CLARIAH project

ACAD: Automatic Coherence Analysis of Dutch

CC 17-002

Deliverable 14: Description of the project

Date: May 23, 2018

Author: Wilbert Spooren
Centre for Language Studies
Radboud University Nijmegen

Introduction

The main goal of ACAD (Automatic Coherence Analysis of Dutch) is to develop an environment in which linguists without programming skills can perform large-scale corpus analyses of discourse coherence phenomena. This document gives a description of the entire project. It makes use of existing deliverables and of two research papers that have been produced during the project (Komen & Hoek, 2018; Hoek & Spooren, 2018).

Composition of the team

The team of ACAD was composed of the following members:
dr. Henk van den Heuvel (Centre for Language and Speech Technology, Radboud University, CLARIAHCLARIN C Centre), director
Micha Hulsbosch (Humanities Lab, Faculty of Arts, Radboud University), programmer
Jet Hoek, MA (UIL-OTS, Utrecht University), text linguist
dr. Erwin R. Komen (Humanities Lab, Faculty of Arts, Radboud University), programmer
prof. dr. Ted Sanders (UIL-OTS, Utrecht University), text linguist
prof. dr. Wilbert Spooren (CLS, Radboud University), text linguist

The team was supported by two student assistents (Patrick Sonsma and Iris Hofstra), who were paid from external funding. Initially, the Meertens Institute was involved as CLARIAH Centre, to host the products of the project but with the approval of CLARIAH it was decided that for practical purposes CLST would act as the the CLARIAHCLARIN Centre involved. From the CLARIAH board, dr. Lex Heerma van Vos was appointed as the CLARIAH contact. He was invited for the project meetings, and attended the closing meeting.

Project management

For the purposes of the project a directory was created at Surfdrive, where files could be exchanged between team members. There was also a subdirectory "Reports on meetings" to which Heerma van Vos had access. The team held various meetings.

June 14, 2017 (opening meeting)
September 29, 2017
November 6, 2017
December 6, 2017
February 6, 2018
March 18, 2018
April 25, 2018 (closing meeting)

The opening meeting took place at Utrecht University, all other meetings were held at Radboud University. All meetings had an agenda, and reported on the progress of different phases of the work. Minutes of the meetings were made available via the ACAD-directory at Surfdrive.

Content of the project
The goal of ACAD is to develop an environment in which computationally naive discourse analysts can carry out an automatic analysis of causal coherence in discourse. The project specifically aims at developing a linguists-oriented search interface for which the case of causal coherence serves as a point-of-proof. In recent years, many proposals have been developed to distinguish between objective and subjective coherence relations like Cause-Consequence (objective) and Claim-Argument (subjective). Dutch has a relatively rich repertoire of causal connectives that can express causal relations that differ in subjectivity (Sanders & Spooren, 2015). Several studies have raised the question to what extent the distribution of these connectives depends on genre and register (e.g., Spooren et al., 2010). Empirical studies touching upon this topic are typically manual analyses of small corpora (for an exception see Bestgen et al., 2006), the results of which have for example been summarized and standardized in the CLARIN DiscAn project (Sanders et al., 2012). ACAD intends to scale up such small-scale analyses. Therefore, the research question of the project is:

> To what extent do the results of causal coherence analysis in different genres in terms of subjectivity based on small-scale studies hold for large datasets?

The answer to this question is formulated in the scientific paper that is milestone 3 in the list of deliverables, below (Hoek & Spooren, 2018).

Design of the project
The aim of ACAD is to use CLARIAH components to automatically analyse different corpora representing different genres for the use of causal connectives and the subjectivity of the environments in which these connectives occur. A causal connection consists of two segments (S1 and S2) connected through a causal connective (a coordinating or subordinating causal conjunction or adverbial causal conjunction). The complete task consisted of several subtasks.

1. Prepare a research corpus. This involves gathering and conversion of existing corpora for most of the genres that are needed, as well as adding new corpora to cover the genres that are missing for our research. During the project it was decided, for the sake of speed and simplicity, to focus the results of our analyses on three genres, which vary in degree of spontaneity with which the utterances are produced (Clark, 1996), the degree of fragmentedness/integratedness of the utterances (Chafe, 1994) and the modality that is used by the speaker/writer (written, spoken, computer-mediated). These genres are spontaneous conversations (represented by the Corpus of Spoken Dutch), newspaper texts (represented by the SoNaR subsection of newspapes) and WhatsApp conversations (represented by two newly added subcorpora). Of course, it is possible to carry out the queries in all the corpora that are available in the search engine.

2. Identification of the relevant cases in the corpus. Queries were formulated to select all causal connectives from existing corpora and from corpora prepared in phase 1. For the sake of simplicity, it was decided to focus the queries and the evaluation of their results on four cases of causal connectives, that belong the most frequent ones in Dutch. The queries that have been formulated can act as models for subsequent queries.

Medial *omdat*: subordinate conjunction that according to the literature expresses relatively objective backward causal relations (like Consequence-Cause and Action-Reason)

Want: coordinating conjunction that expresses relatively subjective backward causal relations (like Claim-Argument)
Daarom: adverbial conjunction that expresses relatively objective forward causal relations (like Cause-Consequence and Reason-Action)
Dus: adverbial conjunction that expresses relatively subjective forward causal relations (like Argument-Claim)

3. Identification of S1 and S2. The span of the discourse segments that are connected by the causal connectives (S1 and S2) was detected automatically on the basis of queries. The queries were formulated in Cesar (**C**orpus **E**ditor for **S**yntactically **A**nnotated **R**esources, described in deliverable D2 and in Komen & Hoek, 2018). Cesar is a web interface that allows the researcher to define searches, local and global variables, conditions to maintain elements in the result set, and output features. Behind the scene these searches are translated into the the Xquery language. The actual search takes place in a component of the CLARIN-NL CorpusStudio web application. The Cesar web interface[1] is hosted at CLST being a CLARIN C Centre.

The quality of this step—the identification of S1 and S2—was evaluated separately. It is reported in Hoek and Spooren (2018).

4. Identification of the directionality of the causal connection. This directionality is a direct function of the type of the connective and its location (initial, medial). In the case of the chosen connectives, *omdat* and *want* are backward connectives (the consequence of the causal relation is presented in the first segment S1), whereas *daarom* and *dus* are forward connectives.

5. Determination of the subjectivity of the connection. The amount of subjectivity of S1 and S2 was determined on the basis of a so-called thematic analysis (Bestgen et al., 2006). For this we made use of a subjectivity lexicon provided by De Smedt and Daelemans (2012), which consists of circa 1000 adjectives with subjectivity scores ranging from 0 to 1. Somewhat arbitrarily, we opted for considering adjectives with scores ≤.20 as objective and adjectives with scores ≥.70 as subjective. S1 and S2 were searched for both adjectival and adverbial uses of these adjectives. A second operationalisation of establishing subjectivity was counting the number of verbs of volition and modal verbs (as specified in Haeseryn et al.,1997). For details see Hoek and Spooren (2018).

Overview of the deliverables

Table 1 provides an overview of the deliverables of ACAD. Several parts had been envisioned as separate deliverables; during the project, it was decided that these could and should be included in M3, the paper on subjectivity and causal coherence. These concern:

D4: evaluation of retrieval of connectives
D9: evaluation of queries to identify directionality of the connectives and the identification of S1 and S2
D11: evaluation of the queries to identify the subjectivity of the segments.

As to the corpus data, ACAD has produced several forms of output.
a) Existing corpora have been enriched. A concrete example is VU-DNC, which has been extended with a syntactic parse. As VU-DNC is maintained by INT[2], it was decided to store

[1] https://cesar.science.ru.nl/
[2] https://ivdnt.org/producten/vu-dnc/index.html

this enrichment at INT. Other existing corpora, such as the CGN and several large parts of Lassy Large's Sonar component, have been transformed into the FoLiA *xml* format.[3]
b) New corpora have been added, along with the lemmatisation, POS-tagging and syntactic parse. This concerns NRC newspaper data from the paper version and the digital version of NRC andWhatsApp data collected in previous research. For all of these materials, IPR statements are available, allowing the use of these data for scientific research. The data and their enrichments will be stored at DANS.

Table 1. Overview of deliverables from ACAD

ID	Deliverable	Type	Person/institute r	Status	Location
D1	Text preparation; identification of the selection of the corpora	Data	CLS	Ready	
M1	Alpino conversion of new texts; description metadata in CMDI	Metadata	CLS-HumLab	Ready	
D2	Evaluation of the results	Document	CLS	Ready	Surfdrive
S1	Creation of front-end onto CorpusStudio end with a syntactic tree visualize	Software	CLS-HumLab	Ready	cesar.science.ru.nl
D3	Test of front-end; retrieval of causal uses of connectives; feedback to developer	Data	UU	Ready	n.a.
D4	Evaluation of the results	Document	CLS	Part of M3	
S2	Embedding in Clariah Centre	Software	CLST	Ready	Cesar is available via CLST; corpora are available via DANS
D5	Formulate queries to identify marker use (S1-conn-S2 versus conn-S1-S2)	Data	UU	Ready	Cesar
D6	Formulate queries to identify directionality of connective	Data	UU	Ready	Cesar
D7	Interaction with Researcher + adaptation of interface	Data/Software	CLS-HumLab	Ready	n.a.
D8	Training of Researcher to evaluate intermediate queries	Data	CLS-HumLab	Ready	n.a.
D9	Evaluation of the results	Document	CLS	Part of M3	
M2	Paper on front-end	Milestone	CLS-HumLab	Submitted to CLIN journal	
D10	Formulate queries to identify subjectivity of hits	Data	UU	Ready	Cesar
S3	Improvement of interface	Software	CLS-HumLab	Ready	n.a.
D11	Evaluation of the results	Document	CLS	Part of M3	
D12	Statistical analysis of subjectivity	Data	UU	Ready	Part of M3
M3	Paper on subjectivity and causal coherence	Milestone	CLS	Almost finished	
D13	Evaluation of Clariah components; suggestions for improvements	Document	CLS-HumLab	Ready	Surfdrive
D14	Description of project	Document	CLS	This document	Surfdrive
D15	Description of desired changes to Clariah components	Document	CLS-HumLab	Ready	Surfdrive

[3] Much of Lassy Large's Sonar component had previously been converted to FoLiA for the CorpusStudio web application project. All the syntactic parses were turned into FoLiA using FoliaParse (Komen, 2015).

Evaluation of the work and future directions

The ACAD project allowed us to build a search engine for automatic corpus analysis that allows us to test text linguistic hypotheses on a scale that was previously unimaginable without programming skills. The combined work of text linguists and programmers proved fruitful, and we believe that we have created a powerful search engine with a lot of scientific potential.

Nevertheless, we also see points of improvements, which in part have been described in other deliverables:

- Speed up the Alpino parser
- The backend of Cesar is hosted at a SurfSara virtual machine. Speed and reliability would be improved if it were hosted elsewhere.
- Improve FoliaParse as described in D13-D15.
- Although Cesar targets an uninitiated user, it still has a steep learning curve. The current project did not budget learning materials and user manuals. We try to produce some examples and manuals for new users, but the full explorative power of Cesar would benefit from a rich learning environment.
- It is tempting to extend the search engine Cesar to new areas. One can imagine using the tool for completely new research questions, not necessarily related to the analysis of causal coherence. One could also imagine using the machinery for the analysis of other languages. Although such extensions are in principle possible, it should be noted that the current version of CESAR has been developed to facilitate searches that start out with *words* that need to be found. It is only in the final month of the ACAD project that searches starting out with *constructions* have been added to CESAR (e.g. pronouns, relative clauses, main or subordinate clauses with a particular word order).

References

Bestgen, Y., Degand, L. & Spooren, W.P.M.S. (2006). Toward automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes, 41*(2), 175-194.

Chafe, W. (1994). *Discourse, consciousness and time*. Chicago: Chicago University Press.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press..

De Smedt, T. & Daelemans, W. (2012). 'Vreselijk mooi!'(terribly beautiful'): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)* (pp. 3568-3572).

Haeseryn, W., Romijn, K., Geerts, G., Rooij, J. de, & Toorn, M. van den, (1997). *Algemene Nederlandse Spraakkunst* (2e, geheel herz. dr. ed.). Groningen: Martinus Nijhoff.

Hoek, J. & Spooren, W. (2018, in preparation). Automatic coherence analysis of Dutch.

Komen, Erwin R. (2015). *Surfacing Dutch syntactic parses*. Amsterdam. http://wordpress.let.vupr.nl/clin26/abstracts/.

Komen, E. & Hoek, J. (2018, submitted). Automatic coherence analysis for non-programmers.

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

Sanders, T., Vis, K., & Broeder D. (2012). Project notes of CLARIN-project DiscAn: Towards a Discourse Annotation system for Dutch language corpora. Downloadable proceedings of the Eigth Workshop on Interoperable Semantic Annotation (isa-8), Pisa, Italy, October 3-5, 2012 (downloadable from http://sigsem.uvt.nl/isa8/proceedings.html).

Spooren, W., Sanders, T., Huiskes, M., & Degand, L. (2010). Subjectivity and causality: A corpus study of spoken language. In S. Rice and J. Newman (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research* (pp. 241-255-14). CSLI/University of Chicago Press.