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.

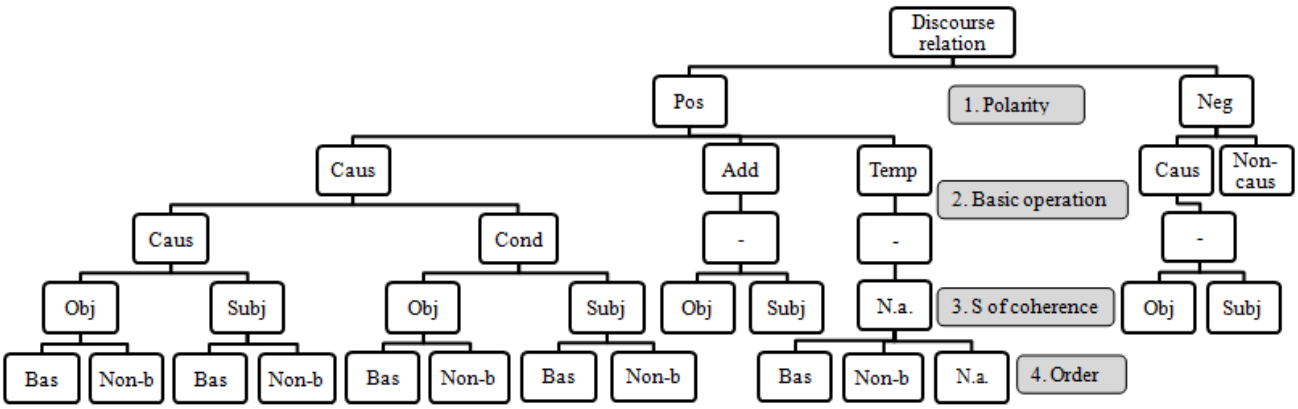

Figure 1. Flowchart of the step-wise annotation instruction.

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 (κ = .73). Agreement is moderate for the primitives basic operation (κ = .42) and order (κ = .47). Agreement on source of coherence is fair (κ = .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 < κ < .78), but almost perfect in the first year explicit condition (κ = .84). Agreement on basic operation is moderate in most conditions (.45 < κ < .50), but it is fair in the first year implicit condition (κ = .33). For source of coherence, agreement is fair in all conditions (.27 < κ < .39). Agreement for the primitive order is fair for the first year implicit condition (κ = .34) and substantial in the third year explicit condition (κ = .66), whereas it's moderate in the other conditions (κ = .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 (κ = .86). Agreement with order was substantial (κ = .61). Agreement of the annotators with the original annotations on basic operation was moderate (κ = .49) and agreement on source of coherence was fair (κ = .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 < κ < .52) and fair agreement for the primitive source of coherence (.25 < κ < .36). For polarity, most conditions show almost perfect agreement (.84 < κ < 91), except for the third year implicit

condition, which shows substantial agreement (κ = .79). For the primitive order, the two explicit conditions show substantial agreement (.62 < κ < .69), but only moderate agreement is found in the first year implicit condition (κ = .50) and third year implicit condition (κ = .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.