Project notes of CLARIN project DiscAn: Towards a Discourse Annotation system for Dutch language corpora

Ted Sanders
University Utrecht
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht
T.J.M.Sanders@uu.nl

Kirsten Vis
University Utrecht
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht
K.Vis@uu.nl

Daan Broeder
TLA - Max-Planck
Institute for Psycholinguistics
Wundtlaan 1
NL-6525 XD Nijmegen
Daan.Broeder@mpi.nl

Abstract

Although discourse is a crucial level in language and communication, many existing corpora of Dutch language lack annotation at this level. This paper describes the recently started DiscAn project, which sets the first step to change this situation for Dutch, in line with international tendencies. The project has five main goals: 1) to standardize and open up an existing set of Dutch corpus analyses of coherence relations and discourse connectives; 2) to develop the foundations for a discourse annotation system that can be used in Dutch natural language corpora; 3) to improve the metadata within European research infrastructure project CLARIN by investigating existing CMDI profiles or adding a new CMDI profile specially suited for this type of analysis; 4) to inventorize the required discourse categories and investigate to what extent these could be included in ISOcat categories for discourse that are currently being developed; 5) to further develop an interdisciplinary discourse community of linguists, corpus and computational linguists in The Netherlands and Belgium, in order to initiate further research on cross-linguistic comparison in a European context.

1 Introduction

Over the years, the notion of “discourse” has become increasingly important in linguistics - a remarkable development, considering that linguistics used to deal almost exclusively with sentences in isolation. Nowadays, the discipline includes the study of form and meaning of utterances in context, and formal, functional, and 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 into genuine discourse is (primarily) their meaning rather than their form. More specifically, there is a shared belief that “discoursehood” is based on the possibility to relate discourse segments to form a coherent message (Kehler, 2002; Sanders, Spooren & Noordman, 1992; Taboada & Mann, 2006; Wolf & Gibson, 2005).

Language users establish coherence by relating the different information units in the text. The notion of coherence has a prominent place in both (text-)linguistic and psycholinguistic theories of text and discourse. When confronted with a stretch of discourse, language users make a coherent representation of it. At the same time, discourse itself contains (more or less) overt signals that direct this interpretation process. In general, two types of coherence and their textual signals are distinghuished: (i) Referential coherence: how does reference to individuals create continuity and (as a result) coherence? The linguistic signals considered involve reference to persons (*Beatrix* , *she*, *the professor*), objects and concepts; (ii) Relational coherence: how do coherence relations like causals and contrastives constitute connectedness? The linguistic signals considered are connectives and lexical cue phrases. This project focuses on the second type of coherence.

Existing corpora of natural language use often lack systematic information on the discourse level. For Dutch corpora like the Corpus of Spoken Dutch (‘Corpus Gesproken Nederlands’, CGN), for instance, lexical, syntactic and even semantic annotations are available, but typical discourse phenomena like referential and relational coherence are not addressed. Still, the discourse level is a crucial level of description for language and communication

Internationally, the last decennium has shown a tendency to change this situation. Initiatives like the Penn Discourse Treebank (Prasad et al., 2008) and the RST treebank (Carlson & Marcu, 2001) aim at creating a level of corpus annotation focusing on discourse structure information. The DiscAn project aims at developing the first step in this direction for the Dutch language community, with the explicit ambition of taking it to a cross-linguistic level. The project, that runs from April 1, 2012 until April 1, 2013, is part of and funded by CLARIN, a large-scale European research infrastructure project designed to establish an integrated and interoperable infrastructure of language resources and technologies, cf. www.clarin.nl.

2 Research data

The first aim of the DiscAn project is to integrate existing corpora of Dutch discourse phenomena in the CLARIN infrastructure, in order to standardize a valuable amount of corpus work on coherence relations and discourse connectives, and to make it available and more easily accessible for a much wider range of researchers in the humanities in general and in linguistics in particular.

The data in the existing corpora take various forms. They typically exist as fragments in doc files from scanned or copied files from newspaper, chat, spoken or child language corpora, which are analyzed on discourse variables using a systematic annotation scheme or code book. The analysis is usually available in the form of excel- or SPSS-files. Table 1 below presents a global overview of corpora, the discourse phenomena analyzed, the type of corpus, as well as the amount of analyzed cases.

3 Annotation Scheme

The various corpora have not been analyzed in identical ways, but large similarities exist with respect to the basic categories that are present in every analysis. An important part of the DiscAn project is the conceptual and text-analytical work that needs to be done, in order to identify overlapping of relevant categories, to make the analyses comparable. Earlier international work (Mann & Thompson, 1988; Sanders et al., 1992; Sanders, 1997; Sweetser, 1990; Taboada & Mann, 2006; Wolf & Gibson, 2005) will be inspiring and leading here. The Penn Discourse Treebank (Prasad et al., 2008) provides a classification, as Bunt et al. (2012) do. We expect to see similarities, but also deviations from these proposals, for both theoretical and empirical reasons. The results from our first applications to corpora will shine a light on the validity of our classification. In sum, based on existing theoretical and analytical work, the basic categories include:

- polarity: positive / negative relation (*because/omdat* and *and/en* versus *but/maar* and *although/hoewel*);
- nature: causal / temporal / additive (*because/omdat*, *then/toen*, *and/en*)
- order: antecedens-consequens or vice versa (*therefore/dus*, *because/omdat*)
- subjectivity: objective / content (*as a result/daardoor*) vs. subjective / epistemic (*therefore/dus*) vs. speech act (*so/dus*)
- perspective: subject of consciousness; first, second person, etc.
- adjacency: how far are the related segments apart?
- linguistic marking of relations: connectives / lexical cue phrase / implicit
- semantic-pragmatic characteristics of segments: modality, tense and aspect.

The discourse analytical data is available in various formats: excel tables, doc files, SPSS files etc. The data in the DiscAn project will be made available in a uniform and acceptable format, both in terms of metadata and discourse annotation categories.

Discourse phenomena	**Author**	**Cases**
Causal connectives	Bekker (2006)	500 explicit (*doordat, want, dus, daarom, nadat, voordat*) / 200 implicit
Causal connectives	Degand (2001)	150 (*want, aangezien, omdat*) from newspapers
Coherence relations	Den Ouden (2004)	70 (causal implicit, non-causal)
Connectives	Evers-Vermeul (2005)	600 historical data / 4400 from Childes
Causal connectives	Pander Maat & Degand (2001)	150 (*dus, daarom*) from newspaper corpora
Coherence relations	Pander Maat & Den Ouden (2011)	795 implicit and explicit relations from a self-assembled corpus of 40 press releases
Causal connectives	Pander Maat & Sanders (2000)	150 (*dus, daarom, daardoor*) from a newspaper-corpus (Volkskrant)
Causal connectives	Persoon (2010)	105 (*omdat, want*) from CGN
Causal connectives	Pit (2003)	200 (*aangezien, omdat, doordat, want*) newspaper / 100 (*omdat, doordat, want*) narrative; from newspaper (Volkskrant) and fictional books
Causal connectives	Sanders & Spooren (2009)	100 newspaper (Volkskrant) / 275 from CGN / 80 from Chat (*want, omdat*)
Coherence relations	Sanders & van Wijk (1996)	100 childrens explanatory texts; ca. 1500 coherence relations
Coherence relations	Spooren & Sanders (2008)	1100 coherence relations (children elicit responses)
Causal connectives	Spooren et al. (2010)	275 (*want, omdat*) spoken, from CGN; 100 (*want, omdat*) written
Causal connectives	Stukker (2005)	300 (*daardoor, daarom, dus*) newspaper / 300 historical data (*daarom, dus*)
Coherence relations	Vis (2011)	135 texts; 643 subjective relations
Connectives	Van Veen (2011)	1951 *waarom- (why-)* questions and their answers (Childes)

Table 1: Overview of DiscAn corpora.

3.1 Importance of DiscAn

The availability of this corpus, with its possibility to search on discourse terms, will be of great importance to many linguists, especially those interested in discourse structure in language use. In addition to the particularly large group of discourse analysts, text linguists and applied linguists working on text and discourse, we can think of theoretical linguists working on the syntax-semantics-discourse interface, language acquisition researchers, sociolinguists interested in language variation, as well as researchers in the field of (language and) communication. However, the merits of the DiscAn project are not limited to the availability of these corpora. The standardized annotation scheme that was used for the subcorpora will be used to further to develop the foundations for a discourse annotation system that can be used to apply in existing Dutch natural language corpora. The standardized discourse category coding scheme developed in the first phase, will be the basis for this second phase. Finally, we expect to be able to contribute to the ISOcat categories for discourse that are currently being developed. The end product of DiscAn will be a set of annotated subcorpora with discourse coherence phenomena which will allow researchers to search for connectives and the way they are used, but also, for instance for a certain type of causal relation in spoken discourse. Researchers interested can be found in linguistics and language use (syntax, semantics, child language) and communication studies (subjectivity, variance across genres and media).

References

Birgit Bekker. 2006. *De feiten verdraaid. over tekstvolgorde, talige markering en sprekerbetrokkenheid.* Doctoral dissertation, Tilburg University, Tilburg, The Netherlands.

Harry Bunt, Rashmi Prasad and Aravind Joshi. 2012. *First steps towards an ISO standard for annotating discourse relations.* Proceedings of ISA-7 workshop (Interoperable Semantic Annotation) at LREC 2012, Istanbul.

Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual.* http://www.isi.edu/ marcu/discourse/

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

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

Hanny den Ouden. 2004. *Prosodic realizations of text structure.* Doctoral dissertation, Tilburg University, Tilburg, The Netherlands.

Jacqueline Evers-Vermeul. 2005. *The development of Dutch connectives: Change and acquisition as windows on form-function relations.* Doctoral dissertation, Utrecht University, Utrecht, The Netherlands

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

Andy Kehler. 2002. *Coherence, reference and the theory of grammar.* Chicago: The University of Chicago Press.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes 18*: 3562.

Alistair Knott and Ted Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics 30*: 135175.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text 8* (3), 243-281.

Henk Pander Maat and Liesbeth Degand. 2001. Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics 12*, 211245.

Henk Pander Maat and Ted Sanders. 2000. Domains of use or subjectivity: The distribution of three Dutch causal connectives explained. In: Elizabeth Couper-Kuhlen and Bernd Kortmann (eds.), *Cause, condition, concession, and contrast: Cognitive and discourse perspectives*, 5782. Berlin et al.: Mouton de Gruyter.

Ingrid Persoon, Ted Sanders, Hugo Quené and Arie Verhagen. 2010. Een coördinerende omdat-constructie in gesproken Nederlands? Tekstlinguïstische en prosodische aspecten. *Nederlandse Taalkunde*, 15, 259-282.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0.* In Proceedings of LREC08.

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

Ted Sanders and Wilbert Spooren. 2009. Causal categories in discourse - Converging evidence from language use. In Ted Sanders and Eve Sweetser (eds.). *Causal categories in discourse and cognition.* (pp. 205-246). Berlin: Mouton de Gruyter.

Ted Sanders, Wilbert Spooren and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1-35.

Ted Sanders and Wilbert Spooren. 2008. The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics, 40* (12), 2003-2026.

Wilbert Spooren, Ted Sanders, Mike Huiskes and Liesbeth Degand. 2010. Subjectivity and Causality: A Corpus Study of Spoken Language. In: Sally Rice and John Newman (eds.) *Empirical and Experimental Methods in Cognitive/Functional Research.* (pp.241-255). Chicago: CSLI publications.

Manfred Stede. 2004. *The Potsdam commentary corpus.* Proceedings of the ACL-04 workshop on discourse annotation. Barcelona, July 2004.

Ninke Stukker. 2005. *Causality marking across levels of language structure: a cognitive semantic analysis of causal verbs and causal connectives in Dutch.* Doctoral dissertation, Utrecht University, Utrecht, The Netherlands.

Ninke Stukker, Ted Sanders and Arie Verhagen. 2008. Causality in verbs and in discourse connectives. Converging evidence of cross-level parallels in Dutch linguistic categorization. *Journal of Pragmatics 40* (7), 1296-1322.

Ninke Stukker and Ted Sanders. 2011. Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics, 44* (2), 169-190.

Eve Sweetser. 1990. *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

Maite Taboada and William Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies, 8* (4), 567-588.

Rosie Van Veen. 2011. *The acquisition of causal connectives: the role of parental input and cognitive complexity*. PhD dissertation. Utrecht University, Utrecht, The Netherlands.

Kirsten Vis. 2011. *Subjectivity in news discourse: A corpus-linguistic analysis of informalization*. Amsterdam: VU University Amsterdam dissertation.

Florian Wolf and Edward Gibson. 2005. Representing Discourse Coherence: A corpus-based study. *Computational Linguistics 31* (2), 249-287.

Florian Wolf and Edward Gibson. 2006. *Coherence in natural language. Data structures and applications.* MIT Press, Cambridge Mass.