# Coding coherence relations: Reliability and validity

WILBERT SPOOREN and LIESBETH DEGAND

*Abstract*

*This paper tackles the issue of the validity and reliability of coding discourse phenomena in corpus-based analyses. On the basis of a sample analysis of coherence relation annotation that resulted in a poor kappa score, we describe the problem and put it into the context of recent literature from the field of computational linguistics on required intercoder agreement. We describe our view on the consequences of the current state of the art and suggest three routes to follow in the coding of coherence relations: double coding (including discussion of disagreements and explicitation of the coding decisions), single coding (including the risk of coder bias, and a lack of generalizability), and enriched kappa statistics (including observed and specific agreement, and a discussion of the (possible reasons for) disagreement). We end with a plea for complimentary techniques for testing the robustness of our data with the help of automatic (text mining) techniques.*

*Keywords:* *coherence relations, discourse, reliability, interrater agreement, corpus analysis*

## 1. Introduction

In recent years the analysis of discourse coherence has advanced considerably. A large number of studies has appeared that show that coherence is a crucial aspect of what makes a text a text (see Sanders and Spooren, 2007 for an overview). Many of those studies rely on corpus data for the analysis of the structure of discourse. Such studies typically hypothesize that different types of cohesive elements (typically coordinating or subordinating conjunctions and conjunctive adverbs) signal different types of coherence relations. An example is Spooren et al. (to appear), who investigate whether the use of Dutch *want* ('for, because') and *omdat* ('because') in spoken language follows the patterns

*Corpus Linguistics and Linguistic Theory* 6–2 (2010), 241–266 1613-7027/10/0006–0241
DOI 10.1515/CLLT.2010.009 

that have been described by Degand and Pander Maat (2003) for written language. To that end more than 250 fragments with *want* or *omdat* from the Corpus of Spoken Dutch have been analyzed for a number of variables, including the propositional attitude of the introductory segment (does it express a fact, an action, general knowledge, individual knowledge, an experience, a perception or a judgment), the type of relation between the two related segments (is it a non-volitional content relation, a volitional content relation, an epistemic relation, a speech act relation or a textual relation), presence of a conceptualizer in the introductory segment (no conceptualizer, a first person conceptualizer, a second person conceptualizer, or a third person conceptualizer) and linguistic expression of that conceptualizer (explicitly present or implicit).

This type of analysis requires a large number of coding decisions, which are usually based on subtle interpretive differences. Hence they raise the issue of intercoder agreement. In previous work we have argued that especially in the field of discourse analysis, with its complex semantic interpretations, intercoder agreement is an important issue (Spooren 2004). Low interrater agreements suggest that the categories used in a theory are vague, in the sense that categorizations are non-replicable, and consequently unfit as a basis for theory building.

Despite its importance, there is presently no tradition in the field of corpus-based discourse studies to report agreement measures. The reasons are probably twofold: on the one hand the community may not be aware of the importance of sufficient interrater agreement, on the other hand agreement may turn out too low to report. The reason for the latter might be that the coherence of a text is not in the verbal material, which gives at best instructions for and restrictions on the interpretation. Textual coherence resides in the mental representation that readers make of a text (Graesser et al. 1997; Sanders et al. 2002; Sanders and Spooren 2007). Consider the following fragment:

(1) (a) Greenpeace heeft in het Zuid-Duitse Beieren een nucleair transport verstoord.
(b) Demonstranten ketenden zich vast aan de rails.
(Telegraaf-i, April 10, 2001)
(a) Greenpeace has disturbed a nuclear transport in the Southern German state Bavaria.
(b) Protestors chained themselves to the tracks.

Among the many inferences we make on the basis of this short electronic news item is the fact that the impediment of the transport was *caused* by the protesters chaining themselves to the rails. This information is not present in the explicit linguistic material. We infer it on the basis of world knowledge and knowledge about the genre (writers of news texts are expected to explain the phenomena they describe).

As there is no explicit indication of the causal link between (a) and (b), we need to rely on interpretation to do the analysis. And interpretation is prone to individual variation, which may result in unreliable classifications. There is another reason why the problem is urgent in the case of coherence relations. It is a well known fact that reliability is more difficult to achieve when the number of categories increases. Mann & Thompson (1988) distinguish over 20 coherence relations. Therefore the classification of a text fragment as a particular instance of a certain coherence relation is a great source of variation (see Den Ouden et al. 1999 for an investigation of the reliability of this type of classification). Nevertheless, explicit accounts of interrater agreement are rare in the corpus linguistic community, even if there are a number of exceptions, especially in the area of (manual) corpus annotation projects (very often built in the context of computational linguistic applications), (e.g. Jovanovic et al. 2006, Pitt et al. 2005, Shriberg et al. 2004, Palmer et al. 2005). Still, several of such annotated corpora are distributed without mentioning interrater agreement accounts (cf. Penn Discourse Treebank, PDTB 2006; Carlson et al. 2003; the Lancaster Speech, Writing and Thought Presentation Project, Semino and Short 2004). When it comes to classifying discourse phenomena, only few researchers report issues related to intercoder agreement (for exceptions see, Steen et al. to appear, on metaphor identification, Miltsakaki et al. 2004, on annotating discourse connectives and their arguments; see also Poesio and Vieira 1998; Rosenberg and Binkowski 2004; Tou Ng et al. 1999; Marcu et al. 1999).

In this paper we want to explore the issue of interrater agreement in analyzing coherence relations. In the following section we will describe the size of the problem on the basis of a sample analysis. This sketch will be put into the context of recent literature from the field of computational linguistics on required intercoder agreement (Section 3). In Section 4 we will describe our view on how to deal with this issue. We will suggest three routes to follow in the analysis of coherence relations.

## 2. The size of the problem

In order to estimate the size of the problem we have taken a subset of the coding procedure of Spooren et al. (2010) and applied it in three different versions to limited sets of data. First we will present a pilot study, in which we calibrate our codebook by carrying out a preliminary analysis. Then we will present a test of our codebook in two phases: a test with relatively few fragments and a considerable number of corpus variables and a test in which we use a larger number of fragments but only a restricted number of variables. The results of these three steps will be described below.

## 2.1 *Pilot Analysis: Calibrating the codebook*

### 2.1.1 *Method*

From the 22 variables analyzed by Spooren et al. (2010) we have made a selection of 10 variables. In this first step we used the original coding instructions that were used by Spooren and colleagues. The materials to be analyzed were a random sample of 20 fragments from the Corpus of Spoken Dutch (Corpus Gesproken Nederlands, CGN; Oostdijk 2000). For present purposes we restricted our analyses to occurrences of *want* ('because', 'for'). We used the orthographic transcription of the corpus for our analyses. We used a Perl script to extract the utterances containing *want* from the corpus. The lines of the transcript are numbered. For each fragment we collected the utterance containing *want* and five lines of preceding and following context. In (2), an example is given of a fragment from CGN.

(2) Example fragment from CGN

```
fragment 1202 fn000640
101 N01113  of ze loopt naar het park bij het kasteel.
            'or she walks to the park near the castle.'
102 N01113  als het mooi weer is gaat ze daar even
            zitten.
            'if the weather is nice she sits down there
            for a while.'
103 N01114  ja.
            'yes.'
104 N01113  dus toch een enorm gevoel van vrijheid omdat
            ze niet of een taxi moet bellen of wat ik net
            noem de bus of of trein te nemen.
            'so still an enormous feeling of freedom
            because she doesn't have to call a cab or
            what I just said take the bus or the train.'
105 N01114  ja ja precies.
            'yes yes exactly.'
106 N01114  WANT ja taxi is dus ook een gedoe.
            'WANT yes taxi is also a hassle.'
107 N01113  WANT uhm bejaardenwoningen die liggen vaak
            aardig buiten het centrum.
            'WANT uhm homes for the elderly they often
            are located quite a long way from the
            centre.'
108 N01114  ja ja ja.
            'yes yes yes.'
```

```
109 N01113  en als ze in Nijmegen uh terecht was gekomen
            dan had ze vaak uh de bus moeten nemen en dan
            zit je twintig minuten in de bus voordat je
            in het centrum bent is dat mens doodmoe als
            ze weer thuiskomt.
            'and if she had ended up in Nijmegen uh then
            she would have uh had to take the bus often
            and in that case you sit in the bus for
            twenty minutes before you are in the centre
            that woman is exhausted when she gets home.'
110 N01114  ja.
            'yes.'
111 N01114  precies.
            'exactly.'
```

The fragment is numbered (1202) and gives the name of the file from which the fragment comes. The utterances are numbered and preceded by a code identifying the speaker (N01113, N01114). Occurrences of *want* have been capitalized for ease of identification. Note that the transcribers of CGN have added punctuation marks and that speaker overlap is not indicated in the transcription.

The variables selected for our analysis are given in Table 1 (possible values are given in parentheses).

Some of these variables focus on interpretation (Var5, Var6, Var9), some on more formal characteristics of the utterances (Var3, Var4, Var7, Var8, var10).

Table 1. *Variables in the first analysis*

| Variable | Content (levels) |
|---|---|
| Var 1 | Name of the coder (values: the names of the two authors) |
| Var 2 | Number of the fragment (the values were present in the fragments) |
| Var 3 | Utterance number(s) of the segment preceding *want* (S1) |
| Var 4 | Utterance number(s) of the segment following *want* (S2) |
| Var 5 | Propositional attitude of S1 (values: action, fact, opinion, observation, knowledge, experience) |
| Var 6 | Propositional attitude of S2 (values: action, fact, opinion, observation, knowledge, experience) |
| Var 7 | Identity of the conceptualizer in S1 (values: speaker/1st person, second person, third person (nominal or pronominal, generic person) |
| Var 8 | Identity of the conceptualizer in S2 (values: speaker/1st person, second person, third person (nominal or pronominal, generic person) |
| Var 9 | Type of relation expressed by *want* (values: non-volitional content, volitional content, explanation of a mental state, epistemic, textual, speech act) |
| Var 10 | Syntactic modification of *want* (values: no modification, coordinating conjunction, intensifier, focus element) |

These variables are presented here for demonstration purposes only; at the same time they represent our current state of thinking with respect to the differences between *want* and *omdat*. As stated in earlier analyses (Degand 1998, 2001, Degand and Pander Maat 2004, Pit 2003, 2007, Spooren et al. 2010), the two connectives seem to differ in degree of subjectivity. Such differences can linguistically be reflected in the type of propositional attitude a segment expresses, the type and presence of a conceptualizer and the type of coherence relations.

As to propositional attitudes expressed in a segment, opinions are expected to be more subjective than actions, which in turn are less objective than facts, observations, experiences and knowledge.

Segments can also be subjective because they reflect the presence of an active conceptualizer (whose mental activities are presented in the text), especially if that conceptualizer is the speaker or the addressee. Therefore, 1st and 2nd person conceptualizers are considered more subjective than 3rd person conceptualizers, but those are less objective than segments that do not reflect the presence of a conceptualizer (Langacker 1990).

The type of coherence relation that is expressed by the connective can also differ in subjectivity: non-volitional content relations are objective in that they reflect physical causality; volitional content relations (for example reasons for an action) are somewhat more subjective in that they reflect conceptual activity of the person for whom the reason holds; speech act relations (in which performance of a speech act is motivated) and epistemic relations (in which the speaker or addressee draws inferences on the basis of evidence) are considered most subjective (Pander Maat and Degand 2001).

The appendix reproduces the codebook that we used and contains a number of examples and operationalization to guide the analysts in their categorizations.

### 2.1.2 *Results and conclusion*

An informal analysis showed a considerable amount of disagreement. Therefore we did not perform any agreement statistics. Instead, we analyzed possible sources of disagreement and made some changes to the codebook and the procedure of coding:

a. The amount of context that was available for each fragment was limited. It was decided to double the amount of context available for the analysis (10 lines per target utterance). It would also have been helpful to be able to listen to the audio versions of the fragment, but for practical reasons this was not possible.
b. Some of the disagreement was caused by the fact that the fragments are sometimes underdetermined with respect to the values of the variables. For example, for 'Type of relation', in some cases, both a coding as expla-

nation for a mental state and as epistemic relation seems defendable. It was decided to allow for such ambiguous coding in the next analysis.

c. For some variables (notably 'Propositional attitude of the segment') the order of applying the coding seems important. In a fragment like (3) the utterance 'uurtje' ('an hour') in line 99 can be coded as a fact (paraphrase: "I have been there for an hour") or as a judgment (paraphrase "I must have been there for an hour"). Some of the disagreements occur because one of the coders followed the strategy that if the first value (fact) of propositional attitude was applicable, she chose that value.

(3) Example fragment from CGN

```
99 N01090   uurtje.
            'an hour.'
100 N01089  oh.
            'oh.'
101 N01090  ja WANT ik heb nog wat gedronken in CC.
            'yes WANT I still had drink in CC.'
```

d. Some of the disagreement was caused by the fact that some distinctions are fairly subtle. It was therefore decided to simplify some of the variables, by aggregating over some of the categories (for example, it was decided to disregard the distinction between an ordinary act and a speech act in the variable 'Propositional attitude').

e. A final decision was to align the strategies of analysis to the extent that coherence relations were assumed to occur within the utterance from the same speaker. Only if it turned out to be impossible to find a relation between utterances from the same speaker, the coherence relation was assumed to exist between utterances from different speakers.

### 2.2 *Test 1: Ten variables, calibrated codebook*

For the second analysis a second set of 20 fragments with *want* was randomly selected from the CGN corpus. The codebook was adapted along the lines discussed above. The two authors coded the fragments independently on the basis of the codebook.

#### 2.2.1 *Results*

Agreement statistics were calculated for variables Var3–Var9. Coding of Var1 and Var2 was obvious, and for Var10 there was complete agreement. For the utterance numbers (Var3, Var4) it was not possible to calculate kappa values, as the values of these variables vary per fragment.

– Utterance numbers (Var3, Var4). The authors agreed in 16 out of 20 cases or 80% on the utterance numbers of S1 and in 17 out of 20 cases or 85% on

the utterance numbers of S2. In other words, determining which specific segments are related by the connective is not 100% reliable.

– Propositional attitude (Var5, Var6). The authors agreed in 15 out of 19 cases[1] or 79.0% on the propositional attitude of S1 (Var5) (κ = .62). For S2 agreement was found in 10 out of 19 cases (52.6%; κ = .32).
– Identity of the conceptualizer (Var7, Var8). The authors agreed in 11 out of 19 cases (57.8%) on the identity of the conceptualizer in S1 (Var7) and in 9 out of 19 cases (47.4%) in S2. κ could not be computed because the coders used different categories.
– Type of coherence relation (Var9). The authors agreed in 10 out of 19 cases (52.6%, κ = .38) on the type of coherence relation that is expressed by *want*.

### 2.2.2 *Conclusion*

Despite the fact that we streamlined the coding procedure and that we had reduced the number of categories to be coded, the intercoder agreement remains exceptionally low. This is especially a problem because we feel that we are experienced coders who have worked with this type of analysis for many years.

## 2.3 *Test 2: Fewer variables, more fragments*

In a third step we decided to repeat the analysis for only two variables for a larger number of fragments. The two variables that were chosen were the ones most central to the analysis of coherence relations, namely the Propositional attitude of S1 (Var5) and the Type of Coherence relation (Var9). 50 Fragments were selected randomly from CGN, with 10 lines of context before and after the target utterance. In order to facilitate coding, the values of Type of Coherence relation were reduced to three: Content relation (volitional or non-volitional, including relations involving reasons for mental states), Epistemic relation and Meta relations (including speech act relations and textual relations), more or less along the lines of Sweetser's (1990) original proposal to categorize coherence relations in a content domain, epistemic domain, or speech-act domain. The aim of this analysis was to maximize the opportunity of reaching a high intercoder agreement.

The two authors coded the 50 fragments independently, using a drastically reduced version of the codebook. The results of the analysis are as follows:

– Propositional attitude of S1 (Var5). The authors agreed in 36 out of 48 cases (75.0%) (two fragments were uninterpretable) (κ = .58).
– Type of coherence relation (Var9, reduced). The authors agreed in 33 out of 45 cases (73.3%) (in five cases one or both coders found the relation impossible to label) (κ = .60).

We conclude that in our enterprise .60 is more or less the maximum kappa value that can be obtained. How can we explain such a low maximum agreement score?

A number of disagreements are due to real interpretation ambiguities. In some cases more than one interpretation seems available, as in fragment (4).

(4) fragment 481 (filename fn000313.sea)
*ik weet niet precies waar hoor WANT ik heb uh ook vandaag ook honderd keer zitten kijken waar ben ik?*
"I don't know exactly where it is mind you WANT I have uh today also already been looking a hundred times where am I?"

The first coder has assigned code 1 (opinion). The basis for that judgment was the presence of *hoor* ('mind you'), which indicates the speaker's point-of-view (possible paraphrase 'I'm warning you that I don't know the exact location'). The second coder focused on the explicit expression of not knowing (*weet niet*), which gets code 2 (fact, knowledge, observation, etc.). The two codings are both defendable.

In a substantial number of fragments there was disagreement between an epistemic relation (code 4, used by coder 1) and motivation for a mental state (code 3, used by coder 2). An example is (5).

(5) fragment 111 (filename fn000264.sea)
*ik denk 't niet WANT daar was dus een beetje bang voor dat zei 'k toch al xxx bang dat Nico zo ging uh ...*
'I don't think so WANT that was a little bit afraid of that's what I said xxx afraid that Nico would go uh ...'
*een beetje bang was dat wij gingen stressen.*
'was a little bit afraid that we would be stressing.'

Coder 1 took as the basis for code 4 (epistemic relation) the paraphrase 'I conclude here and know that we won't be able to empty the room already, WANT Nico was somewhat afraid that we would be stressing'). Coder 2's paraphrase was 'I explain to you why I think that we won't be able to empty the room already: Nico was somewhat afraid that we would be stressing'). One of the tests for choosing an epistemic relation is that, since epistemic relations hold in the here-and-now, they cannot be put in the past tense without significant change of interpretation. This test does not help us here. Under the interpretation provided by coder 2, the example can be put into the past tense 'My explanation why I thought that we would not be able to empty the room already was that Nico was somewhat afraid that we would be stressing'; in the other interpretation it cannot be put into the past tense.

The conclusion is that different interpretations are possible, thus leading to different coding. Since coding disagreement is in this case the result of genuine interpretation ambiguity, we believe that a "minimal" level of disagreement should be accepted as being inherent to language use without ill consequences on theory building.

A different type of disagreement is caused by what may be labeled as coding errors that can be resolved after discussion. An example is (6).

(6) Fragment 489 (filename fn000791.sea)

Speaker N01154
*hoe doe je dat nou als je met je mobiel in het buitenland belt?*
'How do you do that when you call on your mobile phone abroad?'
*als je bijvoorbeeld op vakantie bent.*
'For example if you are on holiday.'
*en je belt met dat ding naar huis?*
'and you use that thing to call home?'
*zitten dat ook al in die belminuten zeg maar.*
'is that also included in those calling minutes [= bundle] so to speak'
Speaker N01155
*ja volgens mij wel maar dat zal morgen ook nog wel dat moet 'k ook nog even navragen.*
'Yes I think so but that will tomorrow also I'll have to check that'
Speaker N01154
*moet je dan tot aan de grens moet je dan uh ...*
'do you until the border do you uh ...'
Speaker N01155
*ja WANT het is wel zo als je belt dan is voor ...*
'Yes WANT it is the case if you call then it is for ...'
*uh tot aan de grens betaal jij.*
'uh until the border you pay.'
*en de rest betaalt thuis.*
'and the rest is paid by home.'

In the analysis of this fragment the two coders disagreed on the propositional attitude of the second segment ('it is the case that if you call, up to the border it is charged to you, and the rest is charged to your home'). One analyst took this to be an instantiation of a general rule (hence code 2), where the other coder took it to be an action (hence code 3). After discussion the two coders

agreed that the general rule interpretation is much more likely. The latter type of disagreement is evitable. It results from a misinterpretation of the categorization variables, which means that these are either not well established enough and should be revised, or not applied correctly by the analyst.

### 2.4 *Preliminary conclusion*

Our experience in the three rounds of coding leads us to a number of conclusions. The first is that the quality of coding improves over time. We did not analyze the results of our first attempt formally because at face value it was clear that we disagreed considerably. Although we had used the codebook earlier, our experience dated from some time ago. We had to get acquainted again with the categories in the analysis. Complex coding like the one at hand apparently requires a warming-up phase. During this warming-up, it is possible that coders develop their own coding strategy, which may lead to disagreements. It is important that analysts acknowledge that such a warming-up phase might exist and that they take it into account in their coding procedure. For instance, one could imagine that the twenty or so first coded occurrences are left aside as training material for the coders involved in the analysis. In addition to a calibration phase of the codebook during which the variables to be coded can be completed or adapted to account for unforeseen phenomena, a "fine-tuning" phase of the application of the codebook would be needed to ensure a "correct" interpretation of the codebook between the different coders.

On the other hand, if quality of coding depends on a warming-up phase, this indicates that coding events are not independent. The dependence of coding raises questions about the statistics used to analyze intercoder agreement. These usually assume independence of coding events, as we will see in the next section.

Another important conclusion to be drawn from our experience is that coding disagreement can be of two fundamentally different types: (i) coding ambiguity, or (ii) coding error. As mentioned before, the first type of disagreement is inevitable because it results from inherent language ambiguity and utterance interpretation problems. If we want to continue analyzing and coding semantic phenomena, a certain margin of disagreement should be allowed for placing a perfect agreement (e.g. a kappa of 1.0) out of reach. The second type of disagreement is the one that is supposed to tell us something about the stability of our coding scheme and the theoretical conclusions that can be drawn from our analysis. It is these disagreements that the agreement statistics are meant to track. Section 4 pursues the discussion of these two types of intercoder disagreement.

## 3. Intercoder agreement in the linguistic literature

Intercoder agreement has been an issue in a wide range of communities (medicine, (educational) psychology, social sciences, etc.). Probably the first to propose a formal measurement of agreement was Jacob Cohen (Cohen 1960). Within linguistics intercoder agreement has most extensively been discussed in the computational linguistic community. The discussion was revitalized with an influential squib by Jean Carletta (1996) in *Computational Linguistics*. Carletta made a plea not to report only percentages of agreement since these do not correct for chance agreements. She recommends reporting kappa scores, which include correction for chance. Perfect agreement is indicated by a kappa of one, and pure chance is indicated by a kappa of zero. Negative kappa indicates disagreement greater than that expected by chance.

Recently, Artstein and Poesio (2008) have dealt with the topic in extenso. They give an overview of the discussion concerning the use of agreement measures since the publication of Carletta's paper. They also give terminological clarifications and introduce the strengths and weaknesses of a large number of agreement measures. All agreement measures in one way or the other are a ratio of the difference between observed agreement and expected agreement ($A_{observed} - A_{expected}$) and the difference between 1 and the expected agreement ($1 - A_{expected}$). The measures differ in whether or not they account for individual differences between coders and whether or not they are able to deal with differences in weight of disagreements (not all disagreements are of equal weight; a measure like Krippendorff's $\alpha$ can deal with such differences in disagreements).[2]

Artstein and Poesio also discuss an important issue frequently observed in the coding of coherence relations, namely category prevalence. Category prevalence occurs when a disproportionate amount of data falls into one category. In that case it can happen that observed agreement is high (e.g., higher than .90), but that the agreement measure is strikingly low. Some have concluded that in case of prevalence agreement measures are misleading, but Artstein and Poesio reject that view, claiming that in case of prevalence, agreement means agreement about the rare categories. It is therefore correct, they claim, that disagreements on rare categories have a strong influence on the agreement measure (in the domain of clinical research, a similar conclusion is reached by Vach 2005).

Artstein and Poesio end their discussion with a recommendation of agreement measures of 0.80 and higher as indications of coding quality. They hasten to add that such a threshold value of agreement cannot be used across the board, and that "useful corpora have been obtained while attaining reliability only at the 0.7 level" (Artstein and Poesio 2008: 37).[3]

Does this mean that we should not use categories which cannot be coded at approximately this level of agreement? No, we don't believe so. What a low

intercoder agreement score says is that the categories cannot be used in a consistent manner, and/or that the coded phenomenon is intrinsically ambiguous (cf. supra). At the same time, these categories may still be interesting. As Craggs and McGee Wood (2005: 293) put it:

> the subjectivity of the phenomena being coded may mean that we never obtain the necessary agreement levels. [ . . . ] However, the fact that we consider these subjective phenomena worthy of study shows that we are, in fact "willing to rely on imperfect data", which is fine as long as we recognize the limitations of a scheme which delivers less than ideal levels of reliability, and use the resulting annotated corpora accordingly.

## 4. Low kappas as a fact of life

Ideally coders work completely independently and agree substantially. What if this ideal cannot be reached, for reasons mentioned earlier? A suboptimal solution is to make use of double coding (i.e. use two coders for all of the data, preferably randomly selected from the corpus) and discuss disagreements. The disadvantage compared to the single coder strategy is mainly pragmatic: it will more than double the amount of time needed for the coding phase. The advantages are many: The amount of coding errors is reduced and any coding strategy that is developed will be a cooperative strategy. A double coding strategy requires that you convince your research associate of the quality of your coding by making explicit the reasoning on which the coding is based. This will undoubtedly increase the quality of coding. Time-consuming as this strategy may be, we have used it in many instances in the past, and will continue to use it in future. For example, Sanders and Spooren (2009) have used complete double coding for their analysis of *want* and *omdat* in three different corpora (written language, spoken language, chat). This was also the case for Degand and Pander Maat's (2003) analysis of the same connectives in written press articles.

A variant of the double coding is the partial overlap coding between two or more coders. The sample to be coded is cut up in several subparts of which some are double coded, while others are coded by only one analyst. For instance, on a sample of 500 occurrences, twice 200 occurrences would be single coded by the two analysts, and 100 would be double coded. This reduces the work load of both analysts two 300 occurrences each instead of 500. The sample of double coded occurrences can serve to calculate agreement statistics, to discuss disagreements as mentioned in the beginning of this section, and as such enhance the reliability of the single coded samples. The disadvantage remains that there is no guarantee that the single coded samples will be analyzed strictly along the same lines as the double coded ones, and that the results reflect a kind of "mean" interpretation between different coders rather than a

constant one. This was the procedure used by Spooren et al. (2010) and Degand and Fagard (2008).

An alternative solution is the one-coder-does-all solution: The complete corpus sample is coded by one and the same coder (again, preferably with a random selection of the sample). Of course the coding will be subject to individual strategies developed by the coder, but these strategies will presumably be systematic and there is no reason to assume that such strategies will be conflated with the phenomena of interest. For example, if a coder tends to overcode for judgments, then it can be assumed that judgments will be overrepresented in the analysis across the board. So if our research question is whether judgments occur more often with *want* than with *omdat*, an overcoding of judgments will not impede an answer to the research question.

Reaching agreement through discussion and the single coding strategy are neither in accordance with the usual standards of assessing coding quality through interrater agreement. But let us consider those standards in more detail. What does it mean for instance if two coders, coding completely independently, meet the standard set by Artstein and Poesio of an interrater agreement statistic of .80? Coding is a process of applying interpretations fixed in the coding instructions to different linguistic forms. A high interrater agreement signals that the interpretations are indeed sufficiently explicitly captured in the coding instructions. If only moderate agreement scores can be reached (say, a maximum of .60, as in our case), this could signal that the interpretation process itself is a source of disagreement. In other words, the coding instructions do not fix the interpretation process.

This leaves us with two possibilities: One is that theories of coherence relations are built on quicksand, and therefore fundamentally flawed, because it seems impossible to get a grip on the basic data. However, we believe that there is ample reason not to be so pessimistic. For even if it is impossible to reach high levels of agreement in the analysis, the results of the various analyses that have been carried out (independently) in the past converge on conclusions like: different languages make use of the same conceptual coherence relations, although the division of labor that various connectives perform varies from language to language (Pit 2003; Degand 2004); connectives can be put on a scale from more to less subjective, with (in Dutch) *doordat* as most objective and *dus* as most subjective (Pander Maat and Degand 2001; Pit et al. 1997).

This leaves us with the other possibility. We think that it is not a coincidence that high agreement scores are rare in the case of the analysis of coherence relations. We think that it reflects the fact that language is to a high degree underdetermined with respect to the resulting interpretations (cf. infra). A coherence relation like Cause-Consequence can be marked explicitly (using a connective like *because*), or it can remain implicit (no connective), in which case the coherence relation has to be inferred; but it can also be expressed with

a so-called underspecific connective (Spooren 1997), for example with a temporal connective like *when* (as in *When John came in, Bill left*). This implies that establishing the coherence relation in a particular instance requires the use of contextual information, which in itself can be interpreted in multiple ways and hence is a source of disagreement.

The analyst who wants to interpret coherence relations in a text should make use of all of the contextual information available. In the case of writing, the contextual information is usually encoded into language: writers expect their readers to infer their communicative intentions on the basis of the text only. In case of spontaneous speech, this is radically different: much of the communicative content is available in the paralinguistic and extralinguistic context. Ideally then, the analyst should have witnessed the original conversation; in corpora of spoken language the sound recording is nowadays usually provided (sometimes even video recordings) as an approximation. In our case, however, no sound files were available, and we were only able to make use of the transcripts of the conversations, which may have added to the amount of disagreement.

### 4.1 *The redundancy of language*

Perhaps the low kappa scores obtained in our types of analysis should not come as a surprise, because language is fundamentally both redundant and economical.[4] It is part of the essence of language that much of the information conveyed between communicative partners is redundant in that it is presented in many different ways (gestures, contextual information, shared background, the linguistic code; see Clark 1996 for a description of these different modes to perform 'joint communicative actions'). Even in writing, e.g. in printed mass media "[m]eaning is conveyed not only by the words in a news item but also by the size of the headline, the position on the page and the page in the paper, the association with pictures, the use of boldface and other typographical devices" (Schramm 1997: 57–58). In this sense, language is redundant: As a rule, people succeed in understanding each other in communication thanks to a variety of indices. Speakers can also "deliberately introduce more redundancy; we can repeat (. . .), or we can give examples and analogies" (Schramm 1997: 54).

At the same time language is semantically underdetermined (see Sperber and Wilson 1995 [1986]; Recanati 2002a, 2002b). What is meant by this is that "in order to ascribe a definite meaning to a sentence, it is necessary (in many cases) to take contextual factors into account." (Vicente and Martinez-Manrique 2005: 537). In other words, "one can determine the content of the speech act only by appealing to pragmatic considerations concerning what the speaker means, what his intentions are" (Recanati 2002b: 116), and this is a

matter of interpretation. As a consequence, if you ask someone to specify the exact meaning of their contributions, this is usually impossible. We relate this to the observation made by Burge that "[w]e seem normally to understand content in a way whose [sic!] unconscious details (. . .) are not accessible via ordinary reflection. To be entitled to believe what one is told, one need not understand or be able to justify any transition from perceptual beliefs about words to understanding of and belief in the words' content." (Burge 1993: 477, cited by Recanati 2002b: 115). Thus, it is not a contradiction to entertain simultaneously the thought that we have understood the fragments, and are not capable of making this understanding explicit by putting it into a category.

What does this imply for any standard criterion for acceptable agreement statistics? Such a question is difficult to answer. What a low agreement score indicates is that there is something awkward that has to be accounted for. Lack of agreement should lead to a suspicious attitude of the researchers with respect to their interpretation of the data and/or their method of analysis. In this sense, we follow Artstein and Poesio that high agreement is better than low agreement. Our research gains from robust data and low agreement is an indication of poor robustness. Therefore, it seems wise not to accept scores that are too low. As a rule of thumb we are tempted to suggest that agreement statistics should minimally reach the level of .70 for coherence phenomena. This is lower than Artstein and Poesio's standard, but higher than "more forgiving scales", to use Di Eugenio's (2000) words. Di Eugenio cites Rietveld and van Hout (1993) who "consider $.41 \leq K \leq .60$ as indicating moderate agreement, and $.61 \leq K \leq .80$ as indicating substantial agreement. [While t]he psychiatric community considers $K \geq .6$ or even $K \geq .5$ as acceptable (Grove et al. 1981)." (Di Eugenio 2000: 1). Referring to Krippendorff (1980: Ch. 12) she adds that the significance of any such standard cannot be absolutely stated, but depends on the usage of the results that one derives from the analysis, and in particular, at the cost of wrong conclusions. Hripcsak and Heitjan (2002) go even further when they argue that "intermediate levels of kappa between zero and one cannot be interpreted consistently (. . .) [because] the interpretation of these levels relies heavily on the tasks and categories, the purpose of the measurement, and the definition of chance, so such guidelines are deceptive and should probably not be used." (Hripcsak and Heitjan 2002: 101).

The reasons for nevertheless choosing a standard of .70 are twofold: On the one hand the task of coding coherence relations is fundamentally determined by its reliance on contextual interpretation. Therefore, we believe a standard of .80 is unrealistic because it is too high. At the same time we feel the need for setting a lower border for the interval below which researchers should pay attention to reliability issues. If agreement is lower than this lower border, we expect researchers to give an account of why there is so little agreement in their analysis of the data. But apart from setting a new standard, it might be

worthwhile to consider alternative ways to account for the robustness of our data.

### 4.2 *An alternative to kappa*

Problems raised by the use of kappa statistics are diverse. Among them are the so-called kappa "paradoxes". These paradoxes show "that the value of a Kappa coefficient is biased downwards due to violations of the Kappa assumptions" (Jung 2003: 478). An assumption commonly violated is that ''the distribution of random ratings is the same as that of systematic ratings, and no symmetric disagreement is involved'' (Guggenmoos-Holzmann and Vonk 1998 cited by Jung 2003, 495). For example, if the observed prevalence of responses in one of two available categories is low, then there is insufficient information in the sample to judge raters' ability to discriminate cases, and kappa may underestimate the true agreement (cf. category prevalence mentioned above).[5] Likewise, Di Eugenio (2000), reporting on the annotation of so-called Forward- and Backward-Looking Functions (approximately illocutionary acts) in a corpus of dialogues, warns that a factor that affects Kappa's computation is the skewed distribution of categories. Thus, she observes that the infrequent use of a given tag results in a poor kappa score because "a very high level of agreement on the tags that do occur is necessary to reach good results." She furthermore adds that "[t]his intuitive explanation is backed up by (Grove et al. 1981), which points out that the low frequency of a tag may lower the maximum K (corresponding to perfect agreement) to a value sometimes much lower than 1." (Di Eugenio 2000: 2).

This is a problem we encountered in our analyses, for instance when one of the coders hardly ever chose the coherence relation "explanation of a mental state", when the other did. In such cases of skewed distribution, Jung, working within the domain of software process assessment, recommends to complement the kappa coefficient with an index of observed agreement, i.e. the proportion of ratings upon which the two raters agree. According to the author, the "index of observed agreement is simple and easy to understand, and it avoids the difficulties of the Kappa paradox in the presence of skewed distributions of agreement and disagreement. However, the value of the index of observed agreement can be deceptively high since it does not correct for chance factors." (Jung 2003: 494).

Hripcsak and Heitjan (2002), within the domain of medical informatics, formulate the same recommendation to report observed agreement, minimally "as an initial descriptive statistic to summarize the sample" (Hripcsak and Heitjan 2002: 108). In the case of nominal categories, commonly used in linguistic categorization, kappa should be used only if the sample is relatively balanced

or if one's goal is to reject the null-hypothesis that agreement doesn't differ from chance (for more than two categories). In all other cases, "observed agreement and specific agreement on each category should be reported as descriptive agreement measures (but not as formal reliability measures)" (Hripcsak and Heitjan 2002: 108). The authors insist that the best approach depends on the goal of the research (see also Banerjee et al. 1999). At the same time, Di Eugenio (2000: 1) warns that "the dialogue and discourse processing community should pay more attention to the **meaning** of the scales used to evaluate Kappa values" (our emphasis).

We would like to conclude from this that the Kappa measure can also be applied in a "soft" way, together with observed and specific agreement, and a discussion of the (possible reasons for) disagreement. But next to agreement measures, we believe there is also room for complementary ways of testing the robustness of our data.

### 4.3 *A plea for complementary techniques*

An interesting possibility provided by the computational power of state-of-the-art computers is that of complementing hand-crafted, and thus possibly less reliable, analyses with analyses based on text mining techniques. Bestgen et al. (2006) provide an example of such additional analyses. They use two types of techniques, Latent Semantic Analysis and Thematic Text Analysis. The first type of analysis is used to determine the semantic relatedness between words, sentences and texts. This is done on the basis of a so-called term by document matrix, which for every term in a corpus gives the frequency of that term in each document. That matrix is converted into a many-dimensional semantic space. This makes it possible to establish semantic relationships between elements in the space. The measure for semantic relatedness between two terms, or between a term and a document is usually calculated as the cosine of the vectors representing the terms or documents. To reduce the influence of low-frequency elements only the first 300 or so dimensions of the semantic space are maintained.

The Thematic Text Analysis aims to establish whether words from a particular semantic or grammatical category are over- or underrepresented in particular text segments. For example, on the basis of a list of opinion words it is possible to test whether or not these opinion words occur more often in first segments of *want*-fragments than in first segments of *omdat*-fragments.

Presently the automatic analyses are not capable of replacing hand-crafted analyses. But they can function as complements to such analyses. The main advantage of these types of analysis is that they are completely automatic. This makes them immune to issues of intercoder reliability. Moreover it is possible to apply the analyses to large amount of data. Results obtained from hand-

crafted analyses of relatively small amounts of data can be corroborated using computer-based, large-scale automatic analyses. For example, Bestgen et al. (2006) have used LSA to test the hypothesis that the semantic relationship between segment 1 and segment 2 in *want*-fragments is less strong than that in *omdat*-fragments (based on the premise that the first segment in a *want* fragment is generally more perspectivized than the first segment in an *omdat* fragment). In addition they have shown that opinion words occur more often in the first segment of *want* fragments than in the first segment of *omdat* fragments. This suggests that *want* creates a more subjective environment than *omdat*.

## 5. Conclusion

We have discussed a fundamental problem with the analysis of coherence relations: it seems impossible to reach the high agreement statistics that the standard literature on this topic requires. Our guess is that this problem is not restricted to the analysis of coherence relations, but that it holds for all those cases where interpretation (as opposed to formal characteristics) of the phenomenon under scrutiny is central. We have suggested that low agreement scores are inevitable. At the same time we have suggested a number of ways to deal with this problem. A first strategy is that of two-coders-discuss: two coders analyze the corpus independently of each other, and afterwards the two coders discuss the differences; this will force the two coders to convince the other of the correctness of their analysis. A second strategy is to use a one-coder-does-all technique: in order to increase consistency of coding, one coder is responsible for coding the entire corpus. A third option is the soft kappa: two (or more) coders analyze the corpus independently of each other; complimentary agreement scores are reported: observed agreement, kappa coefficient, specific category agreement (to account for possible prevalence), and a discussion of disagreements. Finally we have suggested that the use of automatic techniques may complement the quality of our analyses.

We end with a plea: let us be explicit. Let us report our agreement scores, including kappas, let us give explanations for low kappas. And most of all: let us be explicit about the generalizability of our research results.

## Appendix 1

CODING INSTRUCTIONS

**var1 name of the coder** [self-evident]

**var2 number of the fragment** [to be found in the fragment]

**var3 utterance number S1**
the number before the line of S1
NB: in conversations the relation is preferably to be located between utterances from the same speaker. Only if that is impossible/unlikely, the relation is taken to exist between utterances from different speakers.

**var4 utterance number S2**
idem for S2

**var5 propositional attitude of S1**
*1 = judgment*
The segment mentions a subject of consciousness and that which is judged. The segment describes a situation that is not located on a particular time and/or place (it is stative), and contains a scalar predicate (which can be recognized it that it can be strengthened: *very much X*; *what is more, more than X* ).
Paraphrase: "I think that . . ."
*That is a pity* (what is more, dramatic)
*The stalk tastes well too* (what is more, it is delicious)
*I don't want to do that* (what is more, I refuse to do it)

*judgments w.r.t. the desirability or probability of an act are also categorized as judgments*
The segment mentions a subject of consciousness and that what is judged. The segment mentions a deliberate, voltional act. The judgmental character is usually located in such explicit indicators like modal verbs (*want, should* ) and adverbs (*apparently, probably, necessarily*).
*We have to leave, because . . .*
*He wanted to leave early, because . . .*

*2 = fact, individual knowledge, general knowledge, observation or experience fact*
The segment describes a situation or event that is to be located on a particular time (one can attribute a truth value to it; the segment could also have occurred with a past tense). The segment does not contain a subject of consciousness, there is no deliberate (volitional) protagonist in the causal event; there is only a writer/speaker who reports the event.
*The river flooded (because it has been raining for three days)*
*The vase fell. It broke.*
But also:
*He got on the wrong train (, because he hadn't heard the platform change)* (to get on the wrong train is not a deliberate, volitional act).
paraphrase: "it is a fact that . . ."

OR *individual knowledge*
The segment contains a subject of consciousness. It also describes an 'act' of understanding (or lack of understanding). This act can be located on a point of time (one can attribute a truth value to it; the segment could also have occurred with a past tense). The knowledge must occur explicitly.
*Charles knew that that it was of no use.*

OR *general knowledge*
The segment describes a general rule, which is a generalization over times and groups. The situation described cannot be located on a particular point in time, but it is also non-scalar.
*Man is a social animal (, because we cannot live without each other)*
*That wasn't a habit at Woolworth's in those days.*
Paraphrase: "Normally it is/was the case that . . ."

OR *observation*
The segment contains a subject of consciousness (SoC). The SoC is non-agentive (is an experiencer of the situation/event). The observation can be located on a point in time (one can attribute a truth value to it; the segment could also have occurred with a past tense). The segment contains an explicit verb of observation (*see, hear, taste,* etc.)
*John noticed the road signs*

OR *Experience*
The segment contains a SoC, which is non-agentive. The experience can be located on a particular point in time (one can attribute a truth value to it; the segment could also have occurred with a past tense). The experience is an individual event that is true at a particular point in time.
*Arie became ill.*
*They lost their share.*
Paraphrase: "I still remember that . . .", "I found out that . . ."
Differs from a fact in the presence of the SoC, which implies a certain extent of interpretation.

*3 = act*
The segment contains an agentive protagonist, who intentionally carries out a certain act. The act concerns an individual event, that can be located at a particular point in time or the act concerns processes. The acting can be made explicit by strengthening it adding a formula like ". . . and he did that immediately, fantastically well, with much pleasure . . ." etc.
*She went home early* (as she always did)
*I went to the pub* (and I did it on foot)

NB. speech acts are also categorized as acts.

**var6 propositional attitude of S2**
See var5.

**var7 identity of the protagonist/SoC in S1**
The protagonist is the main character in the act. In case of a judgment, experience, observation, the protagonist is the SoC. NB. If the protagonist remains implicit, choose the protagonist that is mentioned closest in the context.

1 = speaker/author
2 = 2nd person
3 = 3rd person
4 = generic 3rd person (*one* in *one does one's best*)
5 = irrelevant (in case of a fact)

*var8 identity of the protagonist/SoC in S2*
see var7

*var9 Type of relation expressed by want*
1 non-volitional
2 volitional: reason for an act; S1 expresses an act (incl. speech acts)
*John chose strawberry ice cream, because he liked that last time*
*I have told you this for the third time because you keep forgetting it.*
3 reason for a mental state/judgment
*I like it, because it gives you a link with reality*
NB. this is different from epistemic relations because this reason relation can be reported in the past tense, whereas epistemic relations are linked to the speaker/author's here-and-now.
4 epistemic: paraphrase "I conclude here and now that S1, on the basis of S2"
*he is ill, because his coat is not here.*
*he is very much suited as a politician, because his father was very eloquent.*
5 speech-act
S1 expresses a speech act that is supported by the information in S2
*What are you doing tonight, because I want to go to the cinema.*

*var10 syntactic modification of connective (want)*
1 no modification
2 connective (*en, dus, maar*)
3 adverbial ( *juist, vooral* )
4 focalizing construction (*het is* connective . . .)
5 interjection ( *ja, nou, zo* . . .)
6 connective + focalizing construction (*maar het komt* connective . . .)
7 salutation (by the use of a name)

**Acknowledgements**

The second author is senior research associate from the Belgian Fund for Scientific Research (FRS-FNRS). This research was supported by grant n° ARC 03/08-301 from the *Communauté française de Belgique* and by grant IUAP P6/44 from the Interuniversity Attraction Pole program from the Belgian Federal Government. We would like to thank two anonymous reviewers from *Corpus Linguistics and Linguistic Theory*, as well as Fleur van der Houwen (VU Amsterdam) for constructive comments on an earlier version of this paper.

**Bionotes**

Wilbert Spooren (*1956, Ph.D. Nijmegen University, 1989) is professor of Language and Communication at the Faculty of Arts of VU University Amsterdam, The Netherlands. His group teaches language and communication at several BA- and MA-levels. His research focuses on the relation between genre, discourse structure and text quality. E-mail: w.spooren@let.vu.nl

Liesbeth Degand (*1967, Ph.D. Université Catholique de Louvain, 1997) is professor in Linguistics at the University of Louvain (Louvain-la-Neuve, Belgium) and senior research associate at the Belgian Fund for Scientific Research (FRS-FNRS). Her main research interests go to the (corpus-based) study of discourse structure, especially (causal, temporal, and contrastive) discourse markers in Dutch and French, both in speech and in writing, from a synchronic as well as a diachronic perspective, and in native as well as learner language. E-mail: liesbeth.degand@uclouvain.be

**Notes**

1. In one case it was impossible to code the propositional attitude of the fragment, because the interpretation of the fragment was unclear.
2. Outside the linguistic domain, inter-coder agreement and reliability is a much debated issue. This is especially the case in biomedical, biometrical and clinical psychological research. The kappa statistics, although widely used and accepted, is regularly put to question (see e.g. Lawlis and Lu 1972; Lehmann et al. 1995; Guggenmoos-Holzmann 1996; Hux et al. 1997; Hripcsak and Heitjan 2002; Jung 2003; Kottner 2008.
3. Note that this is in line with Krippendorff's (1980) scale which "discounts any variable with $K < .67$, allows tentative conclusions when $.67 < K < .8$ K, and definite conclusions when $K \geq .8$." (Di Eugenio 2000: 1).
4. On the tension between the principle of economy and the principle of redundancy in language, see e.g. Horn (1993).
5. But see Vach (2005) for counterarguments.

## References


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

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney & Debajyoti Sinha. 1999. Beyond Kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics | La Revue Canadienne de Statistique* 27(1). 3–23.

Bestgen, Yves, Liesbeth Degand & Wilbert Spooren. 2006. Toward automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes* 41(2). 175–194.

Burge, Tyler. 1993. Content preservation. *Philosophical Review* 102. 457–488.

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

Carlson, Lynn, Daniel Marcu & Mary E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt & Ronnie W. Smith (eds.), *Current directions in discourse and dialogue*, 85–112. Kluwer Academic Publishers.

Clark, Herbert H. 1996. *Using Language*. Cambridge etc.: Cambridge University Press.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20. 37–46.

Craggs, Richard, & Mary McGee Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics* 31(3). 289–295.

Degand, Liesbeth. 1998. Het ideationele gebruik van 'want' en 'omdat': een geval van vrije variatie? [The ideational use of want and omdat: a case of free variation?] *Nederlandse Taalkunde* 4. 309–326.

Degand, Liesbeth. 2001. *Form and function of causation. A theoretical and empirical investigation of causal constructions in Dutch*. Leuven: Peeters.

Degand, Liesbeth. 2004. Contrastive analyses, translation and Speaker Involvement: the case of *puisque* and *aangezien*. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture and mind*, 251–270. CSLI Publications.

Degand, Liesbeth & Benjamin Fagard. 2008. Intersubjectification des connecteurs. Le cas de car et parce que. *Revista de Estudos Linguísticos da Universidade do Porto* 3(1). 119–136.

Degand, Liesbeth & Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In Arie Verhagen & Jeroen van de Weijer (eds.), *Usage based approaches to Dutch*, 175–199. Utrecht: LOT.

Di Eugenio, Barbara. 2000. On the usage of Kappa to evaluate agreement on coding tasks. *LREC2000, the Second International Conference on Language Resources and Evaluation*, Athens, Greece http://www.cs.uic.edu/~bdieugen/PS-papers/lrec00.pdf (accessed 13 November 2009).

Graesser, Arthur C, Keith K. Millis & Rolf A. Zwaan. 1997. Discourse comprehension. *Annual Review of Psychology* 48. 163–189.

Guggenmoos-Holzmann, Irene. 1996. The meaning of kappa: Probabilistic concepts of reliability and validity revisited. *Journal of Clinical Epidemiology* 49(7). 775–782.

Guggenmoos-Holzmann, Irene & Richard Vonk. 1998. Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine* 17. 797–812.

Horn, Laurence. 1993. Economy and redundancy in a dualistic model of natural language. *Yearbook of the Linguistic Association of Finland*. 33–72.

Hripcsak, George & Daniel F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics* 35(2). 99–110.

Hux, Karen, Dixie Sanger, Robert Reid & Amy Maschka. 1997. Discourse analysis procedures: Reliability issues. *Journal of Communication Disorders* 30(2). 133–150.

Jovanovic, Natasa, Rieks op den Akker & Anton Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation* 40(1). 5–23.

Jung, Ho-Won. 2003. Evaluating interrater agreement in SPICE-based assessments. *Computer Standards & Interfaces* 25(5). 477–499.

Kottner, Jan. 2008. Interrater reliability and the kappa statistic: A comment on Morris et al. (2008). *International Journal of Nursing Studies*. 46. 141–142

Landauer, Thomas K., Peter W. Foltz & Darrell Laham, D. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes* 25. 259–284.

Langacker, Ronald W. 1990. Subjectification. *Cognitive Linguistics* 1(1). 5–38.

Lawlis, G. Frank & Elba Lu 1972. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin* 78(1). 17–20.

Lehmann, Michel, Jean-Pierre Daurès, Nicolas Mottet & Henri Navratil 1995. Comparison between exact and parametric distributions of multiple inter-raters agreement coefficient. *Computer Methods and Programs in Biomedicine* 47(2). 113–121.

Mann, William C. & Sandy A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8. 243–281.

Marcu, Daniel, Estibaliz Amorrortu, & Magdalena Romera 1999. Experiments in constructing a corpus of discourse trees. In *Towards Standards and Tools for Discourse Tagging. Proceedings of the ACL'99 Workshop*, 48–57. Maryland, USA: Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W99/W99-0307.pdf. (accessed 13 November 2009).

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi & Bonnie Webber 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*. Boston, MA. 2004, 1–8. Retrieved from: http://www.seas.upenn.edu/~pdtb/papers/frontiers04.pdf (accessed 12 January 2010).

Oostdijk, Nelleke. 2000. The Spoken Dutch Corpus Project. *The ELRA Newsletter* 5(2). 4–8.

den Ouden, Hanny, Carel van Wijk, Jacques M. B. Terken & Leo G. M. Noordman. 1999. *Reliabilty of discourse structure annotations.* IPO Annual Report (Ext. r. no. 33). Eindhoven: IPO.

Palmer, Martha, Paul Kingsbury & Daniel Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71–106. doi:10.1162/0891201053630264.

Pander Maat, Henk & Liesbeth Degand. 2001. Scaling causal relations and connectives in terms of Speaker Involvement, *Cognitive Linguistics* 12(3). 211–245.

PDTB. 2006. *The Penn Discourse TreeBank 1.0. Annotation Manual*. IRCS Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania. March 2006. http://www.seas.upenn.edu/~pdtb/ (accessed 1 September 2009).

Pit, Mirna. 2003. *How to express yourself with a causal connective. Subjectivity and causal connectives in Dutch, German and French*. Amsterdam: Rodopi.

Pit, Mirna. 2007. Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast* 7, 53–82. doi:10.1075/lic.7.1.04pit.

Pit, Mirna, Henk Pander Maat & Ted Sanders. 1997. 'Doordat', 'omdat' en 'want'. Perspectieven op hun gebruik ['As a result of', 'because' and 'for'. Perspectives on their use]. *Taalbeheersing* [Language Use] 3. 238–251.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89–95. doi:10.1016/j.specom.2004.09.001.

Poesio, Massimo & Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics* 24(2). 183–216.

Recanati, François. 2002a. Unarticulated constituents. *Linguistics and Philosophy* 25. 299–345.

Recanati, François. 2002b. Does linguistic communication rest on inference? *Mind and Language* 17(1–2). 105–126.

Rosenberg, Andrew & Ed Binkowski. 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Companion Volume: Short Papers, Student Research Workshop, Demonstrations, Tutorials Abstracts*, 4 pp. Boston, USA. http://acl.ldc.upenn.edu/N/N04/N04-4020.pdf. (accessed 13 November 2009).

Sanders, Ted, Joost Schilperoord & Wilbert Spooren. 2002. *Text representation: Linguistic and psycholinguistic aspects*. Amsterdam: John Benjamins.

Sanders, Ted & Wilbert Spooren. 2007. Discourse and text structure. In Dirk Geeraerts & Hubert Cuykens (eds.), *Handbook of Cognitive Linguistics*, 916–943. Oxford: Oxford University Press.

Sanders, Ted & Wilbert Spooren. 2009. Causal categories in discourse – Converging evidence from language use. In Ted Sanders & Eve Sweetser (eds.), *Linguistic categories of causality in discourse*, 205–246. Berlin: Mouton de Gruyter.

Schramm, Wilburt. 1997. How communication works. In Alan Wells & Ernest A. Hakanen (eds.), *Mass media & society*, 51–66. Greenwich, CT: Ablex.

Semino, Eleno & Mick Short. 2004. *Corpus stylistics: The presentation of speech, writing and thought in a corpus of English writing*. London: Routledge.

Sperber, Dan & Deirdre Wilson. 1995 [1986]. *Relevance: Communication and cognition*. Oxford: Basil Blackwell.

Spooren, Wilbert. 1997. The processing of underspecified coherence relations. *Discourse Processes* 24(1). 149–168.

Spooren, Wilbert. 2004. On the use of discourse data in language use research. In Henk Aertsen, Michael Hannay & Rod Lyall (eds.), *Words in their places: A festschrift for J. Lachlan Mackenzie*, 381–393. Amsterdam: Faculty of Arts, VU. http://www.let.vu.nl/staf/w.spooren (accessed 26 October 2009).

Spooren, Wilbert, Mike Huiskes, Ted Sanders & Liesbeth Degand. 2010. Subjectivity and causality: A corpus study of spoken language. In John Newman & Sally Rice (eds.), *Empirical and experimental methods in cognitive/functional research*. Chicago: University of Chicago Press. 241–255.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang & Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *5th SIGdial Workshop on Discourse and Dialogue. Proceedings of the workshop*, 97–100. Cambridge, Mass.: Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W04/W04-2319.pdf. (accessed 13 November 2009).

Steen, Gerard J., Eva Biernacka, Lettie Dorst, Anna Kaal, Irene López–Rodríguez & Tryntje Pasma. to appear. Praglejaz in practice: Finding metaphorically used words in natural discourse. In Graham Low, Lynne Cameron, Alice Deignan & Zazie Todd (eds.), *Researching and applying metaphor in the real world*. Amsterdam & Philadelphia: John Benjamins.

Sweetser, Eve. 1990. *From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Tou Ng, Hwee, Chung Yong Lim & Shou King Foo 1999. A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX99: Standardizing Lexical Resources*, 9–14. ACL Antology W99-0502. Retrieved from: http://www.aclweb.org/anthology/W/W99/.

Vach, Werner. 2005. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology* 58(7). 655–661.