Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a postprint version which may differ from the publisher's version.

For additional information about this publication click this link.
https://hdl.handle.net/2066/215095

Please be advised that this information was generated on 2021-06-22 and may be subject to change.

The automatic analysis of subjectivity and causal coherence in text

Wilbert Spooren[1][0000-0002-2982-3970] and Ted Sanders[2]

[1] Centre for Language Studies, Radboud University Nijmegen
[2] UIL-OTS, Utrecht University
w.spooren@let.ru.nl

Understanding a text means making a coherent representation of the information in that text. Causal coherence place an important role in making that representation. Dutch has a rich repertoire of causal connectives to express such causal links, the so-called causal DRDs. Previous research has shown that causal DRDs have their own profile: Dutch omdat expresses mostly relatively objective relations, whereas want tends to express more subjective relations. The following examples demonstrate the point.

1. D De velden zijn nat omdat het veel geregend heeft deze week.
   E The fields are wet OMDAT it much rained has this week
   'The fields are wet because it has rained a lot last week.'
2. D De voetbalwedstrijden worden vast afgelast, want het heeft dezeweek erg veel geregend.
   E The soccer games become surely cancelled, WANT it has this week very much rained
   'Surely the soccer games will be cancelled, because it has rained a lot this week.'
3. D Jan kwam terug omdat hij van haar hield.
   E Jan came back OMDAT he from her loved.
   'Jan came back because he loved her.'
4. D Jan hield van haar, want hij kwam terug.
   E Jan loved from her, WANT he came back.
   'Jan loved her, because he came back.'

5. D Wat doe jij vanavond want er draait een goede film.
   E What do you tonight WANT there turns a good movie.
   'What are you doing tonight, because there's a good movie on.'

The differences between examples (1-5) have been described in terms of subjectivity [1]. Subjectivity can be defined as the degree to which the interpretation of an utterance requires that there is an active Subject of Consciousness who is responsible for the truth of the utterance. An utterance is subjective because there is some thinking entity in the discourse who evaluates. For example, the truth of an utterance such as *The height of the Eiffel Tower is 330 meters* can be evaluated directly in reality, and hence it is not subjective. By contrast, an utterance like *The Eiffel Tower is the greatest achievement of modern day architecture* requires the assumption that there is a Subject of Consciousness who is responsible for its truth.

Relations like (1), which Sweetser has termed content relations [2], can be described as objective: they report real world causality and do not assume the presence of a Subject of Consciousness. So-called epistemic relations like (2) and (4) are subjective because they present the outcome of an active reasoning process from the speaker or writer of the utterance. Similarly, speech act relations like (5) are subjective, because the Subject of Consciousness is motivating his or her performance of the speech act. Reason relations like (3) are in-between, because they do require the assumption of a Subject of Consciousness, but that is typically a character that is quoted in the text, whose reason for performing an action is reported.

Dutch has a preference to use *omdat* for more objective relations and *want* for more subjective relations, as in the examples above. The frequency with which *want* and *omdat* occur is very much genre-dependent: *want* is much more frequent in spontaneous conversations whereas *omdat* occurs more often in written newsreports and opinion pieces. At the same time, the subjectivity profile seems to be independent of genre: the difference in subjectivity between *want* and *omdat* is constant for each of the three genres that were investigated by Sanders and Spooren [1].

This type of findings is typically based on manual analyses of relatively small corpora. Such studies generally use a research design in which subsets of 100 instances of *omdat* and of *want* are compared in different genres (see for example, [3] for an analysis of forward causal DRDs in Dutch, and [4] for causality in Mandarin Chinese).

In this paper, we present a tool that makes use of state-of-the-art language technology to carry out such analyses automatically. The tool is the output of an ongoing project ACAD (Automatic Coherence Analysis of Dutch). For details on the project see https://www.clariah.nl/projecten/research-pilots/acad. The project aims at reaching three goals: (i) carry out these analyses automatically, thus preventing intercoder reliability issues; (ii) scale up the analyses by looking at many more instances and many more causal DRDs than is possible in manual analyses; (iii) look at many different genres.

The present study links to work done by Bestgen et al. [5], who used so-called thematic text analysis: the difference in subjectivity between, for example, *want* and *omdat* leads to the prediction that there are more subjective adjectives and adverbs in the segments that are connected by *want*, and more objective adjectives and adverbs in the segments connected by *omdat*. For our list of subjective and objective adjectives and adverbs we made use of the gold1000 list determined by De Smedt and Daelemans [6], who had participants rate the subjectivity of 1012 adjectives on a scale from 0 to 1. We identified those adjectives as subjective that had a score of 0.7 or higher for each of its meanings (650 adjectives, examples: *overweldigend* (‘overwhelming’), *afschuwelijk* (‘horrible’)), whereas objective adjectives had a score of 0.2 or lower (171 adjectives; examples: *visueel* (‘visual’), *zwart* (‘black’)).

The analysis goes through a number of steps: (i) identification of the relevant cases of causal DRDs; (ii) establishing the scope of the segments S_1 and S_2 that are connected by the DRDs; (iii) determining the direction of the causal link (backward, where the first segment expresses the consequent in the causal relation, as in examples (1-5), or forward); (iv) counting the number of subjective and objective adjectives and adverbs in the two segments; and (v) statistically testing the subjectivity hypothesis.

The current study made use of the corpora available in the Clariah environment: the SONAR corpus (a 500M words corpus containing 25 genres varying from newspaper texts, to wiki-pages, chat and texting, cf. [7]); the VU-DNC corpus (a 2M word corpus containing texts from

newspapers from the 1950s and from 2002; [8]); the Corpus of Spoken Dutch (CGN, [9]); and two newly added corpora: WhatsApp messages obtained in a recent study on the relationship between new media use by adolescents and young adults ([10]), and news texts from a Dutch quality newspaper published both on paper and online, matched for topics and genre (the NRC corpus; 1M words).

First results show indeed that the instrument is sensitive enough to detect the expected differences in the subjectivity of the environment of want and omdat. Thereotical implications and urgent next steps will be discussed. The discussion will be related to the the corpus build in DiscAn [11].

References

1. Sanders, T., Spooren, W.: Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat inter- actions and written text. Linguistics 53(1):53–92 (2015).
2. Sweetser, E.: From etymology to pragmatics. Cambridge, Cambridge University Press (1990).
3. Sanders, T.J.M., Stukker, N.:. Causal connectives in discourse: A cross-linguistic perspective. Journal of Pragmatics 44(2), 131-137 (2012).
4. Li, F., Sanders, T.J.M., Evers-Vermeul, J.: On the subjectivity of Mandarin reason connectives - Robust profiles or genre-sensitivity?. In Stukker, N., Spooren, W., Steen, G. (eds.) GENRE IN LANGUAGE, DISCOURSE AND COGNITION, pp. 15-49. De Gruyter Mouton, Berlin/New York (2016).
5. Bestgen, Y., Degand, L., Spooren, W.: Toward automatic determination of the semantics of connectives in large newspaper corpora. Discourse Processes 41(2), 175–193 (2006).
6. Smedt, T. D., Daelemans, W.: '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 (2012).
7. Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I.: The construction of a 500-million-word reference corpus of contemporary written Dutch. In ESSENTIAL SPEECH AND LANGUAGE TECHNOLOGY FOR DUTCH, pp. 219–247. Springer (2013).
8. Vis, K. (2011). Documentation of the VU Diachronic Newspaper texts Corpus. Unpublished ms., retrieved on Feb. 10, 2018 from http://tst-centrale.org/images/stories/producten/documentatie/vu-dnc_doc3-documentation-of-corpus.pdf
9. Oostdijk, N.: The Spoken Dutch Corpus Project. The ELRA Newsletter 5(2), 4–8 (2000).
10. Verheijen, L., Spooren, W., Kemenade, A. van, (submitted). The relationship between Dutch youths' social media use and school writing.
11. Sanders, T., Vis, K., Broeder, D.: Project notes of Clarin project DISCAN: Towards a discourse annotation system for Dutch language corpora. In EIGHTH JOINT ACL - ISO