The use of discourse data in language use research[1]

Wilbert Spooren

1. Linguistics and discourse

Lachlan Mackenzie's current research "[...] is being conducted with the framework of Functional Discourse Grammar (FDG). This is the most recent version of Functional Grammar (FG), the tradition of linguistic research initiated by Simon Dik (1940-1995). [...] In the nineties, researchers felt increasing unease with the relation between the organization of the theory and its ambitions. This led to the emergence of Functional Discourse Grammar (FDG) [...]" (Mackenzie, 2004). The move to discourse data can be considered as a distinctive trend of linguistic research in the past decade. Linguists of every persuasion, be they formal (Asher & Lascarides, 1998; Kamp & Van Eijck, 1996), functional (Hengeveld, 2004) or cognitive (Langacker, 2001) have publicly proclaimed their interest in discourse data. Such a move across the board shows that over the years, the notion of 'discourse' has become more and more important in linguistics, a discipline that used to deal exclusively with sentences in isolation and nowadays includes the study of form and meaning of utterances in context. Presently, formal and functional as well as cognitive approaches exist that consider the discourse level as the core object of study.

There seems to be a consensus that what makes a set of utterances a discourse is primarily due to their meaning rather than to their form. More specifically, there is a shared belief that 'discoursehood' is based on the way in which the meanings of the discourse segments can be related to form a coherent message. The 'discoursehood' or coherence is a cognitive phenomenon rather than a textual phenomenon. A discourse is coherent among other things because referential and coherential links exist between the units of which the discourse is made up. And this creates an interesting issue for linguists, who traditionally are eager to rely on explicit linguistic information to base their analyses on, as these links need not be marked explicitly in the text.

In his paper on the methodology of cognitive linguistics, Langacker (1999) takes as a starting point for linguistic research the analysis of *explicit* linguistic elements. This seems to put the analysis of typical discourse phenomena such as coherence relations outside of the realm of linguistics, by definition. Such a stance is at odds with the previously noted move to the discourse level and I believe that it is precipitate: Linguistic regularities behind coherence relations can and should be studied within linguistics. But at the same time I am convinced that the analysis of discourse coherence is in urgent need of reliable and valid analyses of coherence relations. In this contribution I want to present three examples of discourse analytic research that demonstrate the need for analytic reliability. The three examples all make use of corpus data, which is emblematic for current language usage studies. They also relate all three to the study of coherence relations, which happens to be a field of interest that Lachlan and I share.

[1] I want to thank Ted Sanders for his valuable comments on a draft of this paper.

2. Methodological issues of using corpora in the study of coherence relations

Linguists who want to make use of corpora to test their hypotheses come across a number of questions that need to be answered.

2.1 Choice of corpora

The first question to be answered is what kind of corpus to study. One of the challenging findings of language research is that linguistic regularities should in general be indexed for the text type in which they occur. The discourse organizing principles of a narrative are not the same as those of a news report or of a spontaneous conversation. It is not at all self-evident that coherence relations occur similarly in different text types. And given that most corpora are organized in spoken versus written data (see the ELRA 2004 webpage for an example), at some stage the linguist must answer the question: Will I study spoken or written corpora?

A related issue concerns the context in which the corpora were established. For example, a lot of child language research on discourse coherence and connectives uses so-called naturalistic corpora: corpora of utterances produced in a naturalistic situation The typical case is that of a child conversing with an adult about who was to blame for the row between the child and his sister. In this type of setting it can be expected that a lot of arguing will be going on and relatively little description. Hence we will find many so-called argumentative relations (rebuttals, conclusions and the like) and few descriptive relations (causes, temporal relations and the like). Let us suppose that a particular type of coherence relation doesn't show up in that corpus (for instance the denial of expectation relation, as in *John put his coat on, but he wasn't leaving*). Then you don't know whether this is so because the child doesn't know how to use the relation or because that type of relation simply didn't fit in the context.

2.2 Noise in corpora

Language users make errors. However, the data do not come with a tag signalling whether or not the instance was used correctly or erroneously. Therefore the use of discourse corpora comes with a double risk: For one thing we do not know whether or not an utterance is used correctly, for another we do not know what the errors are and what are the correct uses. Table 1 presents us with an illustrative example. The example stems from the so-called ESF-corpus, which consists of utterances of second language learners. The corpus was created in the eighties of the previous century for the benefit of a large scale comparative investigation of the acquisition of second languages. The Dutch part of the corpus contains utterances from speakers of Turkish and Moroccan descent. The example in Table 1 stems from a Moroccan adolescent, codenamed Mohamed, who at the time of recording had been in Holland for two years. Mohamed's task was to have a free conversation with the interviewer. The topic of discussion is discrimination and Mohamed refers to an experience in a discotheque.

The point that I want to make relates to the use of *en* ('and') in line 10. On the basis of the context and sensible assumptions about Mohamed's communicative intentions, it seems likely that Mohamed meant to express an explanation-relation "Yes, they said 'Stupid Turk' to me because they thought that I was a Turk.". If that interpretation is correct then Mohamed should have used *want* or *omdat* ('because') in stead of *en*, and consequently Mohamed has made an error. This need not be the case however. It may be the case that in line 10 Mohamed

is taking up the utterance he started to produce in line 4 (before he was interrupted by the interviewer), in which case the sequence may even be considered correct.

Table 1. *Sample from the ESF corpus*.

Period: last
I: Interviewer
M: Mohammed, informant (Moroccan)
Genre: conversation
Topic: discrimination

M 1 Bijvoorbeeld wij daar	For example we [were] there [in a
M 2 en zeggen “Stom turk”	discotheque]
M 3 en uh ja “Wat jullie doen hier?”	and [they] say “Stupid turk”
M 4 “Moet terug naar eigen land” en eh	And uh yeah “What are you doing here?”
I Zeggen ze dat tegen jou?	“Must go back to [your] own country” and uh
M 5 Ja.	Do they say that to you?
M 6 Hm?	Yes.
M 7 Niet tegen mij	Hm?
M 8 maar hun zeggen zo he.	Not to me
I Tegen jou zeggen ze “Stomme turk”?	But them say so right
M 9 Ja.	To you they say “Stupid turk”?
M 10 En dachten dat ben ik turk was.	Yes.
M 11 Dat is di/discriminatie.	And [they] thought that am I was [a] turk.
	That is di/discrimination.

2.3 Implicit and underspecified coherence relations

The third issue, the one that I want to elaborate upon, relates to the fact that coherence relations are *cognitive* phenomena, and not *linguistic* phenomena: The coherence of a text is not in the verbal material, it resides in the mental representation that readers make of a text. 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 impeded a nuclear transport in the Southern German state Bavaria. (b) Demonstrators chained themselves to the rails.”

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 on the basis of 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, Van Wijk, Terken & Noordman, 1999 for an investigation of the reliability of this type of classification).

In an unpublished paper, Cragg and McGee Wood (2004) state that if we have established that data are reliable, then we have established two important characteristics of the data: (1) The categories used in the analysis "are not inordinately dependent on the idiosyncratic judgements of any individual coder"; (2) we know what the categories that were used in the analysis mean. By conversion, "unreliable data is worthless" (Cragg and McGee Wood, 2004, p. 2). It has been a number of years since Jean Carletta (1996) stressed the importance of reporting interrater agreement (an important prerequisite of data reliability). Nevertheless, explicit accounts of interrrater agreement are rare in the corpus linguistic community.

So, the question to be asked is: If coherence relations can remain unsignalled in the text and if classifying coherence relations is such a great source of individual variation, how can we achieve reliable classifications of those coherence relations? In the following I will demonstrate the extent of the problem, by discussing three articles the authors of which have been explicit about the agreement of the classification of coherence relations.

3. Three research samples: Disagreement in analysing coherence relations

3.1 Analysis of child language data

In a study on the psychological validity of the classification of coherence relations presented in Sanders, Spooren and Noordman (1992), Spooren and Sanders (submitted) have performed an analysis of over 1,100 utterances by children in two age groups (corresponding to grade 1 and grade 6 of elementary school). Each child was given two tasks: a description task and a conversation task. In the description task, children had to describe a number of 'Where's Waldo?' pictures, pictures that show large numbers of causally related events, such as a man walking under a ladder, whistling and unaware that on the ladder a window cleaner has just dropped a bucket of water. Example (2), in which the child describes a picture of a man with gigantic bicepses, is typical for the kind of utterances produced by the description task.

(2) a. Die heeft eerst allemaal spierballen. [That one first has all these muscles]
b. En dan gaan ze ineens kapot. [And then suddenly they all crack]

Bram, 7 years, 2 months

In the conversation task, the children had to formulate an opinion on a number of controversial topics, such as vetoed tv programs, wearing brand clothing at school, the amount of the children's pocket money, etc. Fragment (3) is a typical a product of the conversation task; the child is talking about violence at t.v.

(3) a. Ja, dat vind ik soms wel mooi. [Yes, I like that sometimes]
b. Maar als het in het echt gebeurt, [But if it happens in reality]
c. dan vind ik het niet goed. [then I don't approve of it]

Jeroen, 12 years, 4 months

The utterances of the children were recorded and transcribed, and subsequently coded for a large number of variables, such as type of coherence relation, and a number of theory-specific concepts like source of coherence (is the coherence relation propositional, speech act or epistemic), basic operation (is the coherence relation additive or causal), polarity (is the

coherence relation positive or negative), and order of the coherence relation (is the 'apodosis' of the relation expressed in the first segment or in the second segment).

Per child an agreement score was calculated, using the Cohen's kappa statistic. This statistic relates the amount of agreement found between the two judges to the amount of agreement that can be expected on the basis of chance. It has a value between 0 (absolute disagreement) and 1 (full agreement). To assess the kappa scores, a subset of the data was analysed by two coders independently, following a training session and using a coding scheme containing explicit instructions for the coding. The coders were all very much experienced in the theory of classifying coherence relations.

In the case of this analysis a *K* score was calculated per child for each of the four categorizations. Agreement was lowest for Order of the segments (with *K* varying from .40 to .83 in grade 1, and from .78 to 1.00 in grade 6). For Basic operation *K* varied from .66 to .91 for grade 1 and from .88 to 1.00 for grade 6. For Source of coherence *K* varied from .66 to .88 for grade 1, and from .90 to .98 for grade 6. Finally, for Polarity *K* varied form .54 to 1.00 for grade 1 and from .87 to 1.00 for grade 6.[2]

The exact interpretation of a kappa score is a matter of dispute. Rietveld and Van Hout (1992) cite an interpretation of *K* scores varying from 'slight' ($.00 < K \leq .20$) to 'fair' ($.20 < K \leq .40$), 'moderate' ($.40 < K \leq .60$), 'substantial' ($.60 < K \leq .80$) and 'almost perfect' ($.80 < K \leq 1.00$). This interpretation suggests that even the lowest agreement scores in Spooren and Sanders' (submitted) study are worth mentioning. By contrast, Krippendorff (1980) suggests that variables should only be reported if their reliability is higher than 0.80, and that data with a reliability between 0.65 and 0.80 should only be presented as tentative. Following this suggestion the data from grade 1 should not be reported upon. Of course, reliability is not the same concept as interrater agreement. Reliability is the degree to which the data are independent of the individual measurement (i.e., the degree to which data are stable, reproducible and accurate). Agreement is the degree to which different coders apply a coding scheme similarly. But agreement can be seen as a precondition for reliability and consequently the two are not unrelated.

Therefore, a substantial amount of the data from grade 1 was coded in a manner that can at best be qualified as moderate, and at worst prevents reporting on these data. The question is: If even such experts, with all the effort put into it, and with the help of an explicit coding scheme cannot get a higher interrater agreement, what is the quality of the data?

3.2 Subjectivity in newspaper discourse

In a study of unmarked and marked backward causal constructions in newspaper texts, Spooren, Bekker and Noordman (2001) analyzed the occurrence of subjectivity signals in 712 fragments, 362 fragments from news texts and 359 fragments from opinionating texts.All of

[2] It is interesting to note that Order of the Segments, which prima facie seems a straightforward category to code, gets the lowest interrater agreement. This is caused by the fact that in the coding schema, Order of the Segments does not apply in the case of additive relations. If one coder opts for a causal relation (like Cause-Consequence) and the other for an additive relation (like Temporal Sequence) the two coders will not reach agreement on Order of the Segments. An example is (2) in the text.

the fragments were so-called backward causal relations, like Consequence-Cause (as in (4)) or Claim-Argument (as in (5)).

(4) (S1) Vijf politiemannen raakten gewond doordat (S2) de demonstranten hen met stenen en flessen bekogelden.
(S1) Five policemen were injured because of the fact that (S2) the demonstrators pelted them with stones and bottles.

(5) (S1) De miljoenenbrand bij pluimveehouderij BPC in Barneveld op 4 januari, waarbij negen vrachtwagens uitbrandden, is hoogstwaarschijnlijk aangestoken. (S2) De technische recherche heeft sporen van brandstichting gevonden.
(S1) The multi-million-guilders fire at the poultry farm BPC in Barneveld at January 4, during which nine trucks were completely burnt out, most probably was started deliberately. (S2) The forensics department has found traces of arson.

The coherence relation was either marked explicitly by a connective, as in (4), or unmarked, as in (5).

The subjectivity signals that were studied were:

(6) Subjectivity signals studied in Spooren et al. (2001).
- expression of mental or communicative activity
- expression of contrast or evaluation
- modal elements
- narrative style (direct narrative, direct speech or thought, indirect speech or thought, free indirect speech or thought, mixed)

In order to establish the perspective interpretation of a segment two tests were used (see Oversteegen & Bekker, 2002):

1. Attribution of responsibility
The person who is responsible for the information presented in the sentence is explicitly mentioned in the text: subjective interpretation. If no source is explicitly mentioned, the decisive answer is given by using paraphrase techniques:
- *Volgens mij* (according to me) + segment: if the interpretation of the segment remains the same, then the segment has a subjective interpretation (author-perspective); if the interpretation of the segment changes into a restricted view, then the non-paraphrased segment has a neutral interpretation;
- *Volgens character X (according to character X)* + segment: If the interpretation of the segment remains the same, then the segment has a subjective interpretation (other-perspective); if the interpretation of the segment changes into a restricted view, then the non-paraphrased segment has a neutral interpretation.

2. The deniability test
The test makes explicit the restriction of the validity of the presented information.
- Segment + *maar ik geloof het niet / ik weet wel beter* ('but I don't believe a word of it / but I know better'): if the continuation is possible and does not change the meaning of the segment, then the segment has a subjective interpretation (other-perspective).

Furthermore, following a distinction introduced by Sweetser (1990), it was coded of each fragment whether there was a content relation (propositional relation) or an epistemic relation between the two segments.

In order to establish the interrater agreement of the codings, a subset of the fragments was coded independently by three text linguists, after a training session. Kappa statistics were calculated to assess interrater agreement (cf. Table 2).

Table 2. *Interrater agreement for the variables in the corpus analysis (3 raters, 60 fragments)*.

Variable	kappa
expression of mental or communicative activity	0.66
expression of evaluation or contrast	0.45
presence of a modal element	0.89
narrative style	0.80
perspectivized interpretation	0.75

As the kappa's indicate, agreement was moderate with respect to the expression of mental or communicative activity and poor with respect to the expression of evaluation or contrast. For other variables it was satisfactory to good. If we apply Krippendorff's less charitable interpretation, only two of the variables should be reported upon.

I consider this a significant result. As in the previous case, despite the fact that trained text linguists were used, even the coding of explicit information like expressions of evaluation and contrast does not lead to a high interrater agreement. In order to resolve this problem, in our research the coding of the categories by only one person were used and these codings were based on a very explicit code book and many sessions to discuss the linguistic data. Therefore it seems safe to assume that any coding idiosyncrasies are equally distributed over the parameters in the research.

3.3 Discourse connectives in the Corpus of Spoken Dutch

In a study of discourse connectives in spoken Dutch, Spooren, Huiskes, Degand and Sanders (2002) have compared the use of *want* and *omdat* in written versus spoken language. It has been claimed that the notion of subjectivity accounts for the distributional patterns of these connectives (Pander Maat & Degand, 2001). Subjectivity can be regarded as the degree to which to the speaker/writer is responsible for the causal relation. Dutch *want* seems to have a preference for subjective contexts, whereas *omdat* prefers more objective contexts.

Empirical evidence for such claims consists of corpus studies of written texts. Spooren et al. (2002) have investigated a new source of empirical evidence: the Corpus of Spoken Dutch (Corpus Gesproken Nederlands, CGN, 2004), a 10 million word corpus, which has recently come available . As spontaneous spoken language differs from planned written language, among others, in the direct presence of speaker and addressee, we expected the degree of subjectivity as encoded by the connectives *want* and *omdat* to differ in the two modalities. As a baseline, the results from Degand and Pander Maat (2003) were used.

From the Corpus of Spoken Dutch we selected 150 fragments with *want* and 125 fragments with *omdat* (from the genres interviews and spontaneous conversations). The fragments were analyzed for a number of subjectivity features, such as modality of the related

segments, the nature of the primary causal participant, the syntactic completeness of the utterances, the type of coherence relation, etc. For the sake of interrater agreement each fragment was analyzed independently by two coders.

The variables for which the fragments were coded were (in parentheses are the labels that will be used later to refer to these variables):

- the modality of the first and second segment of the fragment (do the segments express facts, knowledge (individual or general), perceptions, experiences, judgments or intentional acts from the main protagonist) (M1, M2);
- the identity of the main protagonist in the two segments (is the main protagonist the speaker, the addressee, a third person, or – if the segment expresses a fact – is the identity of the conceptualizer irrelevant) (P1, P2);
- is the identity of the main protagonist the same in first and second segment (P=)?
- are the main protagonists indicated explicitly in the segments (Pex1, Pex2)?
- from whose perspective is causal relation between the two segments presented (the author, the main protagonist, some other person) (Per)?
- what is the coherence relation between the two segments (non-volitional, volitional, epistemic, speech act, textual, other) (CR)?
- what is the size of the segments (major constituent, clause or sentence, more than a sentence) (S1, S2)?
- what is the grammatical form of the segments (declarative clause, interrogative clause, imperative clause, elliptical element, incomplete (= interrupted) element) (GF1, GF2)?
- are the two segments uttered by the same speaker (Spe)?
- is there a syntactic modification of the connective (adverbial, other connective, focal construction, interjection, no modification) (Mod)?

Example (6) gives a typical sequence from the corpus (speakers A and B are talking about the bus lanes in the city):

(6)

a	A	en uh daar komt dan die andere busbaan ook naar beneden die sluit dan aan op die busbaan ja dus d ... and uh there does that other bus lane also come down, that one must connect to that bus lane yes so th ...
b	B	oh op diezelfde? oh to the same one?
c	B	maar hoe komt ie dan bij 't station d'rdoor? but how does it get through at the [railway] station?
d	B	WANT dat hebben ze doorgeknipt. because they cut that

The fragment that was coded here, are the two segments in (6c) and (6d). Both segments were coded as intentional acts (M1, M2), the two main protagonists in the two segments (the bus [driver] in segment 1 and *they* in segment 2)(P1, P2) are expressed as 3rd person pronouns, the two protanogists are indicated explicitly (Pex1, Pex2), the causal relation is presented from the perspective of the speaker (Per), the coherence relation (CR) is a speech act relation, the size of each segment is a clause (S1, S2), the grammatical form of the first segment (GF1) is an interrogative sentence, that of the second segment (GF2) is a declarative sentence, the two

segments are uttered by the same speaker (Spe) and the connective *want* is not syntactically modified (Mod).

In a first run, pairs of coders analysed a subset of the corpus independently, on the basis of a code book. This code book was based on extensive discussions of examples and contained very explicit instructions on how to analyse the data. All coders were experienced discourse analysts. On the basis of the first analyses, Cohen's kappa's were calculated.[3] The following table contains *K* coefficients for the variables mentioned that were obtained by one pair of coders.

Table 3. *Kappa (K), absolute (Agr) and relative (Perc) number of agreements between two coders (max. # of agreements is 18).*

	M1	M2	P1	P2	P=	Px1	Px2	Per	CR	S1	S2	GF1	GF2	Spe	Mod
K	12	32	34	62	-19	22	31	0	18	55	-15	65	47	-8	10
Agr	5	8	12	14	11	10	12	13	6	13	13	14	16	15	13
Perc	28	44	67	77	61	56	67	72	33	72	72	78	89	83	72

Again, as in the previous cases, *K* is very low, despite explicit instructions and clear procedures. Only two out of fifteen variables have a Kappa value that is moderate (in the charitable interpretation). Some of the values are excessively low.

Table 3 contains a number of interesting data. For instance, we see cases where the agreement statistic is extremely low, even negative, even though the agreement itself (absolutely and relatively) if fairly high. This typically is the case where few (say, two) categories were used and one of those categories is favoured by the coders. Consider the size of segment 2 (S2) as an example. In the sample from the corpus that was analysed by the two coders, only two values occurred: a clause or sentence (value 1) and more than a sentence (value 2). The two values were used by the coders as follows:

Table 4. *Coding of 'size of segment 2' by two coders.*

	Value '1'	Value '2'
Value '1'	13	3
Value '2'	2	0

In this case we see that there is strong agreement about value '1' (coder 1 used this value 16 times, coder 2 used it 15 times, the two coders agreed in 13 cases), but absolutely no agreement about value '2'. Therefore, even though there is a considerable amount of agreement (72 %), the agreement statistic is extremely low (even negative). One interpretation is that agreement statistics are not suited for such small-sized, skewed tables: Since one of the categories is strongly preferred, this means an increase in the likelihood that coders agree by chance (by choosing the preferred category) and consequently the overall agreement is lowered.

[3] For the calculations of Cohen's Kappa the website http://www.kokemus.kokugo.juen.ac.jp/service/kappa-e.html was used.

In this study, the researchers have decided to discuss the discrepancies, and to use only those data about which the pairs of coders could reach agreement after discussion. But again the issue is raised: If trained and experienced analysts reach such a low interrater agreement, what is the value of the data used in the analysis?

4. How to deal with these issues?

This paper is not going to provide any solutions to the problems mentioned above. But what it wants to raise is the issue of low interrater agreement, that, as far as I can see, is a frequently occurring problem in discourse analytic accounts of coherence. What does a low Kappa score (or other agreement measure[4]) mean? It means that the theoretical categories cannot be applied with any confidence. At the same time, these categories may still be interesting. The problem is that we cannot use the categories in a consistent manner. As Craggs and McGee Wood (2004) 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 this subjective phenomena worthy of study shows that we are, in fact "willing to rely on imperfect data", which is fine as long as we recognise the limitations of a scheme which delivers less than ideal levels of reliability, and use the resulting annotated corpora accordingly."

I see at least the following venues:[5]

1. We must commit a lot of energy to clear definitions of the variables we are coding for. For the sake of replication, our coding categories should be available to the research community, along with our research results.
2. Researchers are well advised to use a limited number of mutually exclusive coding categories per variable of interest. Coding becomes very difficult if we have to choose between 10 categories or more per variable.
3. As much as possible, we should be coding explicit information. These are easiest to code.
4. From 3. it follows that in case of elusive, semantic categories we should try to approximate our variables of interest. If we want to code a category like 'subjective information' let us try to find explicit operationalizations that approximate such a semantic notion (for instance, by using an exhaustive list of markers of speaker attitude).
5. Visualization of the coding process may help in reducing the amount of errors. In this respect, software like MMAX2 (cf. MMAX 2004) might be useful.
6. Researchers must be explicit about the interrater agreement. Readers have a right to know how stable and robust a coding schema is.

---

[4] Craggs and McGee Wood (2004) argue in fact that Kappa should not be used, as its calculation of chance agreement is dependent on individual preferences. Instead they argue for an agreement score that makes chance agreement independent of individual preferences, such as Krippendorff's (1980) Alpha.

[5] My thinking about these issues has been stimulated much by the discussions at the Discourse Annotation workshop of ACL Barcelona, July 2004 (see Webber and Byron, 2004) and by discussions with Mike Huiskes. Some of the ideas mentioned here stem directly from these discussions.

7. In case of low interrater agreement (say, less than 0.80), we should be explicit about the restricted amount of generalizability of our research results.
8. An interesting way of corroborating a coding schema is by testing it indirectly. Degand, Spooren and Bestgen (2004) used two automatic semantic extraction techniques to test linguistic hypotheses based on hand-coded corpora. As the completely automatic, analyst-independent analysis gave the same linguistic patterns as earlier hand-based analyses, Degand et al. took this as evidence for the correctness of the analyst-dependent research results.

References

Asher, N. and A. Lascarides (1998). Bridging, *Journal of Semantics, 15* (1), 83-113.

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

CGN (2004). Corpus Gesproken Nederlands. http://lands.let.kun.nl/cgn/ehome.htm) (consulted on July 13, 2004).

Craggs, R., & McGee Wood, M. (2004). *Evaluating discourse and dialogue coding schemes*. Unpublished manuscript, Manchester (UK).

Degand, L. & Pander Maat, H. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In A. Verhagen & J. van de Weijer (Eds.), *Usage based approaches to Dutch* (pp. 175-199). Utrecht: LOT.

Degand, L., Spooren, W. & Bestgen, Y. (2004). On the use of automatic toolsfor large-scale semantic analyses of causal connectives. In B. Webber & D. Byron (Eds.), *Proceedings of the 2004 ACL Workshop on Discourse Annotation* (pp. 25-32). East Stroudsberg, PA: Association for Computational Linguistics. ISBN: 1-932432-38-8.

ELRA (2004). The European Language Resource Association. http://www.elra.info/ sub Catalogue (consulted on May 28, 2004).

Hengeveld, K. (2004). The architecture of a Functional Discourse Grammar. In J. Lachlan Mackenzie & M. Á. Gómez-González (Eds), *A New Architecture for Functional Grammar* (pp. 1-21). Berlin etc.: Mouton de Gruyter. Functional Grammar Series 24.

Kamp, H. & Eijck, J. van (1997). Representing discourse in context. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of Logic, Language and Information* (pp. 179-237). Amsterdam etc.: Elsevier.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology* (Vol. 5). Beverly Hills etc.: Sage.

Langacker, R. (1999). Assessing the cognitive linguistic enterprise. In Th. Janssen & G. Redeker (Eds.), *Cognitive Linguistics: Foundations, Scope, and Methodology* (pp. 13-59). Berlin etc.: Mouton de Gruyter. Cognitive Linguistics Research 15.

Langacker, R. (2001). Discourse in Cognitive Grammar. *Cognitive Linguistics, 12*(2), 143-188.

Mackenzie, J.L. (2004). Research interest. Lachlan Mackenzie's personal webpage: http://www.let.vu.nl/staf/jl.mackenzie/index_en.htm (consulted on May 28, 2004).

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*, 243-281.

MMAX (2004). MMAX annotation tool. Available from http://www.eml-research.de/english/research/nlp/index.php (consulted on July 30, 2004).

Ouden, H. den, Wijk, C. van, Terken, J.M.B. & Noordman, L.G.M. (1999). *Reliabilty of discourse structure annotations.* IPO Annual Report (Ext. r. no. 33). Eindhoven: IPO, 10 pp.

Oversteegen, L., & Bekker, B. (2002). Computing perspective: the pluperfect in Dutch. *Linguistics, 40*(1), 111-161.

Pander Maat, Henk and Liesbeth Degand (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics, 12* (3), 211-245.

Rietveld, T., & Van Hout, R. (1992). *Statistical techniques for the study of language and language behaviour*. Berlin etc.: Mouton de Gruyter.

Sanders, T., Spooren, W., & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1-35.

Spooren, W., Bekker, B. & Noordman, L. (2001), *Reversed order and subjectivity in different text types*. In Degand, L., Bestgen, Y., Spooren, W. & Van Waes, L. (Eds), *Multidisciplinary Approaches to Discourse 2001* (pp. 61-72). Amsterdam etc.: Stichting VU Neerlandistiek-Nodus.

Spooren, W., Huiskes, M., Degand, L., & Sanders, T. (2002). Connectieven in geschreven en gesproken taal [Connectives in written and spoken language]. Paper presented at the VIOT 2002 (December 2002, Anterwerpen).

Spooren, W. & Sanders, T. (submitted). *The acquisition order of coherence relations: On cognitive complexity in discourse*.

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

Webber, B. & Byron, W. (2004, Eds.). *Proceedings of the 2004 ACL Workshop on Discourse Annotation*. East Stroudsberg, PA: Association for Computational Linguistics. ISBN: 1-932432-38-8.