nuclearity		Relation	
Match	*F*1	Match	*F*1	Match	*F*1	Match	*F*1
24 of 24	1	36 of 47	0.766	29 of 47	0.617	20 of 47	0.425

Table 11 Comparison using both methods, TERM38_SPA

	Units		Spans				Nuclearity		Relation	
Quanti.	24 of 24	1	37 of 46		0.8043		29 of 46	0.6304	21 of 46	0.4565
	Units		Composition		Attachment		Nuclearity		Relation	
Quali.	24 of 24	1	15 of 23	0.6522	14 of 23	0.6087	17 of 23	0.7391	13 of 23	0.5652

Table 12 Comparison of agreement using both methods for text TERM38

Relation	Quantitative method				Qualitative method			
	A1	A2	Match	%	A1	A2	Match	%
Background	3				3			
Cause	1				1			
Concession	1	1	1	2.17	1	1	1	4,35
Contrast		2				1		
Disjunction	2				1			
Elaboration	10	9	*2*	4.35	10	9	*6*	26,09
Evidence		1				1		
Interpretation		1				1		
List	10	12	*6*	13.04	5	6	*4*	17,39
Means		1				1		
Preparation	1	1	1	2.17	1	1	1	4,35
Result		1				1		
Sequence	2	2	2	4.35	1	1	*1*	4,35
Span	16	15	*9*	19.57	–	–	–	–
Total	46	46	21	45.65	23	23	13	56,52

1. Independent factors are evaluated. A different attachment point of a relation only implies disagreement in attachment point (disagreement described at the same line) and in constituent (disagreement described at a higher level in the tree structure) and not in relation as quantitative method does. Moreover, the qualitative method accounts for the source of disagreement (Disagree).

2. Only rhetorical relations are compared. The description allows for a full coincidence in structure (RCA), or a partial match (RA, RC or R).
3. Reasons for annotator disagreement are captured: *a*) because of differences in the linguistic expression [L] or *b*) because of interpretation [A].
4. Relation pairs in the contingency table are able to better describe agreement and disagreement ("confusion patterns", Marcu 2000a).

For example, in Table 7 we can observe the following types of information on the relation agreement:

1. Match in relation, constituent and attachment point (RCA) in the following nine lines: 1, 6, 12, 16, 17, 18, 19, 20 and 23. We observe that in these lines there was total agreement in the three factors observed, that is, for example, in line 1 an agreement in all factors: same CS (1), relation (PREPARATION), constituent (1S) and attachment point (2–24N).
2. Match in relation and attachment point (RA) in line 4. A partial agreement, but in this case in CS (5), relation (ELABORATION) and attachment point (4N). By contrast, slight disagreement in constituent (A2: 5–7S but A1: 5S).
3. Match only in relation (R) in four lines: 3, 5, 13 and 22. For example, in line 3 there was an agreement only in CS (4) and relation (ELABORATION), whereas there were discrepancies in constituent (A2: 4–24S but A1: 4–10S) and attachment point (A2: 2–3N but A1: 3N).

On the relation disagreement, we can observe the following types of information in Table 7:

1. A different choice in nuclearity (N/S [A]) in four lines: 2, 9, 14 and 15.
2. A N/N–N/S mix-up (N/N–N/S [A]) in two lines: 7 and 10.
3. A different relation label (≠ R [A]) in a line: 21.
4. A Marker Change (MC [L]) in a line: 8.
5. A Clause Structure Change (CSC [L]) in a line: 11.

3 Results

In this section, we first present the results of segmentation, and then we compare the results of rhetorical structure based on two evaluation methods: quantitative method (Marcu 2000a) and our new proposal, a qualitative evaluation method.

3.1 Discourse segmentation results

The initial round of segmentation led to the following number of EDUs: 330 in English, 318 in Spanish, and 323 in Basque. We calculated agreement using F-score and Kappa, in a pairwise manner. First of all, we calculated the total coincidence of EDUs, using the verb of the main clause and its principal arguments (VP). If the main verb was the same in both EDUs, then we tabulated it as a match. As we stated in page 7, one of our segmentation principles is that every EDU should contain a

Table 13 Segmentation agreement

Language	Correct	Match	Wrong	Missing	Candidates	F-measure	Kappa
ENG-SPA	330	230	88	12	731.4	70.99	0.7139
ENG-BSQ	330	226	97	7	742.9	69.22	0.7057
BSQ-SPA	323	230	88	5	731.4	71.76	0.7333

finite verb. The main verb of an EDU indicates the principal action, process, state, condition, etc., in relation to the subject of the clause. Therefore, if two EDUs in different languages contain the same verb (that is, both verbs are translation equivalents), they are expressing the same event and we consider that there is coincidence between EDUs. Thus, in this sense, syntax has an important role to play in the detection of the EDUs to be compared, since we take the main verb of the clausal syntactic structure in each language to carry out the comparison. In this work, we have not used a syntactic parser to perform the analysis. We have done the analysis manually, because it was feasible to do it over our corpus and we also wanted to avoid possible mistakes in the harmonization work.[14] In future work, however, we plan to automate our methodology to compare discourse structures, and, in this case, we could integrate a syntactic parser in the system. We then calculated F-measure and Kappa as presented in Table 13.[15]

3.1.1 *Discourse segmentation harmonization*

In our segmentation, it was often the case that one language used a finite verb, whereas the other language used a non-finite verb or other expression, leading to differences in segmentation. Another source of disagreement was the interpretation of ellipsis, where one annotator decided there was more than subject ellipsis in coordination, and did not break up the two VPs, whereas the other annotator decided to break them up. Two other sources of disagreement were different texts in the two languages (not different formulations, but a completely different text, with one sentence deleted or inserted), and simple human error. The latter accounts for no more than two disagreements per language pair.

Harmonization led to joining or separating EDUs in one of the languages, contravening our general principles for segmentation. The main changes in this harmonization were:

1. When two parallel passages share the same structure and the third passage does not, then we harmonize the segmentation of the third language taking into account the segmentation of the two coincident languages.
2. When the segmentations of the three parallel passages are different, then we harmonize the segmentation taking into account the structure of the simplest passage.

---

[14] This harmonization work can be found at http://ixa2.si.ehu.es/rst/segmentuak_multiling.php.

[15] For Kappa segment candidates were calculated automatically by counting verbs.

In Example (1) a Basque conjunct was translated as a clause in both English and Spanish. In the English example there are three finite verbs (all three of them instances of the verb *is*), as is the case in Spanish (*es*, '[it] is'; *se ubica,* '[it] is located'; and *va*, '[it] goes'). In Basque, however, there are only two finite verbs (*estrapolatuko du,* '[it] will extrapolate [it]'; and *jartzen du,* '[it] places [it]'). The third part of the conjunct contains no verb (*eta hizkuntza erromanikoek ezkerraldean,* 'and the Romance languages on the left side'). In the harmonization we inserted a new segment in Basque, reinterpreting not as coordinated NP, but as a juxtaposed clause with an elided verb.[16]

(1)
(a) [Our hypothesis is that a syntactic characteristic of Basque and the romance languages is extrapolated to their morphology,] [so that in Basque derivations the core of the structure is on the right,] [while in the romance languages it is on the left.]
(b) [Nuestra hipótesis es que una característica sintáctica del euskera y de las lenguas románicas se extrapola hasta la morfología,] [de manera que en euskera, también en derivación, el núcleo de la estructura se ubica a la derecha,] [mientras que en las lenguas románicas va a la izquierda.]
(c) [Gure hipotesiak, euskararen eta hizkuntza erromanikoen ezaugarri sintaktiko bat morfologiaraino estrapolatuko du:] [eratorpenean ere euskarak egituraren burua edo gunean eskuinaldean jartzen du,} {eta hizkuntza erromanikoek ezkerraldean.] TERM50_BSQ

In Example (2) the translation from Spanish into English has led to two separate clauses. The Spanish original segmentation contained only one span, since the first idea (*un aumento cuantitativo de la terminología especializada,* 'an increase in the number of specialist terms') is embedded in a non-finite clause (*además de provocar,* 'in addition to leading to'). The English translation splits the ideas into two coordinated clauses (*factors lead to an increase* and *but also [factors] call into question*). Basque also has two clauses to express these two ideas. Since two of the languages divided this sentence into two clauses, in the harmonization we inserted a new boundary in Spanish.

(2)
(a) [All these factors lead to an increase in the number of specialist terms which enrich terminology] [but also call into question some of its basic concepts, such as the one to one relationship between ideas and names, the concept of mastery of a specialist field and the role of standardization in terminology.]
(b) [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,} {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos, como la univocidad noción-denominación, el concepto de dominio de especialidad o el papel mismo de la normalización en terminología.]

---

[16] In the example, the original segmentation is marked with square brackets and the segmentation after harmonization with curly brackets.

(c) [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;] [eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan jarri ditu terminologiaren oinarrizko zenbait kontzeptu: kontzeptu-izendapen bikotearen adierabakartasuna, espezialitateko eremuen kontzeptua, eta normalizazioak terminologian duen eginbeharra.] TERM19_SPA

We quantified the changes necessary to harmonize the segmentations by counting how many times a change was necessary, per language. Table 14 summarizes those changes (the typical actions are "join" or "break up"), and the number of affected EDUs. To compute the number of affected EDUs, we counted, in the cases where we needed to break down a unit, how many new units were necessary (+). In the cases where we needed to join, we counted how many original units were integrated (−). In the table, "initial spans" refers to the spans proposed by the individual annotator for each language, and "affected spans", to the number of spans that underwent a change, whether to join, or to break up. "Harmonized spans" represents the final agreed upon spans across all three languages, for each text.

We can see from the table that the language with more changes is Basque.[17] We found that the linguistic expression of the same or similar concepts required different syntactic constructions in Basque. This makes sense, given that Basque is a non-Indo-European language, showing considerable typological distance from both Spanish and English (Cenoz 2003). Note that, whereas Spanish and Basque were affected in the same proportion in both directions (when breaking down SPA: 44.44 % and BSQ: 41.46 %; when joining SPA: 55.56 % and BSQ: 58.54 %), harmonization in English involved breaking down in a much lower proportion (when breaking down ENG: 18.18 %; when joining ENG: 81.82 %). This suggest that the corpus abstracts in English (whether translated or original) express clauses as separate units, either as simple sentences or as clear (finite) adjunct clauses, without using non-finite clauses or prepositional complements.

3.2 Rhetorical analysis results

Results of quantitative method were presented in order to show the consistency of this method. To this end, first, we present below the results of the quantitative method; second, we present the results of the qualitative method, and after that we compare results from both methods.

3.2.1 Results of the quantitative evaluation method

Results of the quantitative evaluation are shown in Table 15.[18]

---

[17] One-way ANOVA demonstrated significant differences across the three languages in the corpus ($p = 0.07$). We thought this was quite significant, therefore we performed a post-hoc Tukey's test and we observed that harmonization in Basque is the furthest from the other two.

[18] EDUs are excluded because they are identical after harmonization.

Table 14 Segmentation changes

Text	Initial spans			Harmon.	Affected spans		
	ENG	SPA	BSQ	Spans	ENG	SPA	BSQ
TERM18_ENG	8	11	14	8	0	−3	−6
TERM19_SPA	14	12	13	14	0	+2	+1
TERM23_ENG	15	14	14	14	−1	0	0
TERM25_BSQ	10	11	8	10	0	+1	+2
TERM28_BSQ	16	14	12	15	−1	+1	+3
TERM29_SPA	14	14	13	14	0	0	+1
TERM30_ENG	26	27	33	28	+2	+1	−5
TERM31_BSQ	53	52	44	52	−1	0	+8
TERM32_ENG	13	13	18	13	0	0	−5
TERM34_BSQ	50	45	44	46	−4	+1	+2
TERM38_SPA	27	25	28	24	−3	−1	−4
TERM39_ENG	7	8	9	9	+2	+1	0
TERM40_SPA	8	8	8	8	0	0	0
TERM50_BSQ	34	35	30	30	−4	−5	0
TERM51_SPA	35	29	35	31	−4	+2	−4
Total	330	318	323	316	±22	±18	±41
Change rate					6.67 %	5.66 %	12.69 %

Table 15 Quantitative evaluation results (F-measure)

Language comparison		Evaluation		
1st Lang.	2nd Lang.	Span (%)	Nuclearity (%)	Relation (%)
ENG	SPA	84.06	67.43	56.22
ENG	BSQ	86.22	68.24	53.28
SPA	BSQ	88.61	71.02	54.94

Surprisingly, results for the quantitative evaluation are slightly better when Basque is involved in the comparison, which was not the case for the segmentation Span agreement results (Table 14). Agreement, however, is higher for the Nuclearity criterion when Basque is included (also the case for Span agreement results shown earlier). Finally, the Relation agreement drops when Basque is involved. We point out the source of this change and we discuss the results of the Relation comparison in Sect. 2.4.2, where we present the final results of both evaluation methods (Table 21).

3.2.2 *Results of qualitative evaluation method*

Table 16 and Table 17 include the final results for the entire corpus, which account for agreement and disagreement in a qualitative way. In Table 16 results from the agreement level obtained on the four types of measurements increases as the relaxation of the agreement increases too, being RCA the most demanding agreement, and R the more relaxed one.

In Table 18 we show summarized results of the three sources: total agreement between annotators (Agreement), discrepancies because of annotation decisions (Annotation Discrepancies) and discrepancies because of linguistic differences (Translation Strategies).

As we observe in Table 18, the disagreement is higher when data of both A1 (English) and A2 (Spanish) are compared with A3 (Basque). That could be, as we have interpreted from the results of Table 14, because English and Spanish are typologically closer to each other than Basque is to either English or Spanish (Cenoz

Table 16 Qualitative evaluation results (F-measure): analysis of the sources of agreement

Classification		ENG-SPA		ENG-BSQ		SPA-BSQ	
		%	Gain (%)	%	Gain (%)	%	Gain (%)
Agreement	RCA	44.67		40.33		42.33	
	RC	49.34	4.67	42.66	2.33	45.66	3.33
	RA	51.67	7	48.66	8.33	50.66	8.33
	R	59.67	3.33	54.66	3.67	56.99	3

Table 17 Qualitative evaluation results (F-measure): analysis of the sources of disagreement

Classification		ENG-SPA (%)	ENG-BSQ (%)	SPA-BSQ (%)
Annotator-based discrepancies	Nuclearity	4.00	4.00	3.33
	N/N versus N/S	5.33	8.00	6.00
	Attachment span	2.00	1.33	0.67
	Relation	6.67	4.00	2.67
	Similar relation	1.67	4.33	6.67
	Mismatched relation	6.00	4.67	5.67
	Specificity	0.67	4.33	5.33
	No Match	6.33	6.67	4.67
Language-based discrepancies	Marker change	4.67	3.33	4.67
	Clause structure	1.67	1.67	1.33
	Unit shift	1.33	2.67	1.67

Table 18 Qualitative evaluation results (F-measure): summary of results

Classification	ENG-SPA (%)	ENG-BSQ (%)	SPA-BSQ (%)
Agreement	59.67	54.66	56.99
Annotator-based discrepancies	32.67	37.33	35.01
Language-based discrepancies	7.67	7.67	7.67

2003). But this dispersion is not so large if we take into account the fact that there are more Similar Relations and Specificity when A3's data is compared with A1's and A2's.

After aligning the contingency tables of the qualitative evaluation from all the RS-structure in English, Spanish and Basque, we measured the agreement of rhetorical relations with Fleiss Kappa (see Table 19) for assessing the reliability of agreement between more than two annotators. The agreement attained across the three annotators was moderate with a Kappa (Fleiss 1971) score of 0.484 (300 rhetorical relations, 15 texts). We show in Table 19 the agreement relation by relation between the three annotators.

As we observe in Table 19, Fleiss' Kappa measures show different degrees of understanding rhetorical relations.

1. Almost perfect: PREPARATION.
2. Substantial: SUMMARY and CONCESSION.

Table 19 Qualitative evaluation results (Fleiss' Kappa) for rhetorical relations

Relation	Kappa	z	*p* value
Preparation	0.851	25.528	0.000
Summary	0.712	21.361	0.000
Concession	0.705	21.155	0.000
List	0.554	16.629	0.000
Elaboration	0.531	15.933	0.000
Condition	0.525	15.763	0.000
Sequence	0.499	14.966	0.000
Restatement	0.424	12.723	0.000
Background	0.420	12.589	0.000
Circumstance	0.420	12.586	0.000
Contrast	0.376	11.272	0.000
Cause	0.352	10.552	0.000
Purpose	0.335	10.057	0.000
Result	0.301	9.017	0.000
Means	0.221	6.617	0.000
Conjunction	0.172	5.151	0.000
Motivation	0.136	4.084	0.000
Interpretation	0.080	2.390	0.017
Solutionhood	−0.011	−0.337	0.736
Justify	−0.009	−0.269	0.788
Antithesis	−0.008	−0.235	0.814
Evidence	−0.008	−0.235	0.814
Evaluation	−0.003	−0.100	0.920
Disjunction	−0.001	−0.033	0.973
Unless	−0.001	−0.033	0.973

3. Moderate agreement: LIST, ELABORATION, CONDITION, SEQUENCE, RESTATEMENT, BACKGROUND and CIRCUMSTANCE.
4. Fair agreement: CONTRAST, CAUSE, PURPOSE, RESULT and MEANS.
5. Slight agreement: CONJUNCTION, MOTIVATION and INTERPRETATION.
6. No observed agreement for: ANTITHESIS, DISJUNCTION, EVALUATION, EVIDENCE, JUSTIFY, SOLUTIONHOOD and UNLESS.[19]

Translation Strategies. In carrying out the comparison of rhetorical structures, we observed some language differences. Some of them were produced when authors translated from one language into another (translation strategy),[20] and others were the result of comparing rhetorical structure in a pairwise manner, for instance in comparing English and Spanish with each other, when they are both translations of a Basque source. The latter cannot be regarded as translation strategies, so we will include only the first types under the umbrella term 'translation shift'. And the second type under the umbrella 'different language forms'.

On the one hand, we do not analyze translation strategies which do not lead the annotator to choose a different relation, as in Example (3); where in Basque the rhetorical relation was made explicit with the marker (*izan ere,* 'in fact'), but remains the same relation, a CAUSE relation is in the A1 analysis.[21]

(3)

(a) [In the recent past, a trend has been noted, and reported by many researchers in the area of Serbian scientific terminology, of importing borrowings of lexical and larger structural units from English into specific scientific registers, rather that to opt for translations, calques, etc.]$_{3N}$ [This corresponds closely to the fact that a consensus has been reached among Serbian scientists of various orientations regarding the status of English as the only language of scientific communication in the last several decades.]$_{4S-CAUSE}$

(b) [Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unitate lexikalen maileguak eta unitate-egitura luzeagoen maileguak hartzen dira zientzia-erregistro zehatz baterako, itzulpenak edo kalkoak egin ordez.]$_{3N}$ [Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingelesari eman diote zientzia-komunikaziorako hizkuntza bakarraren estatusa.]$_{4S-CAUSE}$ TERM18_ENG

---

[19] "Values of agreement between $-A\_e/1-A\_e$ (no observed agreement) and 1 (observed agreement = 1), with the value 0 signifying chance agreement (observed agreement = expected agreement)." (Artstein and Poesio 2008, p. 559).

[20] Catford (1965, pg. 73) defines translation shifts as "departures from formal correspondence in the process of going from the SL to the TL" (from the Source Language to the Target Language). Chesterman (1997) states that changes from original to translated text are due to a translation strategy.

[21] Note that here there is another translation strategy (CSC hierarchical upgrading in Basque with a coordination of two finite verbs *lortu dute* '[they] achieve [it]' and *eman diote* '[they] give [him]'), which is not under consideration due to harmonization process.

On the other hand, we do analyze all the directions (ENG > SPA, ENG > BSQ and so on) in Table 20 and three types of translation differences that influence rhetorical relations and reveal local translation strategies:

1. Relation signaling has a different configuration (Marker Change). Within Marker Change, we found three subtypes:
   (a) inclusion of a marker,
   (b) exclusion of a marker, and
   (c) changing a marker.
2. Differences because of the use of a distinct language configuration (Clause Structure Change):
   (a) hierarchical downgrading, and
   (b) hierarchical upgrading.
3. Punctuation is used differently (Unit Shift):
   (a) an independent sentence is integrated in another sentence, and
   (b) a clause is translated in an independent sentence. We detail some of them below.

1. **Marker Change**. In Example (4) a discourse maker (*de ahí*, 'hence') was not translated from Spanish into either English or Basque. In English the marker *por ejemplo* ('for example') was also elided and the punctuation changed (from semicolon into colon). This is why annotators in English and Basque labelled the relation ELABORATION; whereas in Spanish, the marker *de ahí* ('hence') resulted in an annotation with the evidence label.

(4)

(a) [Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados;]$_{9N}$ [de ahí, por ejemplo, los apartados que encontramos en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar.]$_{10S-EVIDENCE}$

(b) [Furthermore, terms can be compiled, discussed and assessed anywhere:]$_{9N}$ [many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.]$_{10S-ELABORATION}$

(c) [Are gehiago, edozein tokitatik biltzen dira terminoak, baita komentatu eta haztatu ere;]$_{9N}$ [adibidez, Interneti buruzko terminoen glosarioak zabaltzen dira Web askotan, eta izendegietarako proposamenak egin ere bai, eta erabiltzaileek botoa eman ahal izaten diete.]$_{10S-ELABORATION}$ TERM38_SPA

2. **Clause Structure Change**. In Example (5) the clauses under the relative used in the original Spanish text were avoided in the same way in English and in

Basque (*que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos,* 'that, although [it] has enriched it, [it] has also called into question some of its basic concepts'), in favour of an adversative coordination using a finite verb in English (*but*), and a conjunction coordination (*eta,* 'and') and a finite verb in Basque (*jarri ditu*, '[it] places [them]'). That was the reason for A1 to annotate a CONTRAST relation, whereas A3 annotated a LIST relation. The relative form[22] analyzed here is a product of the harmonization and it was annotated by A2 as an ELABORATION relation.

(5)

(a.) [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,}$_{6N}$ {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos (...)]$_{7-11S-ELABORATION}$[23]

(b.) [All these factors lead to an increase in the number of specialist terms which enrich terminology]$_{6N-CONTRAST}$ [but also call into question some of its basic concepts (...)]$_{7N-CONTRAST}$

(c.) [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;]$_{6N-LIST}$ [eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan jarri ditu terminologiaren oinarrizko zenbait kontzeptu (...)]$_{7N-LIST}$ TERM19_SPA

3. **Unit Shift**. A different punctuation can lead the annotator to interpret a different relation. In the original text in Spanish in Example (6), the spans were linked with comma, whereas in the English text the punctuation was changed, using a period. The punctuation led A1 to consider a hypotactic relation between the first and the following two spans.

(6)

(a) [En esta comunicación, a partir de la experiencia en trabajos de normalización de terminología catalana, se planteará la necesidad social de la normalización terminológica,]$_{N12-LIST}$ [se comentarán algunas de las dificultades con que se enfrenta y se apuntarán ideas para su enfoque dentro de la sociedad actual.]$_{N13-14-LIST}$

(b) [This paper looks, on the basis of experience in the standardisation of terminology in Catalan, at the social need for standardisation of terminology.]$_{N12}$ [Some of the difficulties faced will be discussed, and

---

[22] Again, this goes against the principles of our segmentation.

[23] Note here the human annotation error which does not follow the modular and incremental annotation that Pardo (2005) proposes.

> ideas will be given for approaching this field in present day society.]$_{S13-14-ELABORATION}$ TERM19_SPA

We present, in Table 20, the influence of translation strategies and different language forms more in depth.

It is worth mentioning that when English is the SL there are not so many translation strategies (10.14 %) as when other languages are SL (Spanish: 23.19 % and Basque: 34.78 %). Another interesting aspect is that the Marker Change translation strategy is the most prominent one (MC: 34.78 % versus CSC: 15.94 % and US: 17.39 %), and changes in discourse markers have an influence on rhetorical annotation.[24] These results are merely describing tendencies, because the corpus is not big enough (although is comparable to other corpora in the literature, such as Scott et al. (1998)). The results are sensitive to segmentation granularity or harmonization decisions and to text characteristics (genre and domain). However what is relevant is that the method presented here can describe and quantify translation strategies.

3.2.3 Comparing quantitative and qualitative methodologies

To determine whether the proposed method is consistent, we compare the quantitative results of the relation factor from both methods in Table 21. In this table, we present the final results from both evaluation methods, providing the F-measure of relation factor.

We can highlight two findings in this comparison:

1. The qualitative method finds slightly higher agreement than the quantitative method. The difference goes from almost 2 to 4 % when we compare results in a pairwise manner.
2. Both methods show the same relative agreement rate per language pair. The pair with the highest agreement corresponds to English-Spanish, second comes the pair Spanish-Basque, and finally the pair English-Basque shows the lowest agreement.

In the rhetorical analysis, unlike those we have achieved in the harmonization (changes made in languages to carry out the alignment of discourse units), we see no significant difference (Translation Strategies in Table 20) between languages typologically more distant. It is worth noting, however, that for the closest languages, the English-Spanish pair, the agreement in relation is higher. Languages with more contact like the Spanish-Basque pair obtain better agreement than the English-Basque pair (Table 21).

We see clear advantages to the use of the qualitative evaluation method. First of all, with a qualitative evaluation, we measure inter-annotator agreement using only RST relations. Relations and nuclearity are phenomena of a different nature, and we believe they ought not to be included in the same factor. Secondly, the qualitative evaluation clearly distinguishes the most relevant sources of disagreement; because

---

[24] This phenomenon (marker change is the first reason to mismatch relations) is repeated when we compare translated texts (TL) among them (MC 20.29 %, CSC 4,35 % and US 7.25 %).

Table 20 Translation strategies and different language pairs

	Translation strategies						Different language forms		
	ENG > SPA (%)	ENG > BSQ (%)	SPA > ENG (%)	SPA > BSQ (%)	BSQ > ENG (%)	BSQ > SPA (%)	ENG-SPA (%)	ENG-BSQ (%)	SPA-BSQ
MC	1.45	–	4.35	7.25	10.14	11.59	14.49	4.35	1.45
CSC	1.45	1.45	2.90	4.35	4.35	1.45	2.90	1.45	–
US	2.90	2.90	2.90	1.45	4.35	2.90	0.00	4.35	2.90
Total	68.12						31.88		

Table 21 Comparison of relation factor in quantitative and qualitative evaluation methods (F-measure)

	Quantitative evaluation (%)	Qualitative evaluation (%)
ENG-SPA	56.22	59.67
ENG-BSQ	53.28	54.66
SPA-BSQ	54.94	56.99

of that, results are more reliable. The translation of discourse structure from one language to another does not result in a one-to-one mapping of relations. As Marcu (2000a) has mentioned, sometimes a particular rhetorical structure has to be translated as a different structure. Moreover, translation strategies can affect the rhetorical structure and annotation, and the qualitative method presented here could be used to identify and measure these translation strategies.

4 Conclusions and further work

The methodology we have proposed has two main implications for RST theory and for annotation methodology. First of all, in terms of RST theory, we have shown that it is possible to conduct cross-linguistic studies using the same set of principles. In our study we have shown that, although RST structures may not be exactly the same across languages, they do show a large similarity. Secondly, we have provided a clear and detailed method to identify where structures differ. Thirdly, the annotated files are available to anyone who wishes to use them and on our website[25] the tagged multilingual corpus can be consulted, as for example: (1) the rhetorical structure of a text (in Rs3 format) and its image (in Jpg format); (2) all instances of a selected rhetorical relation in three languages; (3) discourse units of a text in each language or aligned in three languages.

Ours is, to our knowledge, the first study that provides a rigorous qualitative methodology for comparison of rhetorical structures, which solves the deficiencies of quantitative evaluations and provides a qualitative description of agreement and disagreement. This method distinguishes and locates translation strategies when those strategies are the sources of annotator disagreement, as opposed to simple annotator discrepancies. The methodology helps determine whether the same passage in different languages has different RST structures because those structures correspond to different applications of the theory, or whether the discrepancy in RST structures is due to different linguistic realizations (due to translation strategies, broadly understood).

The study has some limitations with regard to the source of the translation differences that the analysis reveals. We believe that in order to detect these sources a translation theory "must include both a descriptive and an evaluative element", as Chesterman (1993) suggests, so that we can decide whether translation strategies may or may not be well motivated. We have presented some suggestions for the

---

[25] http://ixa2.si.ehu.es/rst.

translation differences that the analysis evidenced, showing that typological differences between the languages affected mostly segmentation. More detail, informed by a rigorous translation theory, is necessary, but is beyond the scope of this paper.

Our results show that RST, in conjunction with our methodological proposal for the comparison of RST annotations, are valid tools for the study of translated corpora. The results of our corpus analysis provide some evidence that, in segmentation, the linguistic distance calculated by change in the harmonization process is very small between languages from the same family such as English-Spanish and it is large between languages from distinct families such as Spanish-Basque and English-Basque. Surprisingly, the dispersion in relation agreement caused by translation strategies was very small when comparing English-Basque and Spanish-Basque with English-Spanish. In the same line, the linguistic distance in rhetorical relations, calculated as the F-score result when comparing RST annotations, is not as large as the segmentation differences. It appears that there is more dispersion in segmentation than in rhetorical relations; this may be due to the fact that there is more distance at the level of clause linking than at the level of discourse relational structure. It is worth noting, however, that each language is affected by a particular translation strategy in this corpus.

Although the results obtained by both methods in the annotations for different languages show that there are different interpretations, this is not due to interlingual differences. The problem of annotation subjectivity arises also when three annotators analyze the same text in a language: this problem is even more important when the annotators do not have the same training (although in our experiment the three annotators started their annotation from the same departure criteria). As we said, the purpose of this paper is to present a methodology to compare RS-trees and not to describe the structure of text in the three languages. To see a description of those texts and a detailed work in these three languages, we recommended consulting the corpora developed by the authors in these three languages (English SFU corpus[26] (Taboada and Renkema 2008), Spanish RST TreeBank[27] (da Cunha et al. 2011b) and Basque RST TreeBank[28] (Iruskieta et al. 2013a)). We are aware that in this work we do not account for the problem of multiple relations in RST (Taboada and Mann 2006b; Marcu 2000b) or all the possibilities comparing RS-trees in parallel corpora.

The qualitative evaluation is in certain respects more complex than Marcu's quantitative evaluation, which has been automated by Maziero and Pardo (2009). Despite its complexity, it solves some inherent problems of the quantitative evaluation and it has advantages when describing the sources of disagreement.

We plan to perform two tasks as future work. First of all, we will carry out a larger RST multilingual corpus analysis, but limited to a smaller number of rhetorical relations, with the objective of detecting translation strategies in order to improve machine translation discourse tasks. Second, we will carry out an automatic

---

[26] SFU corpus is available at http://www.sfu.ca/~mtaboada/download/downloadRST.html.

[27] RST Spanish TreeBank is available at http://corpus.iingen.unam.mx/rst/corpus_en.html.

[28] Basque RST TreeBank is available at http://ixa2.si.ehu.es/diskurtsoa/en/.

implementation of the qualitative rhetorical evaluation that we propose in our work, which will be valid for monolingual (Iruskieta et al. 2013a) and multilingual annotation, so that it can be used by all the scientific community working on RST.

Acknowledgments This work has been partially financed by the Spanish projects RICOTERM 4 (FFI2010-21365-C03-01) and APLE 2 (FFI2012-37260), and a Juan de la Cierva Grant (JCI-2011-09665) to Iria da Cunha. Maite Taboada was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (261104-2008). Mikel Iruskieta was supported by the following projects: OPENMT-2 (TIN2009-14675-C03-01) [Spanish Ministry], Ber2Tek (IE12-333) [Basque Government] and IXA group (GIU09/19) [University of the Basque Country]. We would like to thank the anonymous reviewers for their comments and suggestions, Nynke van der Vliet for her feedback on the evaluation method, Esther Miranda for designing the website, and Oier Lopez de Lacalle for helping with the scripts to calculate the statistics.

Appendix: Discourse segmentation details

The first step in analyzing texts under RST consists of segmenting the text into spans. Exactly what a span is, under RST, and more generally in discourse, is a well-debated topic. RST Mann and Thompson (1988) proposes that spans, the minimal units of discourse—later called elementary discourse units (EDUs) (Marcu 2000a)—are clauses, but that other definitions of units are possible:

> The first step in analyzing a text is dividing it into units. Unit size is arbitrary, but the division of the text into units should be based on some theory-neutral classification. That is, for interesting results, the units should have independent functional integrity. In our analyzes, units are essentially clauses, except that clausal subjects and complement and non-restrictive relative clauses are considered as part of their host clause units rather than as separate units.
>
> (Mann and Thompson 1988, p. 248)

This definition is the basis of our work. From our point of view, adjunct clauses stand in clear rhetorical relations (cause, condition, concession, etc.). Complement clauses, however, have a syntactic, but not discourse, relation to their host clause. Complement clauses include, as Mann and Thompson (1988) point out, subject and object clauses, and restrictive relative clauses, but also embedded report complements, which are, strictly speaking, also object clauses.

Other possibilities for segmentation exist; one of the better-known ones is the proposal by Carlson et al. (2003) for segmentation of the RST Discourse Treebank (Carlson et al. 2002). Carlson et al. (2003) propose a much more fine-grained segmentation, where report complements, relative clauses and appositive elements constitute their own EDUs.

In our work three annotators segmented the EDUs of each corpus (A1 segmented English texts, A2 segmented Spanish texts, and A3 segmented Basque texts). These annotators are experts on RST, since they have been researching in this field since years ago, and they have participated in several projects related to the design and elaboration of RST corpora in the three languages of this work. Annotators performed this segmentation task separately and without contact among them. In our segmentation, we follow then the general guidelines proposed by Mann and Thompson (1988), which we have operationalized for this paper. We detail the principles below.

Every EDU Should Have a Verb

In general, EDUs should contain a (finite) verb. The main exception to this rule is the case of titles, which are always EDUs, whether they contain a verb or not.

Non-finite verbs form their own EDUs only when introducing an adjunct clause (but not a modifier clause, as we will see below). In (7), the non-finite clause *Focussing on less widely...* is an independent EDU, because it is an adjunct clause. Note that in both Spanish and Basque the same proposition was translated as an independent sentence.

(7)

(a) [Focussing on less widely used and taught languages (LWUTLs) including Irish,] [the VOCALL partners are compiling multilingual glossaries of technical terms in the areas of computers, office skills and electronics] [and this involves the creation of a large number of new Irish terms in the above areas.]

(b) [El proyecto está enfocado hacia lenguas minoritarias en cuanto al uso y enseñanza, incluido el irlandés.] [El proyecto VOCALL estáen proceso de recopilación de un glosario plurilingüe de términos técnicos de las áreas de informática, secretariado y construcción,] [y esto supone la creación de una larga serie de nuevos términos en irlandés, en las áreas mencionadas.]

(c) [Gutxi erabiltzen eta irakasten diren hizkuntzetan kontzentratzen da proiektua (LWUTL), irlandera barne.] [Informatika, bulego-lana eta eraikuntzako arloetako termino teknikoen glosario eleanizduna biltzen ari da VOCALL,] [eta horrek esan nahi du arlo horietako irlanderazko termino berri ugari sortzen ari dela.] TERM23_ENG

In some cases, a prepositional phrase (especially one containing a nominalized verb) in one language was realized as an independent clause in another. The final decision in such cases is typically to segment minimally, that is, to unify the segmentation across the three languages, so that the language with the fewer segments determines how the texts in the other languages have to be segmented. See also Sect. 3.1.1, on harmonization of the segmentation, for more examples of our final decisions across the three languages.

Coordination and Ellipsis. Coordinated clauses are separated into two segments, including cases where the subject is elliptical in the second clause. In Spanish and Basque, both pro-drop languages, this is in fact the default for both first and second clause, and therefore we see no reason why a clause with a pro-drop subject cannot be an independent unit. We follow the same principle for English. In (8), the first two EDUs in Spanish are coordinated with an elliptical subject in both cases, referring to the authors (*venimos traduciendo,* '[we] have been translating' and *queremos expresar,* '[we] wish to indicate'). They constitute separate EDUs. In the English and Basque versions, the two clauses are expressed as separate sentences.

(8)

(a) [To attain this goal we have been translating doctrinal texts in law at the University of Deusto since 1994.] [We wish to indicate the difficulties we have had over the years and also our achievements,] [if there can be said to be any.]

(b) [Para poder alcanzar ese objetivo en la Universidad de Deusto venimos traduciendo textos doctrinales del campo del Derecho desde 1994] [y queremos expresar las dificultades que hemos tenido a lo largo de estos años y, asímismo, también los logros conseguidos,] [si es que realmente los ha habido.]

(c) [Xede hori iristeko, 1994. urteaz geroztik, Deustuko Unibertsitatean Zuzenbidearen inguruko testu doktrinalak itzultzen dihardugu.] [Esperientzia horretan izandako zailtasunak eta,] [halakorik izanez gero,][29] [lorpenak ere azaldu nahi ditugu.] TERM25_BSQ

Coordinated verb phrases (VPs) or verbs do not constitute their own EDUs. We differentiate coordinated clauses from coordinated VPs because the former can be independent clauses with the repetition of a subject; the latter, in the second part of the coordination, typically contain elliptical verbal forms, most frequently a finite verb or modal auxiliary.

Relative, Modifying and Appositive Clauses. We do not consider that relative clauses (restrictive or non-restrictive), clauses modifying a noun or adjective, or appositive clauses constitute their own EDUs. We include them as part of the same segment together with the element that they are modifying. This departs from RST practice, where (restrictive) relative clauses are often independent spans, as seen in many of the examples in the original literature and the analyzes on the RST web site (Mann and Thompson 1988; Mann and Taboada 2010). We found that relative clauses and other modifiers often lead to truncated EDUs, resulting in repeated use of the Same-unit relation (see Truncated EDUs in 5 section), and thus decided that it was best to not elevate them to the status of independent segments.

An example is presented in (9), where the relative clause is in parentheses in the Spanish original. Note, however, that the coordinated clauses (with an elliptical subject in all cases) are independent segments, as explained above. In Basque, on the other hand, the relative clause is translated as an independent clause with a finite verb (*mugatzen da,* ‘[it] is limited to’). We have not segmented it in Basque, to agree with the other two languages.

(9)

(a) [...] [Internet terminology extends beyond the bounds of its specialist field (which by definition is part of the lexicon of science and technology)] [and breaks into general language.]

(b) [...] [la terminología de Internet traspasa los límites del área de especialidad (a la que se circunscribe por definición el léxico científico y técnico)] [e irrumpe en la lengua de uso general,] [...]

(c) [...] [espezialitateko eremuaren mugak gainditzen dituela Interneteko terminologiak (espezialitatera mugatzen da, definizioz, lexiko zientifiko

---

[29] Truncated EDU. English translation: ‘if there can be said to be any’ (see Sect. 5).

eta teknikoa),] [eta erabilera orokorreko hizkeran sartzen dela indartsu;] [. . .] TERM38_SPA

Parentheticals. The same principle applies to parentheticals and other units typographically marked as separate from the main text (with parentheses or dashes). They do not form an individual span if they modify a noun or adjective as in Example 10, but they do if they are independent units, with a finite verb. Such is the case in (11), with a full sentence in the parenthetical unit (in English, composed of three finite clauses: *can... be represented*, *is* and *are*).

(10)

(a) The analysis of the data at hand—international terms most of which have not yet been standardized in Serbian—indicate that a hierarchy of criteria for evaluating the terms, (...). TERM18_ENG

(11)

(a) [The design and management of terminological databases pose theoretical and methodological problems] [(how can a term be represented?] [Is there a minimum representation?] [How are terms to be classified?),] (. . .)

(b) [Efectivamente, el diseño y la gestión de las bases de datos terminológicos plantean problemas diversos tanto de índole teórica y metodológica] [(¿cómo se representa un término?,] [¿existe una representación mínima?,] [¿cómo se clasifican los términos?)] (. . .)

(c) [Hala da, terminologiako datu-baseak diseinatzeak eta kudeatzeak hainbat arazo dakar bai teoria eta metodologiaren aldetik] [(nola adierazi terminoa?] [Ba al da gutxieneko adierazpenik?] [Nola sailkatu terminoak?),] (. . .) TERM29_SPA

Reported Speech. We believe that reported and quoted speech do not stand in rhetorical relations to the reporting units that introduce them, and thus should not constitute separate EDUs, also following clear arguments presented elsewhere (da Cunha and Iruskieta 2010; Stede 2008a). This is in contrast to the approach in the RST Discourse Treebank (Carlson et al. 2003), where reported speech (there named ATTRIBUTION) is a separated EDU. There are, in any case, no examples of reported speech in our corpus.

Truncated EDUs. In some cases, a unit contains a parenthetical or inserted unit, breaking it into two separate parts, which do not have any particular rhetorical relation between each other. In those cases, we make use of a non-relation label, Same-unit, proposed for the RST Discourse Treebank (Carlson et al. 2003).

We see one such example in (11) above. The element that corresponds to the third unit in English is, in fact, inserted in the middle of the second unit in Basque. In order to align or harmonize segmentation and to preserve the integrity of that unit, we use the Same-unit (non) relation, as shown in Fig. 8, which follows the Basque word order.

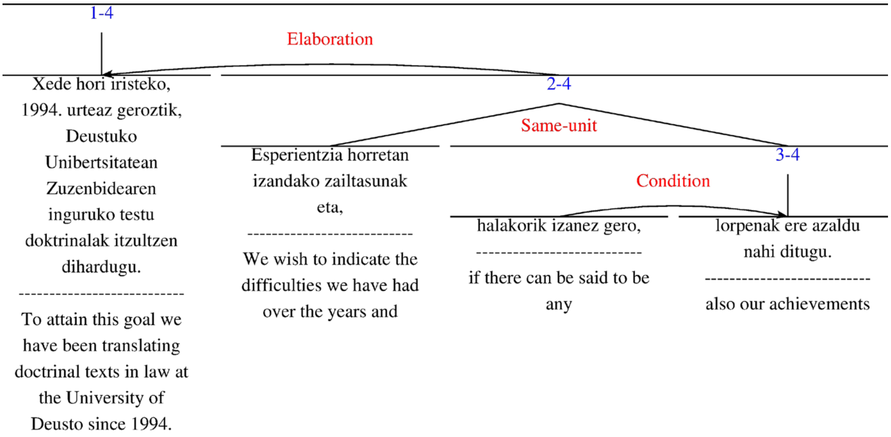

1-4
Elaboration
2-4
Same-unit
3-4
Condition
Xede hori iristeko, 1994. urteaz geroztik, Deustuko Unibertsitatean Zuzenbidearen inguruko testu doktrinalak itzultzen dihardugu.
To attain this goal we have been translating doctrinal texts in law at the University of Deusto since 1994.
Esperientzia horretan izandako zailtasunak eta,
We wish to indicate the difficulties we have had over the years and
halakorik izanez gero,
if there can be said to be any
lorpenak ere azaldu nahi ditugu.
also our achievements

Fig. 8 Example of a Same-unit (non) relation

Once our segmentation criteria were established and the three annotators carried out the segmentation, the three segmentations were compared in terms of precision and recall. In this way, we quantified agreement and disagreement across segmentations. Moreover, we analyzed the main causes of the disagreements. Results are shown in Sect. 3. After the segmentation agreement evaluation, we harmonized the segmentation, ensuring that units were comparable across the languages. At this point, we also calculated linguistic distance between the pairs of languages, We understand linguistic distance as "the extent to which languages differ from each other" (Chiswick and Miller 2005, pg. 1). Although this concept is well known among linguists, there is not a single measure to evaluate this distance Chiswick and Miller (2005). In our work, in order to measure this distance we calculated which language required the most changes in the harmonization process. This harmonization process was necessary to start out the analysis with similar units, and to avoid confusing analysis disagreement and segmentation agreement. Marcu et al. (2000) and Ghorbel et al. (2001) also align (which we termed harmonize) their texts, decreasing the granularity of their segmentation to avoid complexity. With this decision, we lose some rhetorical information at the most detailed level of the tree. This does not, however, affect higher levels of tree structure. The results of this harmonization are shown in Sect. 3.1.

References

Abelen, E., Redeker, G., & Thompson, S. A. (1993). The rhetorical structure of US-American and Dutch fund-raising letters. *Text*, *13*(3), 323–350.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

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

Bateman, J. A., & Rondhuis, K. J. (1997). Coherence relations: Towards a general specification. *Discourse Processes*, *24*(1), 3–49.

Carlson, L., Okurowski, M. E., & Marcu, D. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus].* Philadelphia, PA: Linguistic Data Consortium.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In van Kuppevelt, C. J. Jan & R. W. Smith (Eds.), *Current and new directions in discourse and dialogue* (pp. 85–112). Berlin: Springer.

Catford, J. C. (1965). *A linguistic theory of translation: An essay in applied linguistics* (Vol. 8). New York: Oxford University Press.

Cenoz, J. (2003). The role of typology in the organization of the multilingual lexicon. In J. Cenoz, B. Hufeisen & U. Jessner (Eds.), *The multilingual lexicon* (pp. 103–116), New York: Springer.

Chesterman, A. (1993). From 'is' to 'ought': Laws, norms and strategies in translation studies. *Target*, *5*(1), 1–20.

Chesterman, A. (1997). *Memes of translation: The spread of ideas in translation theory* (Vol. 22). Amsterdam and Philadelphia: Benjamins.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, *26*(1), 1–11.

Cristea, D., Ide, N., & Romary, L. (1998). Veins theory: A model of global discourse cohesion and coherence. In C. Boitet & P. Whitelock (Eds.), *17th international conference on Computational linguistics* (Vol. 1 pp. 281–285). Montreal, Canada: Association for Computational Linguistics.

Cui, S. (1986). A comparison of English and Chinese expository rhetorical structures. Ph.D. thesis, UCLA.

da Cunha, I., & Iruskieta, M. (2010). Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, *12*(5), 563–598.

da Cunha, I., Torres-Moreno, J. M., & Sierra, G. (2011a). On the Development of the RST Spanish Treebank. In *5th Linguistic annotation workshop. 49th annual meeting of the association for computational linguistics, ACL* (pp. 1–10). Portland, Oregon, USA.

da Cunha, I., Torres-Moreno, J. M., Sierra, G., Cabrera-Diego, L. A., Castro-Rolón, B. G., & Rolland-Bartilotti, J. M. (2011b). The RST Spanish Treebank On-line Interface. In *International conference recent advances in NLP* (pp. 698–703), Bulgaria.

Delin, J., Hartley, A. F., Paris, C., Scott, D. R., & Linden, K. V. (1994). Expressing procedural relationships in multilingual instructions. In *Seventh International Workshop on Natural Language Generation* (pp. 61–70), Association for Computational Linguistics.

Delin, J., Hartley, A. F., & Scott, D. R. (1996). Towards a contrastive pragmatics: Syntactic choice in English and French instructions. *Language Sciences*, *18*(3–4), 897–931.

Egg, M., & Redeker, G. (2010). How complex is discourse structure? In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)* (pp. 1619–1623), Valletta, Malta.

Fetzer, A., & Johansson, M. (2010). Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse. *International Journal of Corpus Linguistics*, *15*(2), 240–266.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Flowerdew, J. (2010). Use of signalling nouns across l1 and l2 writer corpora. *International Journal of Corpus Linguistics*, *15*(1), 36–55.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English–Chinese corpus. In *3rd workshop on very large Corpora*, (Vol. 78, pp. 173–183). Boston, MA.

Ghorbel, H., Ballim, A., & Coray, G. (2001). ROSETTA: Rhetorical and semantic environment for text alignment. In: *Corpus Linguistics*, Lancaster University (UK) (pp. 224–233).

Gomez, X., & Simoes, A. (2009). Parallel corpus-based bilingual terminology extraction. In *8th international conference on terminology and artificial intelligence* Toulouse.

Granger, S. (2003). *The corpus approach: A common way forward for Contrastive Linguistics and Translation Studies* (pp. 17–29). Rodopi, Corpus-based approaches to contrastive linguistics and translation studies. Amsterdam/New York.

House, J. (2004). *Explicitness in discourse across languages. Neue Perspektiven in der Übersetzungs-und Dolmetschwissenschaft* (pp. 185–208), Bochum: AKS.

Iruskieta, M., Aranzabe, M. J., Díaz de Ilarraza, A., Gonzalez, I., Lersundi, M., & Lopez de la Calle, O. (2013a). The RST Basque TreeBank: An online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, Brasil.

Iruskieta, M., Díaz de Ilarraza, A., & Lersundi, M. (2013b). Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 1–32.

Kanté, I. (2010). Mood and modality in finite noun complement clauses: A French-English contrastive study. *International Journal of Corpus Linguistics*, *15*(2), 267–290.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: *MT summit*, Phuket, Thailand.

Kong, K. C. C. (1998). Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text & Talk*, *18*(1), 103–141.

Mann, W. C., & Taboada, M. (2010). RST web-site. http://www.sfu.ca/rst/. Accessed 30 September 2012.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281.

Marcu, D. (2000a). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, *26*(3), 395–448.

Marcu, D. (2000b). *The theory and practice of discourse parsing and summarization*. Cambridge: MIT press.

Marcu, D., Carlson, L., & Watanabe, M. (2000). The automatic translation of discourse structures. In *1st North American chapter of the Association for Computational Linguistics conference* (pp. 9–17), Seattle (USA): Morgan Kaufmann Publishers.

Maxwell, M. (2010). Limitations of corpora. *International Journal of Corpus Linguistics*, *15*(3), 379–383.

Maziero, E. G., & Pardo, T. A. S. (2009). Automatização de um método de avaliação de estruturas retóricas. In: *RST Brazilian meeting*, São Paulo, Brazil.

Mitocariu, E., Anechitei, D. A., & Cristea, D. (2013). *Comparing discourse tree structures* (pp. 513–522). Berlin: Springer. Computational Linguistics and Intelligent Text Processing.

Mohamed, A. H., & Omer, M. R. (1999). Syntax as a marker of rhetorical organization in written texts: Arabic and English. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *37*(4), 291–305.

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2007). Bilingual terminology mining-using brain, not brawn comparable corpora. In *Annual meetings ACL* (Vol. 45, pp. 664–671). Prague.

Mortier, L., & Degand, L. (2009). Adversative discourse markers in contrast: The need for a combined corpus approach. *International Journal of Corpus Linguistics*, *14*(3), 338–366.

O'Donnell, M. (2000). RSTTool 2.4: A markup tool for rhetorical structure Theory. In *First international conference on natural language generation INLG'00* (Vol. 14, pp. 253–256). Mitzpe Ramon: ACL.

Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. Ph.D. thesis, Instituto de Ciências Matemáticas e de Computação, São Carlos-SP: Universidade de São Paulo.

Ramsay, G. (2000). Linearity in rhetorical organisation: A comparative cross-cultural analysis of newstext from the People's Republic of China and Australia. *International Journal of Applied Linguistics*, *10*(2), 241–258.

Ramsay, G. (2001). Rhetorical styles and newstexts: A contrastive analysis of rhetorical relations in Chinese and Australian news-journal text. *ASAA E-Journal of Asian Linguistics and Language-teaching*, *1*(1), 1–22.

Salkie, R., & Oates, S. L. (1999). Contrast and concession in French and English. *Languages in Contrast*, *2*(1), 27–56.

Sarjala, M. (1994). Signalling of reason and cause relations in academic discourse. *Anglicana Turkuensia*, *13*, 89–98.

Scott, D. R., Delin, J., & Hartley, A. F. (1998). Identifying congruent pragmatic relations in procedural texts. *Languages in Contrast*, *1*(1), 45–82.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *2003 conference of the North American Chapter of the Association for Computational Linguistics on human language technology* (Vol. 1, pp. 149–156). Association for Computational Linguistics.

Stede, M. (2008a). Disambiguating rhetorical structure. *Research on Language and Computation*, *6*(3), 311–332.

Stede, M. (2008b). *RST revisited: Disentangling nuclearity* (pp. 33–57). Amsterdam and Philadelphia: John Benjamins. 'Subordination' versus 'coordination' in sentence and text.

Taboada, M. (2004a). *Building coherence and cohesion: Task-oriented dialogue in English and Spanish*. Amsterdam and Philadelphia: John Benjamins.

Taboada, M. (2004b). *Rhetorical relations in dialogue: A contrastive study* (pp. 75–97), Amsterdam and Philadelphia: John Benjamins. Discourse across Languages and Cultures.

Taboada, M., & Mann, W. C. (2006a). Applications of rhetorical structure theory. *Discourse Studies*, *8*(4), 567–588.

Taboada, M., & Mann, W. C. (2006b). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, *8*(3), 423–459.

Taboada, M., & Renkema, J. (2008). *Discourse relations reference corpus*. Simon Fraser University and Tilburg University. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html. Accessed 30 September 2012

Trask, R. L. (1997). *The history of Basque*. London: Routledge.

Usoniene, A., & Soliene, A. (2010). Choice of strategies in realizations of epistemic possibility in English and Lithuanian: A corpus-based study. *International Journal of Corpus Linguistics*, *15*(2), 291–316.

UZEI and HAEE-IVAP. (1997). *International congress on terminology*. Donostia and Gasteiz: UZEI; HAEE-IVAP.

van der Vliet, N. (2010). Inter annotator agreement in discourse analysis. http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/.

Wu, D., & Xia, X. (1994). Learning an English–Chinese lexicon from a parallel corpus. In *First conference of the AMTA* (pp. 206–213). Citeseer, Columbia.

Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35.

DE GRUYTER MOUTON Linguistics 2015; 53(1): 53–92

Ted J. M. Sanders* and Wilbert P. M. Spooren

Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text

Abstract: Many languages of the world have connectives to express causal relations at the discourse level. Often, language users systematically prefer one lexical item (*because*) over another (even highly similar) one (*since*) to express a causal relationship. Such choices provide a window on speakers' cognitive categorizations, and have been modeled in previous work in terms of subjectivity. However, a broader empirical basis and a more specific operationalization of subjectivity are urgently needed. This paper provides in these needs by developing an integrative empirical approach to the analysis of the Dutch connectives *omdat* 'because' and *want* 'since/for' in written text, conversation, and chat interactions. These can be considered a case in point for linguistic categorization since related European languages show similar distinctions. The construct of subjectivity is decomposed into characteristics like type of relation and s*ubject of consciousness* (who can be considered responsible for the causality?). The use of statistical methods specifically suitable for hypothesis testing in natural language corpora produces results that provide new insights into the division of labor between the two connectives, as well as into the notion of subjectivity.

Keywords: causality, connectives, subjectivity, Dutch, discourse

DOI 10.1515/ling-2014-0034

***Corresponding author: Ted J. M. Sanders:** Utrecht Institute of Linguistics, Universiteit Utrecht, Trans 10, NL-3512 JK Utrecht, The Netherlands. E-mail: t.j.m.sanders@uu.nl
Wilbert P. M. Spooren: Discourse studies, Radboud Universiteit Nijmegen, Erasmusplein 1, NL-6525 HT Nijmegen, The Netherlands. E-mail: w.spooren@let.ru.nl

1 Introduction

1.1 Discourse, causality, and connectives

People use language to communicate in various contexts and in various media, be it in spontaneous conversations, in writing and reading texts, or in chat interactions. Discourse is a crucial level in all types of human linguistic communication; it is impossible to communicate without understanding the coherence between utterances. One type of coherence relation language users often want to express is causality, for instance in the case of a *reason* or a *consequence-cause* relation (see (1)) between events in the world or by connecting a claim and an argument (see (2)). In English, both relations can be made explicit with the connective *because*.

(1) *The fields are wet because it has rained a lot this week.*

(2) *Surely all soccer games will be cancelled, because it has rained a lot this week.*

In this paper, we focus on this backward causality – that is, the order "S1, CONNECTIVE S2", where S stands for discourse segment, which is minimally a clause.

Many languages of the world have connectives to express causal relations at the discourse level (see Diessel and Hetterle 2011, who analyzed causal clauses in 60 languages from typologically different language families). Speakers of English, for example, can choose between *because* and *since* or *for*. We are interested in the system behind the meaning and use of such connectives. Also, we ask how different these choices in English are from the ones made by speakers of other languages, such as Dutch *omdat* versus *want*, German *weil* versus *denn*, and French *parce que* versus *car*. It seems as if language users often systematically prefer one lexical item over another (even highly similar) one to express a certain type of causal relationship. Systematic use of a particular lexical item to express a certain type of causal relationship implies that people distinguish between several types of causality. Hence, such choices could provide a window on speakers' cognitive categorizations of causality. The linguistic study of the meaning and use of causal connectives may reveal insights into human categorization of causality (see, among others, Sanders and Sweetser, 2009). In other domains, like the study of metaphor or that of causative verbs, similar studies of linguistic categories apparent in people's everyday language use have already produced many interesting insights into the working of the mind (see, for instance, Lakoff 1987; Verhagen and Kemmer 1997).

In her seminal work, Sweetser (1990) presents a domain approach in which she argues that a conjunction like English *because* is used in the content-domain when one event causes another in the real world (3). Epistemic use (4) concerns the speaker's reasoning and (5) illustrates the speech act use.

(3) *John came back because he loved her*. (i.e., the loving caused the return)

(4) *John loved her, because he came back*. (i.e., the observation that he came back is an argument for the claim that the loved her)

(5) *What are you doing tonight, because there's a good movie on*. (i.e., I invite you to come and I give a reason for performing this speech act).

In more recent years, we have seen related proposals in which distinctions like content, epistemic and speech act domains are described in terms of the subjectivity of *speaker involvement* (Pander Maat and Degand 2001; Pander Maat and Sanders 2000, 2001). In such an approach, content relations such as Cause-Consequence are objective (because the Speaker is not involved), whereas epistemic and speech act relations are subjective (because the Speaker is clearly involved) (see Section 1.2 for a further elaboration).

The distinction between, on the one hand, coherence between events in the world – named objective, semantic, propositional, internal, or content relations – and on the other hand coherence realized by the communicative acts or reasoning of the speaker – subjective, pragmatic, external relations – can be found in virtually all taxonomies and categorizations of coherence relations (Kehler 2002; Knott and Dale 1994; Mann and Thompson 1988; Martin 1992; Sanders et al. 1992; Sanders 1997). In addition, crosslinguistic studies suggest similar distinctions are useful to describe the organization of the lexicon of causal connectives in languages like Dutch, German, and French (Pit 2006; Evers-Vermeul et al. 2011). These languages show a more differentiated repertoire of connectives to express backward causality than English, where *because* can be used across the three domains (Ford 1993; Knott and Sanders 1998; Sweetser 1990). Dutch *want* can be used to express speech act and epistemic relations and is therefore considered more subjective than *omdat* (Degand 2001; Pit 2006; Verhagen 2005). Similarly, French *car* and *puisque* would be specifically used in the epistemic domain, whereas *parce que* sticks to content relations (*Groupe λ* 1975; Anscombre and Ducrot 1983; Degand and Pander Maat 2003; Zufferey 2010). Similarly, several German linguists have suggested that *denn* can only be used to express epistemic relations (Pasch 1983; Günthner 1993; Keller 1995, but see also Wegener 2000).

Since the mid 1990s we have witnessed a rise in corpus studies to investigate these and related ideas about the organization of the lexicon of connectives in several languages, seeking to find the system behind the meaning and use of (causal) connectives, which has lead to an empirical test in actual language use of these challenging theories and hypotheses (see Sanders and Spooren 2009 and other contributions to Sanders and Sweetser 2009).

1.2 Subjectivity as categorization principle: inherent characteristic or context-dependent?

The corpus studies mentioned above have marked an important step forward in the field. However, there are fundamental challenges left, both of a theoretical and an empirical nature. The first issue concerns the notion of subjectivity, a term that is often used for different phenomena, which can be a source of confusion (Nuyts 2012). The second concerns the question whether the semantic profiles of causal connectives can actually be characterized in terms of subjectivity as a stable characteristic of their semantic profile, or whether this subjectivity is a context-dependent characteristic. A third issue concerns the empirical basis of the corpus studies, which is insufficient. We aim to address these three issues in this paper, in which we propose an integrative empirical approach to subjectivity in discourse. Below, we first address three and two, to end up with the first issue: the operationalization of subjectivity.

1.2.1 The empirical basis of current corpus studies

The empirical domain of the corpus studies that have been conducted until now is limited in many respects. First of all, there is a substantial amount of work on German (e.g., Günthner 1993; Keller 1995) and French (Anscombre and Ducrot 1983; Groupe λ 1975) causal connectives, some of which specifically investigates spontaneous conversation, but these studies have a limited empirical basis: they analyze only small amounts of cases, and statistical evaluation is often lacking. For Dutch, some recent studies providing such evaluations are available but these studies are dominated by analyses of written text (e.g., Pander Maat and Degand 2001; Pit 2006; Stukker and Sanders 2012); in fact, the only study on non-written media is a small pilot-study in which we compared spontaneous conversation and chat with written text (Spooren et al. 2010).

There is a certain urgency to add other than written data to the empirical foundation of theories on the categorization of connectives. Several studies of spontaneous conversations suggest a typical usage pattern of causal connectives in conversations. Günthner (1993) and Keller (1995) demonstrated that German *weil* can express epistemic relations in spontaneous conversations, whereas in written language it seems to be reserved for the content domain. Recently, the work on German causal connectives has grown considerably, both in terms of quantity and in terms of empirical base; recent extensive and statistically evaluated corpus studies include Frohning (2007), Breindl and Walter (2009) and Volodina (2011a). In some of these studies insights from spoken language data have shown the profiles of causal connectives to be less specialized than was concluded on the basis of the analysis of only written texts (Breindl and Walter 2009). For French, Zufferey (2010, 2012) concludes that *puisque* has a strong preference for epistemic use in telephone conversations. Such results show that written language as the basis for analysis may lead to a distorted picture. This observation is the basis for addressing our second issue: how stable or context-dependent are characterizations of connectives in terms of subjectivity.

A principled point is that written language deviates from the prototypical communicative situation that spontaneous conversations provide in several respects (Clark 1996): Written language generally shows a large distance between authors and readers. In written language the communicators are relatively invisible. As a consequence, written language is detached from the deictic center of communication (Sanders et al. 2009). Authors can (re)consider and revise their lexical choices and formulations; the focus is usually on content and not on interpersonal issues. And whereas spoken language is fragmented, written language is integrated (Chafe 1994). These considerations lead to the conclusion that the use of causal connectives should be investigated systematically in different media. Such investigations are scarce (but see Zufferey 2010, 2012).

In this paper we present a systematic study of the use of *want* and *omdat* as a case in point of how European languages encode backward causal relations that differ in subjectivity. Our research question is: do *want* and *omdat* have a clear semantic profile that is constant across media, or do they have a vague profile and is their use mainly determined by the context in which they appear? The answer to this question is not obvious, given the limited empirical basis for Dutch. And it is not improbable that context plays a great role, given the frequency data presented in Table 1. This table shows strong differences in frequencies between *want* and *omdat* across the media in which they occur.

Why is that? Do *want* and *omdat* indeed have clear semantic profiles and do language users express different causal relations in different media? Or are the semantic profiles in fact not so clear, and do people use *want* and *omdat* in a less systematical way to simply express all kinds of causals? Is *omdat* relatively frequent in newspapers because it is a subordinating conjunction and thus supports the integrative nature of written language? Conversely, is *want* relatively frequent in chat not so much because speakers in chat express different types of causal

Table 1: Relative frequency of *omdat* and *want* (per million words).

Medium	Connective	
	Omdat	**Want**
Newspaper[a]	920	660
Spoken Discourse[b]	521	1640
Chat interaction[c]	445	1032

[a] based on pilot version of D-Coi; [b] based on Corpus of Spoken Dutch; [c] based on VU Chat corpus.

relations than in newspapers, but rather because the coordinating conjunction *want* fits in with the fragmented nature of chat language? In this paper we investigate whether there is a systematical relationship between the semantic profile of the connectives and the medium in which they occur.

1.2.2 The analysis of subjectivity in natural discourse: towards an integrative approach

The final fundamental issue that needs to be addressed is the operationalization of subjectivity. Broadly speaking, three fundamental approaches to subjectivity can be distinguished (Nuyts 2012): those by Lyons (1977), by Traugott (1995) and by Langacker (1990).[1] The three approaches highlight different aspects of the complex notion of subjectivity. Our aim is not to present a totally new approach but rather to combine crucial aspects of all three approaches, in order to operationalize subjectivity as a discourse phenomenon. In fact, we make use of the possibility that the different notions of subjectivity are to some extent "co-applicable" (Nuyts 2012: 22). Subsequently, we will test this integrative approach to subjectivity in an empirical way.

Reference to the speaker is generally considered a core component of linguistic subjectivity: "[Subjectivity is] the property (or set of properties) of being either a subject of consciousness (i.e., of cognition, feelings, and perception) or a subject of action (an agent). It denotes the property of being what Descartes called 'a thinking entity' " (Lyons 1995: 337). Traugott (1995: 31) defines subjectification as the process through which "meanings become increasingly based in the speak-

---

[1] We are aware that these notions were intended to refer to different phenomena and show differences (Nuyts 2012). At the same time, we think that the basic insights represented in these approaches can be combined to develop a valid account of subjectivity in discourse connectives.

er's subjective belief state/attitude toward the proposition".[2] Consequently, an utterance is subjective if it requires reference to the speaker in its interpretation, and objective if it does not. In this paper we follow this characterization of subjectivity, as we will argue below. Additionally, we use Langacker's (1990) insight that subjectivity is defined by the way in which an entity is construed. It is construed with maximal subjectivity when it remains implicit and "off stage", and with maximal objectivity when it is put onstage as an explicit focus of attention. Crucial to this aspect of subjectivity is the way in which the speaker figures in the form of the utterance: explicit reference is objective, whereas an implicit reference is subjective. This implies that Langacker's categories of "subjective" and "objective" are both speaker-related, and therefore "subjective" in Traugott's terms (De Smet and Verstraete 2006: 369; see also Vis 2011; Nuyts 2012).

We consider both speaker-relatedness (Lyons; Traugott) and implicit presence of the speaker (Langacker) as important aspects of subjectivity. Furthermore, we draw attention to the distinction between speaker-subjectivity and character-subjectivity, as discussed for example by Sanders et al. (2012). Finally, we also include the nature of the causal relation itself as a characterization of subjectivity, since our topic of research is causal coherence relations in discourse.

In line with earlier work on causal connectives (Pander Maat and Sanders 2000), we define an utterance as subjective when its interpretation requires an active *Subject of Consciousness* (from now on SoC). A SoC crucially involves an animate subject, a person, whose intentionality is conceptualized as the ultimate source of reasoning, evaluating, describing, or acting "in the real-world". An utterance is subjective because there is some thinking entity in the discourse who evaluates. For instance, *He thought Paris was nice* is subjective because it involves an evaluation by a character in the discourse. Compare this with an utterance like *Paris is in France*, which is presented as a fact in the world that does not depend on the evaluation by a SoC. To be more precise, in the utterance *He thought Paris was nice* the validity of the proposition "Paris is nice" depends on the SoC *He*, whereas in the utterance *Paris is in France* the proposition "Paris is in France" can be verified directly in the non-linguistic reality.

Obviously, each utterance in a discourse comes from a speaker or author, and therefore each utterance is dependent on a SoC. However, in some utterances, the

---

[2] This definition differs from the definition that Nuyts (2012) gives of subjectivity, a difference that is also noted by Nuyts, who considers an utterance like *They may have well left already* not subjective, but neutral, as it does not contain an explicit reference to the first person "assessor". For Traugott, such an utterance is indeed subjective.

SoC is manifest because the sequence cannot be interpreted without reference to a SoC. Such cases – typically feelings and evaluations of all kinds – are considered subjective; they simply cannot be interpreted without making reference to the SoC, her thoughts and feelings. In contrast, utterances that do not depend on such a manifest reference to the SoC are considered objective.

Authors/speakers can be SoC's, but characters can also function as such. The author/speaker is the first voice in the discourse who has constant access to her feelings and thoughts. She does not have access to the feelings and thoughts of a third person. As a result, *I think Paris nice* can be a direct report of an inner feeling, whereas *he thinks Paris nice* is a description of an evaluation. It follows that first person evaluations are more subjective than third person evaluations. The difference, then, between the speaker/writer versus a character as SoC is that the first type concerns a first voice, which is grounded in the Deictic Center of Communication (Sanders et al. 2009). This resembles Traugott's (1989, 1995) view on subjectivity as closeness to the communicative "here and now": the speaker here and now asserts that a particular state of affairs holds. By contrast, the character type concerns a third person in the discourse, which is more distant from the Deictic Center of Communication.

The examples mentioned so far, be it first person SoC's or third person SoC's, are descriptions of evaluations and consequently they are more or less objective. In terms of Sweetser this type of subjectivity may still be in the content domain. Yet, evaluations are often much more implicit. Especially when the speaker/author is first person SoC and the evaluation concerns the here and now, spontaneous evaluations typically are of the type *Paris is great*, i.e., a first person SoC. Such utterances express an evaluation and the SoC remains implicit. Indeed these are the most subjective type of utterances: those in which the speaker is SoC in first person, but remains off stage (Langacker 1990).

In sum, central to our integrative approach is that we consider utterances subjective when they cannot be interpreted without reference to a SoC; the SoC's thoughts, feelings, point-of-view are simply necessary for interpretation. This approach acknowledges that subjectivity ultimately is a cognitive notion. Although it can be signaled through linguistic expressions – such as explicit reference to the speaker / SOC, modal expressions, evaluative verbs, scalar predicates –, the Subjectivity of an utterance does not *depend on* the occurrence of such signals. It is typical for our take on Subjectivity to consider the author/speaker as the default SoC, who will often remain implicit and off stage, often producing subjective utterances.

In our analysis so far, we have presented subjectivity as a property of utterances. However, this is not enough, as subjectivity can also reside in the nature of relations between utterances, as the examples in (3–5), repeated below for convenience, have demonstrated: the link can be of a content, epistemic or speech act type. In fact, it is this kind of subjectivity that is the main focus of our interest (Sanders et al. 1992). Causal links in the content domain (example 3) are objective. Epistemic causality is inherently subjective because the speaker is actively reasoning towards a conclusion or concluding something on the basis of an observation (example 4) in the here-and-now. Cases of speech act causality (example 5) are also subjective: the speaker is performing a speech act and motivating that act on the basis of an observation.

(3) *John came back because he loved her.*

(4) *John loved her, because he came back.*

(5) *What are you doing tonight, because there's a good movie on.*

In our model of analysis we take the relational nature of subjectivity into account in two ways: we distinguish between different types of causal relation, and we specify the SoC that is responsible for the causal link. Example (3) is a content relation, with a third person SoC, *John*. Examples (4) and (5) are of an epistemic and speech act nature, respectively, with the speaker as the SoC. Note that this does not mean that there is a perfect correlation between relation type and SoC. Consider example (6), based on the epistemic relation in example (2).

(6) *That Saturday morning, Willem was sad. [S1 Now all soccer games would be cancelled], because [S2 it had rained a lot that week].*

Example (6) is an epistemic relation: the SoC concludes on the basis of an argument that the games will be cancelled. What is special about this example is that the SoC is not the speaker (as in examples (2) and (4)), but a third person (Willem). This, then, is a case of free indirect speech ("an unspeakable sentence," Banfield 1982; Sanders 2010) and shows that epistemic relations can occur in a third person SoC context. Authors like Banfield (1982) and Schlenker (2004) have argued that complex cases like Free Indirect Speech and the Historical Present constitute a challenge for the linguistic analysis of subjectivity. Our description of example (6) shows the potential of our approach for such cases.

1.2.3 An empirical approach

To summarize the discussion so far, we decompose the complex construct of subjectivity in terms of four characteristics of causal connections. In answer to our research question about the semantic profile of Dutch causal connections, we investigate whether and to what extent these characteristics co-occur. To that end, we make use of statistical methods specifically suitable for hypothesis testing in natural language corpora. Such methods allow us to test whether these characteristics help to distinguish between two Dutch backward causal connectives, *omdat* and *want,* and to what extent the differences between these two connectives are determined by the context in which they occur. Thus we follow Gries (2012), who strongly advocates using corpus data in developing "a psycholinguistically informed, (cognitively-inspired) usage-based linguistics which should be located, firmly and deliberately, in the social/behavioral sciences".

The integrative empirical approach that we develop provides us with new insights into the notion of subjectivity. On top of that, we advance the research into linguistic categorization by a detailed and rigorous empirical study of a relatively large corpus of naturally occurring language from various media. The detailed study of Dutch *want* and *omdat* can be considered a case in point for linguistic categorization since related European languages show similar distinctions.

1.3 Backward causality in Dutch

Causality can be expressed using backward and forward causal connectives. In a forward causal construction the first segment introduces a cause or an argument, and the second segment expresses a consequence or a claim. In backward constructions, the first segment expresses a claim or a consequence, and the second segment expresses the argument or the cause. In backward constructions, the connective typically occurs at the beginning of the second segment. In Dutch the connective signaling a backward causal relation can be a coordinating conjunction (like *want*) or a subordinating conjunction (like *omdat, aangezien,* or *doordat*). The most frequently used causal connectives are *want* and *omdat*.[3] Usually,

---

[3] This paper focuses on backward causality only. The reason is that forward causal connectives are treated in other work (cf. Stukker and Sanders 2012; Stukker, Sanders and Verhagen 2008). The reason that we confine ourselves to *want* and *omdat* is that these are by far the most frequent causal connectives in the Dutch language. For example, the connective *aangezien* occurs in our newspaper corpus only 59 times per million words, and most of these occurrences are forward causal connectives. The connective *doordat* is mostly used as a backward causal connective, but it occurs only 98 times per million words in our newspaper corpus.

the subordinator *omdat* is translated in English as *because*, whereas the coordinator *want* is often translated as *since* or *for*. Both translations of *want* seem too "formal" for *want*, which is very frequently used in relatively informal contexts, especially in spoken discourse, see table 1. The prototypical use of the Dutch backward causal connectives *want* and *omdat* can be illustrated by translating the English examples used so far.

(1) D *De velden zijn nat omdat het veel geregend heeft deze week.*
'The fields are wet OMDAT it has rained a lot this week'

(2) D *De voetbalwedstrijden worden vast afgelast, want het heeft deze week erg veel geregend.*
'Surely all soccer games will be cancelled, WANT it has rained a lot this week'

(3) D *Jan kwam terug omdat hij van haar hield.*
'Jan came back OMDAT he loved her'

(4) D *Jan hield van haar, want hij kwam terug.*
'Jan loved her, WANT he came back'

(5) D *Wat doe jij vanavond, want er draait een goede film.*
'What are you doing tonight, WANT there's a good movie on'

These examples illustrate how *want* 'for/since' is typically used to express epistemic and speech act relations, whereas *omdat* 'because' is typically used to express content relations. More specifically, *omdat* has a preference for *volitional content* relations, in which the intentions of a human actor propagate the actions. Several studies have shown that these characteristics are robust, and vary from strong preferences to clear restrictions on the relations they can express. Taken together, these observations show how the Dutch language "cuts up" backward causality (Degand 2001; Degand and Pander Maat 2003; Pit 2006).

The clearest case of this "cutting up" concerns a specific connective for non-volitional content relations: *doordat* ("as a consequence of the fact that"), see (7).

(7) *De temperatuur steeg, doordat de zon scheen.*
'The temperature rose, DOORDAT the sun was shining'

There are clear restrictions on its use: it only expresses non-volitional content relations. In fact, Dutch *doordat* can never be used to express the relations (3)–(5).

Other divisions of labor are tendencies rather than clear-cut restrictions. For example, the relation in (3) can also be expressed by *want*, which gives the sequence a more epistemic flavor. And in an example like (8) *omdat* is used in an epistemic context (although this use requires a pause before *omdat*; Huiskes 2010; Persoon et al. 2010).

(8) *Het moet wel een slechtvalk zijn, omdat hij met een enorme snelheid omlaag dook.*
'It must be a peregrine falcon, OMDAT it dove downwards with an enormous speed.'

Corpus studies also indicate that *want* regularly expresses volitional content relations, whereas *omdat* can express epistemic relations in a minority of cases (Bekker 2006; Degand 2001; Pit 2006). Hence, volitional content and epistemic relations are regularly lexicalized by the same connectives, an observation which can be taken as an argument against a strict domain-specific hypothesis, in which each connective correlates with a specific domain (content, epistemic, speech act). The study reported in this paper will shed new light on this discussion.

Before we move on to summarize the main research questions, it is important to elaborate somewhat on the syntactic differences between *want* (a coordinating conjunction) and *omdat* (a subordinating conjunction). It is known from the literature that there is a correlation between grammatical status and discourse function. For example, coordination constructions are syntactically independent and hence are better suited to express complete speech acts. Subordinating constructions are integrated, and therefore a less plausible site for expressing complete illocutions (van Dijk 1979; Verhagen 2005; Volodina 2011b). This raises the issue to what extent the grammatical difference between *want* and *omdat* is confounded with a difference in subjectivity. Admittedly, there is a strong correlation between syntactic integration and the tendency to express objective relations. Nevertheless, there are good reasons to assume that a difference in subjectivity between connectives cannot be reduced to a difference in their grammatical status. One reason is that despite the syntactic differences, it is not impossible for *want* to express objective relations (as will be shown in the results section), and similarly we find non-content subjective uses of *omdat* (cf. the use of *omdat* in spontaneous conversations to express subjective epistemic relations; Huiskes 2010; Persoon et al. 2010). A final reason is that in Dutch the subjectivity distinction is also relevant for categorizing connectives that do not differ grammatically, notably in the forward causal domain, where we have adverbial *dus* (*so*, *therefore*) preferring subjective epistemic and speech act relations, and the adverbial *daarom* preferring objective volitional content relations (Pander Maat and Sanders 2001; Pander Maat and Degand 2001; Stukker et al. 2008).

1.4 Research questions and hypotheses

If subjectivity is the right notion to analyze the difference between these connectives, it should go across the modalities of written, spoken and chat language. Therefore our main hypothesis is that *want* occurs in more subjective contexts than *omdat*, irrespective of the medium.

Generalization over media
Hypothesis 1: Across all media, *want* is used more often to express subjective relations (epistemic, speech act) than *omdat*.
Hypothesis 2: Across all media, *want* is used more often to support a judgment than *omdat*.
Hypothesis 3: Across all media, *want* is used more often with first and second person SoC's than *omdat*.
Hypothesis 4: Across all media, *want* is used more often with an implicit SoC than *omdat*.

In addition to these specific hypotheses, we formulate two explorative research questions. The first relates to differences between media. It is an open question whether the medium affects the degree of subjectivity, although one might argue that in relatively spontaneous media (spoken conversation, chat interaction) the Deictic Center of Communication is more salient than in relatively detached media (written text): a communicative situation of direct speaker-hearer interaction (spoken, chat) can be expected to be more subjectively grounded than written communication, because of the direct availability of Speaker and Addressee. The second question relates to the strength of the various subjectivity characteristics: are all characteristics of subjectivity (as formulated in the hypotheses) equally important in predicting the choice speakers or writers make between *want* and *omdat*?

In the following we further develop our integrative empirical approach to test these hypotheses and answer these research questions. By doing so, we tackle the two issues introduced in Section 1.2 – the need for broadening the empirical domain and the operationalization of subjectivity.

2 Method

2.1 Model of analysis

Below we present the specific discourse characteristics that were analyzed. Illustrative examples are taken from our corpora.

I. Propositional attitude of the first segment (S1)

As we argued above, evaluations are central for linguistic subjectivity. A prominent way in which evaluations manifest themselves in discourse is in the form of judgments. Consequently, each segment was analyzed as expressing either a judgment or another propositional attitude (fact, general knowledge, an intentional act, individual knowledge, a perception, an experience). Our analysis focuses on the first segment, because that is the site where relational subjectivity is most manifest in backward causals: for example, in a Claim-Argument relation the argument can be very factual and objective.

A segment expresses a judgment if it presents or implies a SoC – the person responsible for the causal relation; the Subject of Consciousness – and expresses what is judged. The segment expresses a state and uses a so-called scalar predicate (a predicate that can be modified with degree expressions, such as *very much X*; *more than X*), which is a judgment because it can be paraphrased with “I believe/feel that …”. Fragment (9) gives an example.[4]

(9) Judgment in S1
A *[S1 en ik vind ’t niet meer leuk op die manier te*
and I find it not more nice on that way to
werken ook dat nog ’ns een keer.]
work even that again
‘and I don’t like it any longer to work that way also’
B *nee ja.*
‘no yes’
A *hè?*
‘right?’

---

[4] In our presentation we use the following conventions: first and second segments in the relation are delimited by [S1] and [S2]; the interlocutors are indicated by capitals (A, B etc.); the English translations of the Dutch examples are rather idiomatic translations unless a more precise gloss is needed. In the translation the connective is indicated in capitals.

A	*omdat*	*[S2*	*er*	*geen*	*uh*	*geen*	*wisselwerking*	*is.]*
	because		there	no	eh	no	interaction	is

‘OMDAT there is eh no interaction’

Judgments were considered more subjective than other modalities.

II. Relation type

The causal relation expressed in each fragment was analyzed in terms of domains (Sweetser 1990): content (in which the speaker describes a causal relation in the world), epistemic (in which the speaker infers a conclusion on the basis of an argument) and speech act relations (in which the speaker motivates a speech act). Furthermore, within the content relations we distinguish between volitional and non-volitional relations (see Stukker et al. 2008): Does the relation involve an intentional act or not? Examples are:

(10) Non-volitional content
[S1 De vogelstand gaat hard achteruit] omdat [S2 een hele voedselketen stelselmatig wordt vergiftigd].
‘The bird population decreases fast OMDAT a complete food chain is poisoned systematically.’

(11) Volitional content
A *[S1 dan gingen ze Albert-Jan vragen of ie de achtste in de boot kon zijn] omdat [S2 Bas naar die inauguratie van die pastoor moest].*
‘then they went to ask Albert-Jan whether he could be the eighth man in the boat OMDAT Bas had to go that priest’s inauguration.’

(12) Epistemic
A *[S1 en en wa ik het nu heb dat is geen noodsituatie] want [S2 ik kan donders goed inschatten kwart over vier half vijf dat er dan geen student meer boven gaat kijken.]*
‘and and what I have it now that is not an emergency WANT I can estimate very well quarter past four half past four that no student will go and look upstairs then anymore.’

(13) Speech act
A *[S1 en uh a als iemand mij belt ja dan ben ik er niet] want [S2 ik ben bezig met dit werk en dat moet vandaag af punt].*
‘and uh i if someone calls me yes then I am not in WANT I am busy working at this and it has to be finished today period.’

On the basis of our reasoning in Section 1.2, the causal relations can be ordered from least subjective to most subjective, as follows:

Non-volitional content < Volitional content < Epistemic / Speech act

III. Type of SoC in the first segment (S1)

The SoC is the person responsible for the causal relation that is constructed. There can be either no SoC (as in example (10)), or the SoC is a third person (example (11)), a second person (example (14)) or a first person (examples (12) and (13)).

(14) Second person SoC
Speaker A *en dat is de enige manier via mij krijgen ze hun boeken.*
‘and that is the only way through me they get their books’
Speaker B *ja.*
‘yes’
Speaker A *ik ben de leverancier als het ware.*
‘I am the supplier so to speak’
Speaker A *ja.*
‘yes’
Speaker B *ja ja ja ja ja.*
‘yes yes yes yes yes’
Speaker B *uh koopt u dan ook alleen maar gebonden uitgaven OMDAT dat mooier is in de boekenkast of …*
‘eh does that mean that you only buy hardcover editions BECAUSE that is more beautiful on the bookshelves or …’
Speaker A *nee.*
‘no’
Speaker A *neen niet altijd niet altijd.*
‘no not always not always’

These options can be ordered in degree of subjectivity, as follows:

No SoC < Third person < Second person, First person

In the analysis presented below we only use the distinction between third person SoCs and first/second person SoCs. Cases without a SoC were deemed irrelevant as these express facts. Cases of first and second person SoCs were collapsed. We did not distinguish between 1st and 2nd person SoCs, because both introduce subjectivity in the *hic et nunc* of the speech/writing situation (speaker/writer subjectivity in case of 1st person SoCs and addressee subjectivity in case of 2nd person SoCs).

IV. Linguistic realization of the SoC

The final property we will report on is the linguistic realization of the SoC. We have followed Langacker's (1990) suggestion that an explicit reference to the SoC objectifies the SoC. Consequently, implicit reference to the SoC is considered more subjective than explicit reference:

Explicit reference to the SoC < Implicit reference to the SoC

2.2 Materials

For our analysis we used three corpora. For the written medium we made use of the pilot version of the D-COI corpus, a preparatory project which aimed at producing a blueprint and the tools needed for the construction of a 500-million-word reference corpus of contemporary written Dutch (D-COI 2006). The size of the corpus part that we used was 1,8 million words. We randomly selected 100 occurrences of *omdat* and 100 occurrences of *want*.[5] For the spoken medium we made use of the Corpus of Spoken Dutch (Corpus Gesproken Nederlands, CGN). CGN is a 10 million words corpus of completely digitalized material, annotated in several ways (Oostdijk 2000). From the spontaneous conversations and interviews in this corpus we randomly selected 100 fragments with *want* and 100 fragments with *omdat*. For the chat medium we have used the VU Chat corpus, a small corpus of chat conversations between secondary school children, collected at VU University Amsterdam. The size of this corpus is 217,000 words. From this corpus we selected all occurrences of *omdat* (39 cases) and *want* (90 cases). Because of the limited size of the corpus we had to add occurrences from other chat data: we selected all 27 occurrences of *omdat* in a pilot version of the CONDIV corpus (Grondelaers et al. 2000) and added ten randomly selected occurrences of *want*

---

[5] Some corpus fragments contain more than one instance of the connective under analysis. In that case we have analyzed both instances. Consequently sometimes we have more than 100 instances per corpus per connective. Note that *omdat* can occur in sentence-initial position ("Omdat S1, S2") and in sentence-medial position ("S1, omdat S2"). As we are dealing with backward causals, we have only included the latter type of cases in our corpus.

from the same corpus. Only 12 *omdat*-instances from the CONDIV corpus could be used in the analyses reported below, as the IRC chat in the CONDIV corpus is extremely difficult to interpret and has many instances of *omdat* without an apparently appropriate context.

2.3 Procedure

In our analysis we followed the “complete double coding” strategy (Spooren and Degand 2010), in which the two authors coded the fragments independently and discussed discrepancies. We determined the subjectivity in the corpus examples by analyzing a number of properties of the discourse context (i.e., the segments surrounding the connectives) that provide information on the subjectivity of the relation. First we determined the size of the related segments. Then we analyzed the type of causal relation, the propositional attitude of the first segment, the SoC (if present), and the linguistic realization of the SoC.

2.4 Statistical analysis

As indicated earlier, the frequency of the two connectives differs per medium. The samples from which we collected the fragments also differed in size. For example, *omdat* is more frequent than *want* in the written corpus, and in the chat corpus there were not enough instances of *omdat* to create a sample of 100 occurrences (which was our initial target).

In order to compensate for the difference in size of the samples and the corpora from which these stem, we did not analyze the raw frequencies to test our hypotheses, but the logits of these frequencies. A logit is the natural logarithm of the frequency of a phenomenon, divided by the corpus size minus the frequency of the phenomenon (in formula: ln(frequency/(corpus_size – frequency)). We used the data in Table 1 to estimate the size of the corpus from which our samples were chosen.

To test the hypotheses that differences between *want* and *omdat* generalize over media, we carried out logit analyses, in which contributions from the variables to account for the variation in the data are evaluated in order to establish the best fitting model. Four separate analyses were carried out, one for each indicator of subjectivity (type of relation, propositional attitude in the first segment, type of SoC, linguistic realization of the SoC).

To answer the research question concerning the relative weight of each indicator of subjectivity, a so-called CART (Clustering and Regression Tree) analysis was carried out (Baayen 2008: 148–154). With this analysis we tried to set up a model that predicts whether a fragment uses *omdat* or *want* on the basis of such factors as the type of relation, the type, and linguistic realization of the SoC, the propositional attitude in the first segment and the type of medium.

3 Results

For ease of reading we present the statistical details of the analyses in the appendices. In the main text we will present those parts of the analyses that directly test our four hypotheses. In footnotes we will present additional significant parts of the analysis.

3.1 Type of Relation

Our overall hypothesis is that, irrespective of medium, *want* occurs more often in subjective contexts than *omdat*. For Type of Relation this means that we expect to find more Epistemic/Speech Act relations with *want* than with *omdat*. The results are summarized in Table 2.

Examples (15)–(19) illustrate the findings as summarized in Table 2. Examples (15), (16), and (17) are prototypical examples of *omdat* expressing content (15) and *want* expressing an epistemic (16) and a speech act (17) relation. (18) and

Table 2: Type of relation in spoken, chat and written data, by connective (percentages are column percentages per medium).

		Omdat	Want
Spoken			
	Content	89 (89.9)	40 (40.4)
	Epist./Speech Act	10 (10.1)	59 (59.6)
Chat			
	Content	45 (88.2)	35 (35.0)
	Epist./Speech Act	6 (11.8)	65 (65.0)
Written			
	Content	95 (95.0)	39 (39.0)
	Epist./Speech Act	5 (5.0)	61 (61.0)

Note: Four fragments had relations with different possible readings and were coded as missing.

(19) are non-prototypical cases of *omdat* expressing an epistemic and *want* expressing a content relation.

Fragment 15 is from the spoken corpus, more specifically an interview with a school teacher who explains how he arrived at this school. S1 expresses a volitional action, which is explained in S2; the two segments are connected with *omdat*, expressing a content-volitional relation ("the reason was …").[6]

(15) *Omdat* expressing a volitional content relation
maar [S1 ik ben m wel hier meteen uh op school uh terecht gekomen na mijn examen van de PA].
'but I did m manage uh to go to this school immediately uh after my final examination at the teacher training college'

omdat [S2 mij dat gevraagd werd om hier les te komen geven en ik daar wel trek in had.]
'OMDAT I was asked to teach here and I felt like doing it'

Fragment (16) was taken from a Dutch newspaper story about English football player Tony Adams, who is the SoC and Speaker in this fragment. In S1 Adams (Speaker=SoC) draws a conclusion about someone else's behavior (*he* – notably football player David Beckham) and explains this conclusion on the basis of knowledge of an ongoing state of affairs, signaled by *want*, expressing an epistemic relation.

(16) *Want* expressing an epistemic relation
[S1 Ik weet niet meer wat hij zei maar hij moet het gewaardeerd hebben], want [S2 hij heeft er sindsdien vaak over gesproken]
'I don't know what he said but he must have appreciated it WANT he spoke of it often since then'

Fragment (17) is part of a chat conversation between two middle school students, in which one asks a question and subsequently provides the reason for asking this question. This is a prototypical example of a speech act use of *want* in chat. The relation can be paraphrased as "I ask you what your address is and the reason for my asking (speech act) is that I do not have the address."

---

[6] Note: "ggg" stands for guttural sounds, "xxx" means uninterpretable.

(17) *Want* expressing a speech act relation
maarre tim ... [S1 wat's jou egte adres]
'But eh tim ... what is your real address'

want [S2 die heb ik niej]
'WANT that I don't have'

Fragment (18) is a case of an epistemic relation, but expressed in an *omdat*-construction. It is from the written corpus, and an interviewee is quoted.

(18) *Omdat* expressing an epistemic relation
De oefenmeester, zelf nog een groentje in het Europese topvoetbal, klampt zich maar vast aan de ervaring van vorige week op Old Trafford, toen zijn ploeg de offensieve intenties van Manchester United met verbluffend positiespel ontregelde.
'The trainer, himself a newcomer in European top football, clings to his experience from last week at Old Trafford, when his team disorganized the offensive intentions of Manchester United using astonishing positional play.'

"Het spel van Manchester United ligt ons wel.
'Manchester United's type of play suits us nicely.'

Bovendien [S1 zullen zij wederom op de aanval speculeren], omdat [S2 ze normaal gesproken moeten winnen".]
'Moreover, they will again speculate on attacking, OMDAT normally speaking they have to win".'

Fragment (19) is a volitional content relation from a chat conversation between students. The speaker explains why Miranda's face was completely red, using a *want*-coordination.

(19) *Want* expressing a content relation
A *en Miranda's gezicht was helemaal rood. WANT ze had gemische peeling gehad ofzo.*
'and Miranda's face was completely red. WANT she had had a chemical peeling or something.'

The logit analysis is summarized in Table A1 in the appendix (a short introduction to the interpretation of these tables is provided at the beginning of the appendix). The data are best described with a model containing main effects of Connective and Type of Relation and interactions of Connective * Medium and

Connective * Type of Relation (model 6 in Table A1). The fit of the resulting model is adequate ($\chi^2(4) = 3.43$, $p = .49$). Directly related to our hypothesis is the interaction between Connective and Type of Relation: In *omdat*-fragments there are relatively few Epistemic/Speech Act relations (21 out of 250 relations or 8.4%), in *want*-fragments the majority are Epistemic/Speech Act relations (185 out of 299 relations or 61.9%).[7]

3.2 Propositional Attitude

The Propositional Attitude hypothesis states that irrespective of medium, first segments of *want* fragments more often express an opinion, compared to *omdat* fragments. The data are summarized in Table 3.

Fragment (20), from the spoken corpus, illustrates a judgment in S1, which is the dominant propositional attitude for *want*-connections.

(20) Judgment in S1
[S1 dat is gewoon krankzinnig].
'that is simply insane'

Table 3: Type of propositional attitude in spoken, chat and written data, by connective (percentages are column percentages per medium).

		Omdat	**Want**
Spoken	Judgment	43 (43.4)	54 (54.0)
	Other propositional attitudes	56 (56.6)	46 (46.0)
Chat	Judgment	12 (23.5)	35 (35.0)
	Other propositional attitudes	39 (76.5)	65 (65.0)
Written	Judgment	42 (42.0)	73 (71.6)
	Other propositional attitudes	58 (58.0)	29 (28.4)

Note: One case is missing because it allowed for multiple readings.

---

[7] The parameter estimates for model 6 (in Appendix, Table A2) allow for an interpretation of the other effects. The main effect of Connective shows that there are somewhat more *want*-fragments than *omdat*-fragments. The main effect of Type of Relation indicates that overall there are less instances of Epistemic and Speech Act relations compared to Content relations. The interaction between Connective and Medium reflects the fact that there are relatively few fragments with *want* in the Written medium (relative to the size of the corpora).

want [S2 als hij uhm mensen goed inschat moet ie ook weten dat ik m'n uiterste best doe om dat zo snel mogelijk voor elkaar te krijgen.]
'WANT if he uhm is such a good judge of character then he should also know that I am doing my very best to take care of that as soon as possible.'

In (21), a *want*-construction without a judgment in S1, taken from a chat-conversation, a pupil explains why he cannot always watch his favorite TV-series. This is a non-volitional causal relation, which even could have been expressed by a *doordat*.

(21) *Want* expressing a non-volitional content relation
alleen [S1 kan t niet altijd kijken] want [S2 mn vader wil altijd journaal kijken]
'only cannot always watch it WANT my father always wants to watch the news.'

(22) shows the typical *omdat*-pattern: the propositional attitude in S1 is other than judgment and is presented in a construction expressing a volitional relation.

(22) *Omdat* expressing a non-judgment in S1
[S1 Drie vrouwen van middelbare leeftijd worden achterna gezeten] omdat [S2 ze het waagden te protesteren.]
'Three middle-aged women are chased OMDAT they dared to protest.'

Fragment (23), from a newspaper, shows a non-typical and infrequent occurrence of an *omdat*-construction with a clear judgment, expressing an epistemic relation.

(23) *Omdat* expressing a judgment in S1
[S1 Sint Maarten kan hier niet afgebeeld zijn] omdat [S2 het desbetreffende portaal (...) aan de martelaren gewijd is en dat was Maarten niet.]
'It cannot be Saint Martin who is depicted here OMDAT the portal in question (...) is devoted to martyrs and Saint Martin wasn't a martyr.'

The logit-analysis is summarized in Table A3 (Appendix). The data are best described with a model containing a main effect of Connective and interactions of Connective * Medium, Connective * Propositional Attitude and Medium * Propositional Attitude (model 7 in Table A3). The fit of the resulting model is adequate ($\chi^2(2) = 4.34$, $p = .11$). Directly related to our hypothesis is the significant interaction between Connective and Propositional Attitude: In *omdat*-fragments there are relatively few judgments (97 out of 250 relations or 38.8%), in *want*-fragments judgments are the majority (162 out of 302 relations or 53.6%).[8]

3.3 SoC-type

Our next analysis concerns the relationship between SoC, medium, and connective. In this analysis we compared 1st and 2nd person SoC on the one hand with 3rd person SoC on the other, see Table 4. As there are relatively few 2nd person SoCs in the medium, we grouped them together with 1st person SoCs. In the analysis we disregarded first segments without a SoC (facts) and fragments in which the SoC is a secondary speaker (a quoted character).

The logit analysis shows that the data in Table 4 are best described by a model containing all three variables (Table A5 in the Appendix shows that the fit of this model was perfect: $\chi^2(0) = 0.00$). Central to our research question is the two-way interaction between Connective and SoC, which means that the predominance of first/second person SoCs is much larger for *want*-fragments

Table 4: Type of SoC in spoken, chat and written data, by connective (percentages are column percentages per medium).

		Omdat	**Want**
Spoken	1st/2nd person	76 (82.6)	73 (76.8)
	3rd person	16 (17.4)	22 (23.2)
Chat	1st/2nd person	39 (81.2)	90 (91.8)
	3rd person	9 (18.8)	8 (8.2)
Written	1st/2nd person	19 (24.7)	57 (72.2)
	3rd person	58 (75.3)	22 (27.8)

Note: 64 cases are missing (either the first segment does not have a SoC because it is a fact, or the SoC is a quoted character).

---

[8] The parameter coefficients for model 7 (table A4) suggest the following interpretation for the other effects. The main effect of Connective shows that there are relatively more *want*-fragments than *omdat*-fragments in this analysis. The interaction between Connective and Medium reflects the fact that there are relatively few fragments with *want* in the Written medium (relative to the size of the media). The interaction between Medium and Propositional Attitude reflects the fact that the predominance of other propositional attitudes over s is largest in the chat medium.

than for *omdat*-fragments (in conformity with our hypothesis), see examples (16), (17), and (18). However, this result is modified by a significant three-way interaction between Medium * SoC * Connective showing that this predominance of first/second person SoCs in *want* fragments is even stronger in the written medium, because there *omdat*-fragments generally occur with third person SoCs.[9]

Overall, *want* shows a consistent pattern over the media: it has predominantly 1st person SoCs. *Want*-cases in the chat medium have an even higher amount of 1st person SoCs than in the other media, see example (17), and (24) below.

(24) *[S1 geen praatjes he kleine man]*
'no big mouth ay little man'

want [S2 anders zet ik je in der prullenbakk]
'WANT otherwise I will put you in the wastepaper basket'

Omdat has a clearly different behavior: in the spoken and chat medium, it resembles *want* with its abundance of 1st person SoCs as in example (9); in the written medium, however, this predominance has reversed, in that there are mainly 3rd person SoCs, as illustrated in example (25) below.

(25) *[S1 Maar de technocraten wilden per se aan de macht blijven], omdat [S2 ze hun economisch model in stand wilden houden].*
'But the technocrats absolutely wanted to maintain power, OMDAT they wanted to hold on to their economic model.'

---

[9] The main effect of Connective shows that there are relatively more *want*-fragments than *omdat*-fragments (note that this analysis is restricted to fragments in which SoCs (first/second person or third person) occur. The main effect of Medium reflects the fact that there are relative few fragments with a first or third person conceptualizer in the written medium, indicating that the written medium had relatively many factual relations like (10). The main effect of SoC shows that overall there were relatively less third person than first/second person SoCs. The interaction between Connective and SoC can be interpreted as follows: the predominance of *want* fragments in which conceptualizers occur is higher in the written medium than in the other two media.

The two-way interaction between Medium and SoC can be interpreted as follows: in the spoken medium and the chat medium there are relatively few third person SoCs compared to first/second person SoCs, whereas in the written medium there are more third person SoCs than first/second person SoCs. In other words, we see more first and second person SoCs in chat and spoken language.

Table 5: Linguistic marking of SoC in spoken, chat and written data, by connective (percentages are column percentages per medium).

		Omdat	Want
Spoken	explicit	67 (71.3)	57 (58.2)
	implicit	27 (28.7)	41 (41.8)
Chat	explicit	30 (62.5)	52 (53.1)
	implicit	18 (37.5)	46 (46.9)
Written	explicit	48 (53.9)	37 (36.6)
	implicit	41 (46.1)	64 (63.4)

Note. 25 cases are missing (the first segment does not have a SoC because it expresses a fact).

3.4 Linguistic marking of the SoC

Our next analysis concerns the relationship between connective, medium, and linguistic realization of the SoC. Remember that a SoC (if present) can be referred to explicitly in the first segment or that it can remain implicit. The latter is judged to be more subjective than the former. For that reason it is expected that implicit SoCs occur more often in the first segment of *want*-fragments than in that of *omdat*-fragments. The data are summarized in Table 5.

The logit analysis shows that the data are best described with a model containing all main effects and all two-way interactions. The fit of the resulting model is acceptable ($\chi^2(2) = 0.46$, $p = 0.80$). The parameter estimates for this model are presented in Table A8 (Appendix). Directly of interest for our hypothesis is the two-way interaction between Connective and Linguistic Marking: The predominance of explicit markings is less strong for *want* than for *omdat*. In other words, and as predicted, *want*-fragments have more implicit marking of the SoC than *omdat*-fragments (*omdat*: 86 out of 231 cases or 37.2%; *want*: 151 out of 297 or 50.8%).[10]

---

[10] The main effect of Connective reflects the overall predominance of *want* fragments in this analysis of linguistic marking. The main effect of Medium reflects the fact that there are relatively few fragments with implicit or explicit conceptualizers in the chat medium. The main effect of Linguistic Marking shows that overall explicit marking is predominant. The interaction of Connective and Medium reflects the fact that the predominance of *want* fragments in this analysis does not hold for the written medium.

Example (20), repeated here for convenience, is a clear prototypical example of *want* with a (first person) implicit SoC, expressing a judgment in S1.

(20) First person implicit SoC in S1
[S1 dat is gewoon krankzinnig].
'that is simply insane'

want [S2 als hij uhm mensen goed inschat moet ie ook weten dat ik m'n uiterste best doen om dat zo snel mogelijk voor elkaar te krijgen.]
'WANT if he uhm is such a good judge of character then he should also know that I am doing my very best to take care of that as soon as possible'

Fragment (26) shows a *want*-case from the spoken corpus, with a first person explicit SoC.

(26) *Want* with explicit 1st person SoC
en dan wil je mensen zo snel mogelijk helpen.
'and then you want to help people as soon as possible'

en en [S1 wa ik het nu heb dat is geen noodsituatie] want [S2 ik kan donders goed inschatten kwart over vier half vijf dat er dan geen student meer boven gaat kijken.
'and and wha I am having it now that is not an emergency because I can estimate damned well quarter past four half five that no student is going to look upstairs anymore'

3.5 The relative importance of the indicators of subjectivity

In the previous subsections we have shown that all four indicators of subjectivity play a significant role in characterizing the difference between *want*- and *omdat*-fragments. An important question is whether each characteristic is equally important for this characterization. In order to answer that question we have made use of a so-called CART (classification and regression tree) analysis, described in Baayen (2008: 148–154). A CART-analysis predicts the classification of an object on the basis of a number of factors. In our case, it produces a tree (see Figure 1) that outlines a decision procedure for determining the realization of *want*. Each split in the tree is labelled with a decision rule. The leaf nodes of the tree specify a partition of the data into a series of non-overlapping subsets. The analysis stops when a new split does not have enough explanatory value or when there are too few observations left to make a new split. The length of the branches is an

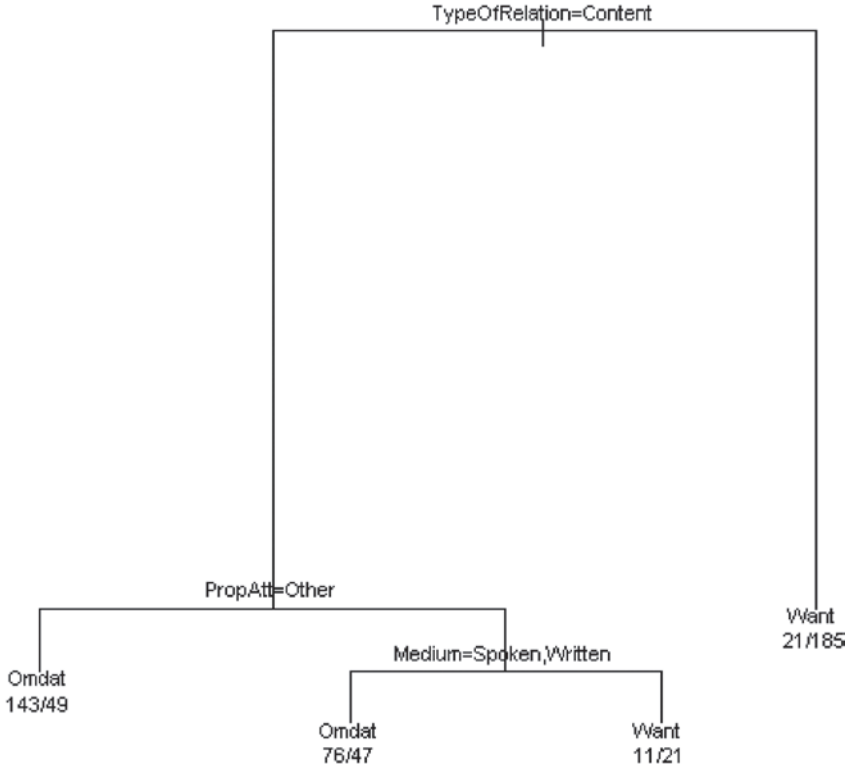


Fig. 1: A decision procedure for determining the realization of *want*, as produced by a CART (classification and regression tree) analysis (Baayen 2008: 148–154).

indication of the explanatory value of a split (the longer the branch, the higher the explanatory value).

The tree should be read as follows. The first decision concerns the type of relation. If it is not a content relation then the connective is predicted to be *want*. This leads to a correct prediction for 185 out of 206 cases. If it is a content relation then the next decision concerns the propositional attitude of the first segment. If that attitude is not a judgment the connective is predicted to be *omdat* (143 correct predictions out of 182 cases). The final decision concerns the medium: if the medium is written or spoken (as opposed to chat), the connective is predicted to be *omdat* (76 correct predictions, 47 incorrect predictions), if it is chat, it is predicted to be *want* (21 correct predictions, 11 incorrect). Note that groupings are determined by the program and were not predetermined by the investigators.

Overall, the analysis makes 425 correct predictions out of 553 cases or 76.9% correct. Compared to a minimal model in which the most frequent connective (*want*) is the only predictor, this leads to an improvement of 22.3%.

The analysis clearly shows that Type of Relation is by far the most important predictor of the connective choice. Other factors only matter for fine-tuning the prediction of the connective for content relations. We also see that Linguistic Marking and Type of SoC do not contribute substantively to the quality of the prediction.

4 Discussion

In this paper we investigated the system behind the meaning and use of backward causal connectives in discourse. Our starting point was the distinction between content / semantic / objective relations versus epistemic-speech act / pragmatic / subjective relations, which is well-known in text linguistics (ever since van Dijk 1979), functional (Degand 2001; Martin 1992) and cognitive linguistic (Sweetser 1990), and from work on (causal) connectives, as well as from cognitive approaches to coherence relations (Sanders et al. 1992; Sanders 1997). In addition, several linguists have suggested that the choices made by speakers of languages like Dutch *omdat* versus *want,* German *weil* versus *denn*, and French *parce que* versus *car* and *puisque*, are not only systematical, but also exactly reflect this distinction (see Pit 2006).

Here, we have argued in favor of a principle of subjectivity to explain the systematic differences between connectives. We have set out on a corpus investigation of the meaning and use of the Dutch connective pair *want* and *omdat*. The corpus that we analyzed was larger and more varied in media than in previous work, and we analyzed it using state-of-the-art statistical methods. But that was not the only innovative aspect of our study. We adopted an integrative empirical approach in order to solve two major challenges in the field. First, we decomposed the complex construct of subjectivity in several characteristics, which were analyzed separately. Second, we argued that it is important to investigate whether insights from existing work on written corpora can be generalized to other media, especially because claims concerning *cognitive reality* are at stake. After all, our most natural and spontaneous way to communicate is not merely via discourse, but through *spoken* discourse. Now that spoken corpora have become available in many languages, it is possible to test hypotheses against spoken corpus data. Chat data also provide an interesting case, because they are spontaneous, like spoken language, but chat lacks the immediate feedback, the intonation, and prosody of face to face spontaneous conversations. A more specific reason to be interested in more spontaneous, less-edited language data is directly related to the interpretation of causal connectives as acts of categorization: How "basic" is this act? Are the distinctions only realized by highly proficient language users in a production context with many editing opportunities? Or are the same differences realized in the totally different production context of spontaneous conversations, characterized by limiting time constraints and few planning and editing options?

We analyzed a corpus of *omdat*- and *want*-cases from written, spoken and chat discourse. We expected subjectivity to go across the modalities of written, spoken and chat language. Therefore our main hypothesis was that *want* occurs in more subjective contexts than *omdat*, irrespective of the medium. We formulated four specific hypotheses on the way in which the connectives *want* and *omdat* would show differences in terms of subjectivity. When we summarize the main results of our corpus research, we can say that all four hypotheses repeated below were confirmed. Across all media

- *want* is used more often to express subjective relations (epistemic, speech act) than *omdat* (hypothesis 1);
- *want* is used more often to support a judgment than *omdat* (hypothesis 2);
- *want* is used more often with first and second person SoCs than *omdat* (hypothesis 3);
- *want* is used more often with an implicit SoC than *omdat* (hypothesis 4).

We can conclude from this that the subjectivity hypothesis and the generalization over media hypothesis are supported by the data: *Want* and *omdat* show a clearly different pattern over all media we investigated.

In addition to the hypotheses, we formulated two explorative research questions. The first was whether the medium affects the degree of subjectivity. The results are not unequivocal. We did not find a significant interaction between Medium and Type of Relation. We did find a significant interaction between Medium and Propositional Attitude, but it is not easy to interpret that interaction: Contrary to expectation there are relatively few judgments in chat. A plausible explanation is that chat is subjective not because of its abundance of judgments but because it has relatively many speech act relations (spoken: 20 out of 198 or 10.1% speech act relations; written: 11 out of 200 or 5.0% speech act relations; chat: 42 out of 151 or 27.8% speech act relations): Even though speech act relations can be considered very subjective, the first segment in a speech act relation is not a judgment. There was also a significant interaction between Medium and SoC: the written fragments are heavily dominated by Third Person SoC's, suggesting that written texts have many objective reports of causal relations. Finally, there was also a significant interaction between Medium and Linguistic Realization of the SoC: surprisingly the written corpus has relatively many implicit SoC's. However, those implicit SoC's predominantly are first person SoC's in the written corpus, which constitute the minority. In sum, although the various interactions with Medium can be interpreted, it is not the case that there is a straightforward relationship between Medium and Subjectivity. This obviously is an area for further research.

The second explorative question concerns the relative strength of the various subjectivity characteristics: are all operationalizations of subjectivity (as formulated in the hypotheses) equally important? The CART-analysis provides a clear answer to that question: Type of Relation (content versus epistemic/speech act) is by far the most important predictor of the connective. The second important factor is propositional attitude of the first segment: If that attitude is not a judgment the connective is almost certainly *omdat*. The final decision concerns the medium: if the medium is written or spoken, the connective is predicted to be *omdat*.

Hence, our corpus study clearly corroborates the hypotheses formulated for the two causal connectives we have studied. There are substantial differences in the meaning and use of *omdat* and *want*: *want* is subjective in that it typically signals an epistemic or speech act relation, whereas *omdat* typically signals a content relation. In addition, *want* often has a judgment in the first segment, a first person conceptualizer, which is more often implicitly realized than *omdat*. This pattern roughly replicates earlier results reported by Pander Maat and Degand (2001) and Pit (2006) on the distribution of these connectives in written language. These differences between *want* and *omdat* survive across media, as they are found in spontaneous conversations, chat communication as well as written text.

What is the theoretical interpretation of our main findings? First of all, it shows the relevance of the notion of subjectivity, which we have defined, operationalized and actually used in corpus analysis in such a way that it indeed explains the differences between the two connectives. However, there is a fundamental issue to address here. Even though distinctions like objective-subjective or content-epistemic/speech act seem relevant across languages, many studies have observed that the causal categories are not *always* reflected in connective use. Like our study, earlier corpus studies have shown how, in a minority of cases, causal connectives that seem to specialize in one type of relation, *can* in fact be used to express other causal categories. For instance, even though Dutch *want* specializes in expressing epistemic relations, it can be and – as our study and earlier corpus studies show – actually *is* used to express content relations (Pit 2006). Similar observations exist for French and German connectives (Stukker and Sanders 2012). Apparently we are not dealing with black-and-white distinctions, but rather with tendencies. A crucial question is what consequences such empirical observations should have for a theory of causal connectives as categorization devices.

In our view, the conceptual basis of linguistic categories offers a natural explanation of the fact that causal connectives in actual language use do not always directly reflect conceptual categories of causality (Stukker et al. 2008, 2009). A crucial insight here is that causal categories show prototypicality structure. Classical categorization theory (Rosch 1973) argued that robins are better examples of the category of birds than ospreys and puffins are. Similarly, connective uses that seem counter-examples against our categorization hypothesis, should be regarded as less prototypical members of the same category to which the "normal" uses belong. In that respect, it is not a coincidence that we often used terms like prototypical and less prototypical when characterizing patterns and examples presented above. More specifically, we expect the non-prototypical uses to have a different status in the language user's mental representation of the connectives' meaning and use (Bybee 2007; Stukker et al. 2008, 2009).

Such a position requires a more detailed analysis in various respects, which goes beyond the scope of this paper. We briefly mention two. One is the crosslinguistic comparison of patterns in connective meaning and use. Results of a meta-analysis of existing corpus studies indeed suggest highly similar patterns indicating a prototypicality structure for French, German, and Dutch (Stukker and Sanders 2012). A second issue is to show in (qualitative) linguistic analyses how exactly *want* and *omdat* result in different conceptual representations (Sanders et al. 2012), and perhaps even more importantly, how non-prototypical examples still show resemblance to their prototype. For instance, when we observe that, in a minority of cases, *omdat* can express epistemic relations, it is important to explain that this use of *omdat* is not a coincidence, but that the *omdat*-context shows, for instance, more objective characteristics than a *want*-context does (Degand 2001; Sanders and Spooren 2013; Stukker and Sanders 2012). This could even be done using automated large-scale quantitative analyses like the one Bestgen et al. (2006) used to study the subjective nature of the context of *omdat* and *want* in newspaper language.

In conclusion, we believe that causal categories are fundamental to human cognition and natural language at the discourse level. Causality and subjectivity are two cognitive principles that organize our knowledge of coherence relations. Notions like causality and subjectivity can help us explain the system and use of causal relations and their linguistic expressions in everyday language use, and following the methodological principle of converging evidence (cf. Gonzalez-Marquez et al. 2007, and contributions to this volume), we have shown elsewhere that they explain the acquisition of connectives and relations (Evers-Vermeul and Sanders 2009, 2011; Spooren and Sanders 2008; Sanders and Spooren 2009 and the references cited there) as well as discourse processing and representation (Canestrelli et al. 2013). Furthermore, it seems worthwhile to feed theories of connectives and coherence relations with corpus studies of spontaneous language use in communicative situations that allow for direct interaction; for one thing, we have never before seen so many attested speech act relations as in our chat corpus. Systematic comparison of various communicative situations is imperative. Finally, we analyzed the causal connections in terms of detailed characteristics and subsequently investigated whether and to what extent they co-occur. We made use of statistical methods specifically suitable for hypothesis testing in natural language corpora. Such an enterprise provides new insights into the notion of subjectivity. We would like to see such methods used in studies that could reveal whether other, less-related languages, also encode such categories of causality, or other types of coherence relations. Recent results on Mandarin Chinese (Li et al. 2013), and results from studies looking into parallel corpora of translated texts (Cartoni et al. 2013) are promising.

Acknowledgments: This research was enabled by the Netherlands Organization for Scientific research, through NWO-VICI-grant 277-70-003, awarded to Ted Sanders. We thank the editors and two anonymous reviewers for their constructive suggestions on a previous version of this paper. Earlier versions of this paper were presented at the IPrA conference in Manchester (July 2011) and the UK Cognitive Linguistics Conference in Brighton (2008), as well as at the BeNeCLA-conference in Antwerp (2010). We thank the colleagues at these meetings for providing us with feedback.

References

Anscombre, Jean-Claude & Oswald Ducrot. 1983. *L'argumentation dans la langue*. Brussels: Pierre Mardaga.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Banfield, Ann. 1982. *Unspeakable sentences: Narration and representation in the language of fiction*. Boston: Routledge & Kegan Paul.

Bekker, Birgit. 2006. *De feiten verdraaid. Over tekstvolgorde, talige markering en sprekerbetrokkenheid* [The twisted facts: About text order, linguistic marking and speaker involvement]. Tilburg: Tilburg University dissertation.

Bestgen, Yves, Liesbeth Degand & Wilbert Spooren. 2006. Towards automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes* 41(2). 175–194.

Breindl, Eva & Maik Walter. 2009. *Der Ausdruck von Kausalität im Deutschen: Eine korpusbasierte Studie zum Zusammenspiel von Konnektoren, Kontextmerkmalen und*

Diskursrelationen (Arbeiten und Materialien zur deutschen Sprache). Mannheim: Institut für Deutsche Sprache.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Canestrelli, Anneloes, Willem Mak & Ted Sanders. 2013. Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes* 28(9). 1394–1413.

Cartoni, Bruno, Sandrine Zufferey & Thomas Meyer. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse* 4(2). 65–86.

Chafe, Wallace. 1994. *Discourse, consciousness and time*. Chicago: Chicago University Press.

Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.

D-COI. 2006. The Dutch Language Corpus Initiative. http://lands.let.ru.nl/projects/d-coi/ (accessed 28 June 2010).

Degand, Liesbeth. 2001. *Form and function of causation: A theoretical and empirical investigation of causal constructions in Dutch*. Louvain: Peeters.

Degand, Liesbeth & Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In Arie Verhagen & Jeroen van de Weijer (eds.), *Usage based approaches to Dutch,* 175–199. Utrecht: LOT.

De Smet, Hendrik & Jean-Christophe Verstraete. 2006. Coming to terms with subjectivity. *Cognitive Linguistics* 17. 365–392.

Diessel, Holger & Katja Hetterle. 2011. Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. In Peter Siemund (ed.), *Linguistic universals and language variation*, 23–54. Berlin & New York: De Gruyter Mouton.

van Dijk, Teun. 1979. Pragmatic connectives. *Journal of Pragmatics* 3. 447–456.

Evers-Vermeul, Jacqueline, Liesbeth Degand, Benjamin Fagard & Liesbeth Mortier. 2011. Historical and comparative perspectives on subjectification: A corpus-based analysis of Dutch and French causal connectives. *Linguistics* 49(2). 445–478.

Evers-Vermeul, Jacqueline & Ted Sanders. 2009. The emergence of Dutch connectives; how cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language* 36(4). 829–854.

Evers-Vermeul, Jacqueline & Ted Sanders. 2011. Discovering domains: On the acquisition of causal connectives. *Journal of Pragmatics* 43(6). 1645–1662.

Ford, Celia E. 1993. *Grammar in interaction: Adverbial clauses in American English conversations*. Cambridge: Cambridge University Press.

Frohning, Dagmar. 2007. *Kausalmarker zwischen Pragmatik und Kognition: Korpusbasierte Analysen zur Variation im Deutschen*. Tübingen: Niemeyer.

Gonzalez-Marquez, Maria, Irene Mittelberg, Seana Coulson & Michael J. Spivey. 2007. *Methods in cognitive linguistics*. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan T. 2012. Corpus linguistics, theoretical linguistics, and cognitive/ psycholinguistics: towards more and more fruitful exchanges. In Joybrato Mukherjee & Magnus Huber (eds.), *Corpus linguistics and variation in English: Theory and description*, 41–63. Amsterdam: Rodopi.

Grondelaers, Stefan, Katrien Deygers, Hilde van Aken, Vicky van den Heede & Dirk Speelman. 2000. Het CONDIV-corpus geschreven Nederlands [The CONDIV corpus of written Dutch]. *Nederlandse Taalkunde* 5(4). 356–363.

Groupe λ. 1975. Car, parceque, puisque. *Revue Romane* 10. 248–280.

Günthner, Susanne. 1993. '… weil – man kann es ja wissentschaftlich untersuchen' – Diskurspragmatische Aspekte der Wortstellung in WEIL-Sätzen. *Linguistische Berichte* 143. 37–55.

Huiskes, Mike. 2010. *The role of the clause for turn-taking in Dutch conversations*. Utrecht: Utrecht University dissertation.

Kehler, Andrew. 2002. *Coherence, reference and the theory of grammar*. Chicago: University of Chicago Press.

Keller, Rudi. 1995. The epistemic *weil*. In Dieter Stein & Susan Wright (eds.), *Subjectivity and subjectification: Linguistic perspectives*, 16–30. Cambridge: Cambridge University Press.

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

Knott, Alistair & Ted Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics* 30. 135–75.

Lakoff, George. 1987. *Women, fire and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Langacker, Ronald. 1990. Subjectification. *Cognitive Linguistics* 1. 5–38.

Li, Fang, Jacqueline Evers-Vermeul & Ted Sanders. 2013. Subjectivity and result marking in Mandarin: A corpus-based investigation. *Chinese Language and Discourse* 4. 74–119.

Lyons, John R. 1977. *Semantics*. Cambridge: Cambridge University Press.

Lyons, John R. 1995. *Linguistic semantics: An introduction*. Cambridge: Cambridge University Press.

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

Martin, James R. 1992. *English text: System and structure*. Amsterdam & Philadelphia: John Benjamins.

Nuyts, Jan. 2012. Notions of (inter)subjectivity. *English Text Construction* 5(1). 53–76.

Oostdijk, Nelleke. 2000. The Spoken Dutch Corpus Project. *The ELRA Newsletter* 5(2). 4–8.

Pander Maat, Henk & Liesbeth Degand. 2001. Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics* 12(3). 211–245.

Pander Maat, Henk & Ted Sanders. 2000. Domains of use or subjectivity? The distribution of three Dutch causal connectives explained. In Elisabeth Couper-Kuhlen & Bernd Kortmann (eds.), *Cause, condition, concession and contrast: Cognitive and discourse perspectives*, 57–81. Berlin & New York: Mouton de Gruyter.

Pander Maat, Henk & Ted Sanders. 2001. Subjectivity in causal connectives: An empirical study of language in use. *Cognitive Linguistics* 12(3). 247–273.

Pasch, Renate. 1983. Die Kausalkonjunktionen "da", "den", und "weil": drei Konjunktionen – drei lexikalische Klassen. *Deutsch als Fremdsprache* 20. 332–337.

Persoon, Ingrid, Ted Sanders, Hugo Quené & Arie Verhagen. 2010. Een coördinerende *omdat*-constructie in gesproken Nederlands? Tekstlinguïstische en prosodische aspecten [A coordinating because-construction in spoken Dutch? Text-linguistic and prosodic aspects]. *Nederlandse Taalkunde* 1. 259–282.

Pit, Mirna. 2006. Determining subjectivity in text: The case of backward causal connectives in Dutch. *Discourse Processes* 4. 151–174.

Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4. 328–350.

Sanders, José. 2010. Intertwined voices: Journalists' modes of representing source information in journalistic subgenres. *English Text Construction* 3(2). 226–249.

Sanders, José, Ted Sanders & Eve Sweetser. 2012. Responsible subjects and discourse causality. How mental spaces and perspective help identifying subjectivity in causal connectives. *Journal of Pragmatics* 44(2). 191–213.

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. 2009. Causal categories in discourse – Converging evidence from language use. In Ted Sanders & Eve Sweetser (eds.), *Causal categories in discourse and cognition*, 205–246. Berlin: Mouton de Gruyter.

Sanders, Ted & Wilbert Spooren. 2013. Exceptions to rules: a qualitative analysis of backward causal connectives in Dutch naturalistic discourse. *Text & Talk* 33(3). 399–420.

Sanders, Ted, Wilbert Spooren & Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15(1). 1–35.

Sanders, Ted & Eve Sweetser (eds.). 2009. *Causal categories in discourse and cognition*. Berlin: Mouton de Gruyter.

Sanders, Ted, José Sanders & Eve Sweetser. 2009. Causality, cognition and communication: A mental space analysis of subjectivity in causal connectives. In Ted Sanders & Eve Sweetser (eds.), *Causal categories in discourse and cognition,* 21–60. Berlin: Mouton de Gruyter.

Schlenker, Philippe. 2004. Context of thought and context of utterance: A note on free indirect discourse and the historical present. *Mind & Language* 19(3). 279–304.

Spooren, Wilbert & Liesbeth Degand. 2010. Coding coherence relations: reliability and validity. *Corpus Linguistics & Linguistic Theory* 6(2). 241–266.

Spooren, Wilbert & Ted Sanders. 2008. The acquisition of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics* 40(12). 2003–2026.

Spooren, Wilbert, Ted Sanders, Mike Huiskes & Liesbeth Degand. 2010. Subjectivity and causality: A corpus study of spoken language. In John Newman & Sally Rice (eds), *Empirical and experimental methods in cognitive/functional research,* 241–255. Stanford CA.: CSLI.

Stukker, Ninke & Ted Sanders. 2012. Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics* 44(2). 169–190.

Stukker, Ninke, Ted Sanders & Arie Verhagen. 2008. Causality in verbs and in discourse connectives. Converging evidence of cross-level parallels in Dutch linguistic categorization. *Journal of Pragmatics* 40. 1296–1322.

Stukker, Ninke, Ted Sanders & Arie Verhagen. 2009. Categories of subjectivity in Dutch causal connectives: a usage-based analysis. In Ted Sanders & Eve Sweetser (eds.), *Causal categories in discourse and cognition*, 119–171. Berlin & New York: Mouton de Gruyter.

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

Traugott, Elizabeth. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 57. 33–65.

Traugott, Elizabeth. 1995. Subjectification in grammaticalization. In Dieter Stein & Susan Wright (eds.), *Subjectivity and subjectivisation: Linguistic perspectives*, 31–54. Cambridge: Cambridge University Press.

Verhagen, Arie. 2005. *Constructions of intersubjectivity: Discourse, syntax and cognition.* Oxford: Oxford University Press.

Verhagen, Arie & Suzanne Kemmer. 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27. 61–82.

Vis, Kirsten. 2011. *Subjectivity in news discourse: A corpus-linguistic analysis of informalization*. Amsterdam: VU University Amsterdam dissertation.

Volodina, Anna. 2011a. *Konditionalität und Kausalität im Diskurs: Eine korpuslinguistische Studie zum Einfluss von Syntax und Prosodie auf die Interpretation komplexer Äußerungen*. Tübingen: Narr.

Volodina, Anna. 2011b. Sweetsers Drei-Ebenen-Theorie: Theoretische Überlegungen vor dem Hintergrund einer korpuslinguistischen Studie über konditionale und kausale Relationen. In Gisella Ferraresi (ed.), *Konnektoren im Deutschen und im Sprachvergleich: Beschreibung und grammatische Analyse*, 127–155. Tübingen: Narr.

Wegener, Heide. 2000. *Da, denn* und *weil* – der Kampf der Konjunktionen: Zur Grammatikalisierung im kausalen Bereich. In Rolf Thieroff, Matthias Tamrat, Nanna Fuhrhop & Oliver Teuber (eds.), *Deutsche Grammatik in Theorie und Praxis*, 69–81. Tübingen: Niemeyer.

Zufferey, Sandrine, 2010. *Lexical pragmatics and theory of mind: The acquisition of connectives*. Amsterdam & Philadelphia: John Benjamins.

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

Appendix

Introduction: How to read the tables

Each logistic regression analysis has resulted in two tables: a summary of the analyses and an estimate of the parameters in the resulting model. In the summary, a progressively more complex model is tested for its fit to the data. For example, the first line in Table A1 states that a model with only the constant as a predictor does not fit the data well, because its χ^2 is highly significant ($p < .001$). The second line states that adding the factor “Connective” improves the model significantly ($\chi^2(1) = 30.31$, $p < .001$), but the resulting model is still not a very good fit ($\chi^2(10) = 286.32$, $p < .001$). The table shows in line 7 that adding the interaction Medium*Type of relation does not improve the model compared to the model specified in line 6. That is why the model specified in line 6 is the resulting model.

Table A2 gives the parameters for this model. Positive estimates indicate that the odds that the fragment is of the described type go up, compared to the reference category, whereas negative estimates indicate that the odds go down. For example, the fifth line states that when the relation type is Epistemic/SpeechAct, the estimate is –2.39 and that this estimate significantly differs from zero ($p < .01$). This means that the number of epistemic/speech act relations is significantly lower than that of the reference category, content relations.

Analysis 1: Type of relation

Table A1: Summary of logit analysis

Component	χ^2 model	df	p model	χ^2 factor	df	p factor
1. constant	316.63	11	<.001			
2. + Connective	286.32	10	<.001	30.31	1	<.001
3. + Medium	280.88	8	<.001	5.44	2	0.07
4. + Type of Relation	246.31	7	<.001	34.57	1	<.001
5. + Connective*Medium	188.31	5	<.001	58.00	2	<.001
6. + Connective*Type of Relation	3.43	4	0.49	184.88	1	<.001
7. + Medium*Type of Relation	1.97	2	0.37	1.46	2	0.48
8. + Connective*Medium*Type of Relation	0.00	0	1.00	1.97	2	0.37

Table A2: Parameter estimates for model 6 (Table A1).

Coefficients	Estimate	SE	z value	p
(Intercept)	−7.65	0.10	−74.26	<.01
Connect:Want	0.27	0.16	1.68	n.s.
Medium:Chat	−0.16	0.17	−0.92	n.s.
Medium:Written	0.57	0.14	4.01	<.01
RelType:Epist/SA	−2.39	0.23	−10.48	<.01
Connect:Want;Medium:Chat	−0.31	0.22	−1.37	n.s.
Connect:Want;Medium:Written	−1.48	0.20	−7.38	<.01
Connect:Want;RelTtype:Epist/SA	2.87	0.26	11.17	<.01

Analysis 2: Propositional attitude

Table A3: Summary of loglit analysis

Component	χ^2 model	df	p model	χ^2 factor	df	p factor
1. constant	143.01	11	<.001			
2. + Connective	112.76	10	<.001	30.25	1	<.001
3. + Medium	107.12	8	<.001	5.64	2	0.06
4. + Prop.Att.	105.02	7	<.001	2.10	1	0.15
5. + Connective*Medium	46.65	5	<.001	58.37	2	<.001
6. + Connective*Prop.Att.	34.49	4	<.001	12.16	1	<.001
7. + Medium*Prop.Att.	4.34	2	0.11	30.15	2	<.001
8. + Connective*Medium*Prop.Att.	0.00	0	1.00	4.34	2	0.11

Table A4: Parameter estimates for model 7 (Table A3).

Coefficients	Estimate	SE	z value	p
(Intercept)	−8.50	0.14	−58.65	<.01
Connect:Want	1.54	0.17	8.95	<.01
Medium:Chat	−0.78	0.24	−3.22	<.01
Medium:Written	0.76	0.18	4.15	<.01
PropAtt:NoJudg	0.44	0.17	2.56	<.05
Connect:Want;Medium:Chat	−0.15	0.23	−0.64	n.s.
Connect:Want;Medium:Written	−1.54	0.20	−7.56	<.01
Connect:Want;PropAtt:NoJudg	−0.77	0.18	−4.22	<.01
Medium:Chat;PropAtt:NoJudg	0.89	0.23	3.83	<.01
Medium:Written;PropAtt:NoJudg	−0.34	0.20	−1.66	n.s.

Analysis 3: Type of SoC

Table A5: Summary of logit analysis

Component	χ^2 model	df	p model	χ^2 factor	df	p factor
1. constant	297.21	11	<.001			
2. + Connective	263.47	10	<.001	33.74	1	<.001
3. + Medium	257.27	8	<.001	6.20	2	<.05
4. + SoC	155.58	7	<.001	101.69	1	<.001
5. + Connective*Medium	107.20	5	<.001	48.38	2	<.001
6. + Connective*SoC	85.09	4	<.001	22.10	1	<.001
7. + Medium*SoC	22.99	2	<.001	62.11	2	<.001
8. + Connective*Medium*SoC	0.00	0	1.00	22.99	2	<.001

Table A6: Parameter estimates for model 8 (Table A5).

Coefficients	Estimate	SE	z value	p
(Intercept)	−7.75	0.11	−67.55	<.01
Connect:Want	1.08	0.16	6.56	<.01
Medium:Chat	−0.17	0.20	−0.89	n.s.
Medium:Written	−0.64	0.26	−2.49	<.05
SoC:3rd	−1.56	0.28	−5.67	<.01
Connect:Want;Medium:Chat	−0.11	0.25	−0.44	n.s.
Connect:Want;Medium:Written	−0.33	0.31	−1.07	n.s.
Connect:Want;SoC:3rd	0.36	0.37	0.98	n.s.
Medium:Chat;SoC:3rd	0.09	0.46	0.20	n.s.
Medium:Written;SoC:3rd	2.68	0.38	7.01	<.01
Connect:Want;Medium:Chat;SoC:3rd	−1.31	0.64	−2.06	<.05
Connect:Want;Medium:Written;SoC:3d	−2.43	0.52	−4.69	<.01

Analysis 4: Linguistic marking of the SoC

Table A7: Summary of logit analysis

Component	χ^2 model	df	p model	χ^2 factor	df	p factor
1. constant	122.38	11	<.001			
2. + Connective	92.70	10	<.001	29.68	1	<.001
3. + Medium	86.27	8	<.001	6.43	2	<.05
4. + Ling.Mark.	80.73	7	<.001	5.54	1	<.05
5. + Connective*Medium	25.93	5	<.001	54.80	2	<.001
6. + Connective*Ling.Mark.	16.14	4	<.01	9.79	1	<.01
7. + Medium*Ling.Mark.	0.46	2	n.s.	15.68	2	<.001
8. + Connective*Medium*Ling.Mark.	0.00	0	1.00	0.46	2	n.s.

Table A8: Parameter estimates for model 7 (Table A7).

Coefficients	Estimate	SE	z value	p
(Intercept)	−7.90	0.12	−68.20	<.01
Connect:Want	0.94	0.16	5.94	<.01
Medium:Chat	−0.24	0.19	−1.26	n.s.
Medium:Written	0.26	0.17	1.52	n.s.
Mark:Implic	−0.91	0.18	−4.96	<.01
Connect:Want;Medium:Chat	−0.34	0.23	−1.49	n.s.
Connect:Want;Medium:Written	−1.59	0.21	−7.58	<.01
Connect:Want;Mark:Implic	0.58	0.18	3.15	<.01
Medium:Chat;Mark:Implic	0.27	0.23	1.17	n.s
Medium:Written;Mark:Implic	0.82	0.21	3.85	<.01