Automatic coherence analysis of Dutch: Testing the subjectivity hypothesis on a larger scale

Jet Hoek (University of Cologne)
Ted J.M. Sanders (Utrecht University, Utrecht Institute of Linguistics OTS)
Wilbert P.M.S. Spooren (Radboud University Nijmegen)

Abstract
With the increasing availability of large corpora, quantitative corpus analysis is becoming more and more popular as a method for doing linguistic research. This paper uses a new research tool that makes it possible to search syntactically annotated corpora without extensive programming knowledge (CESAR) to study the subjectivity patterns of four Dutch causal connectives. Analyzing a large set of causal relations marked by four of the most frequent Dutch causal connectives (*daarom*, *dus*, *omdat*, and *want*), the case study aims to corroborate the subjectivity hypothesis established on the basis of smaller scale studies that used manual annotation. The automatic analysis of the subjectivity patterns of Dutch causal connectives illustrates the usability of CESAR in particular and the feasibility of automatic coherence analysis in general. In addition, it generates new insights into the subjectivity patterns of *daarom*, *dus*, *omdat*, and *want*.

1 Introduction

With the increasing availability of large corpora, quantitative corpus analysis is becoming more and more popular as a method for doing linguistic research. Annotated corpora provide even more opportunities for linguistic analyses than corpora that contain no meta-linguistic annotations. The Penn Discourse Treebank (Prasad et al., 2008) and the RST Treebank (Carlson, Okurowski, & Marcu, 2002) are examples of corpora that contain discourse-level annotations; these corpora have been invaluable resources for the study of discourse coherence in recent years. However, the road to doing large-scale coherence analysis can be steep. Creating a discourse-annotated corpus requires a lot of time and resources, since annotation largely has to be done manually. In addition, the resulting corpus is usually restricted to a single language (i.e., the language of the corpus data) and to the framework used in the annotation of the corpus data (e.g., Rhetorical Structure Theory [Mann & Thompson, 1988] for the RST Treebank), which means that a corpus may not be annotated for all the distinctions relevant to research questions posed by researchers working within another discourse annotation framework. As a result, the number of available large discourse-annotated corpora is limited, and for many languages, including Dutch, no extensive discourse-annotated corpora exist. Large syntactically annotated corpora, on the other hand, are more widely available.

In this paper, we make use of a new research tool that can search syntactically annotated corpora and allows for the automatic analysis of coherence relations in Dutch without extensive

programming knowledge: CESAR, a web-based search interface.[1] Using this tool, we conduct a large-scale corpus study to assess subjectivity of Dutch causal connectives across three genres. Mirroring manual discourse annotation, our automatic analysis consists of two main steps: segmentation and annotation. After providing a brief overview of CESAR and the automatic analyses it facilitates in Section 2, we describe our approach to the automatic segmentation of Dutch causal relations, segment a large set of causal relations (marked by four highly frequent Dutch causal connectives), and assess the quality of the segmentation in Section 3. We then automatically determine the subjectivity of the segmented relations (i.e., 'annotation') in Section 4 and compare the subjectivity patterns found in our analysis to subjectivity patterns established on the basis of manual annotations. Finally, we discuss the results of our automatic subjectivity analysis in relation to previous findings of smaller scale studies involving manual annotations and formulate recommendations for further development of automatic coherence analysis in Section 5. The aim of this paper is thus two-fold: to assess the feasibility of automatically segmenting and annotating coherence relations using a rule-based approach, and to assess the subjectivity patterns of four frequent Dutch causal connectives at a larger scale.

2 CESAR

To automatically analyze coherence relations in Dutch, we use CESAR, a web-based interface that enables linguistics researchers or students to conduct quantitative corpus searches without the programming skills such projects often require. CESAR was developed as part of the ACAD project (Automatic Coherence Analysis of Dutch, Clariah project CC 17-002) and consists of pre-programmed functions that can be used to search available corpora on the basis of syntactic annotations.[2] As such, automatic analyses in CESAR are rule-based. The general approach taken in the automatic detection and segmentation of coherence relations in this paper is laid out in Sections 3.1 and 3.2; the selected automatic annotation method, i.e., the way in which we determine the subjectivity of a coherence relation, is explained in Section 4.2. The search projects used in this paper are shared publicly on the CESAR interface.[3]

2.1 Corpora available in CESAR

Search projects in CESAR can be executed on several different corpora, ranging in language mode (written versus spoken), register (formal versus informal), spontaneity (edited versus spontaneous). All corpus data come with syntactic annotations: POS tags, syntactic parses, and lemma information.[4] An overview of the corpora used in the current study can be found in Table 1.

[1] https://cesar.science.ru.nl

[2] A detailed description of the specifications of the CESAR interface can be found in Komen and Hoek (*submitted*) and in the CESAR documentation available in the CESAR portal.

[3] The specific search projects used for this study have been made publicly available in the CESAR portal: Hoeketal2018_daarom for *daarom*, Hoeketal2018_dus for *dus*, Hoeketal2018_omdat for *omdat*, and Hoeketal2018_want for *want*.

[4] For some of the corpora, lemma information is limited.

Table 1
Overview of corpora used in the current study

Genre	Corpus	Reference	Number of words
Newspaper	VU DNC	Vis, Sanders, & Spooren (2012)	786,374
	NRC 2011	Spooren et al. (2018a)	962,097 ('hard' only)
Spoken	Corpus Gesproken Nederlands (Corpus of Spoken Dutch)	Oostdijk (2002)	999,576
Chat	WhatsApp corpus 'Lieke'	Verheijen & Stoop (2016); Spooren et al. (2018c)	354,744
	WhatsApp corpus 'Manon'	Spooren et al. (2018b)	WILBERT?

The syntactic annotations are crucial to the search project reported in this paper (see Sections 3 and 4). While the syntactic annotations are indispensable, the quality of any project in CESAR will in part be dependent on the quality of the POS tagging, syntactic parses, and lemmatization of the corpus texts. Whenever a search project outputs results that are unexpected on the basis of the input variables, it may therefore be due to a mistake in the syntactic annotation of the fragment. Because the syntactic annotations are sensitive to the quality of the corpus texts, the quality of the annotations may be lower for more spontaneous, unedited text (e.g., chat data).

In addition to being dependent on the quality of the syntactic annotations, which in most cases were automatically generated, researchers using the CESAR portal for quantitative linguistic analyses involving genre are dependent on the reliability of the genre-tagging. Genre tags have in most cases been generated by researchers during the compilation of the corpora.

3 Automatic segmentation of Dutch coherence relations

Any stretch of natural text, be it written or spoken, displays coherence; the elements of the texts are connected by so-called coherence relations, such as *Cause-consequence*, *Contrast* or *Temporal overlap.* A coherence relation typically consists of two segments, and the relation may or may not be explicitly marked. An example of a coherence relation is given in (1).

(1) De velden zijn nat omdat het veel geregend heeft deze week.
'The fields are wet because it has rained a lot this week.'

In this example we have a segment 1 (S_1) *De velden zijn nat* ("the fields are wet"), which expresses the consequence of a segment 2 (S_2), the cause *het veel geregend heeft deze week* ("it has rained a lot this week"); S_1 and S_2 are connected via the connective *omdat* ("because"). An automatic analysis of coherence relations requires the correct identification of the connective, and of S_1 and S_2.

The automatic approach to identifying connectives and their associated discourse segments reported in this paper was developed for all types of connectives, though the main focus has been on causal relations. In Section 4, we analyze the subjectivity of coherence relations signaled by the four most frequent Dutch causal connectives: *daarom* 'that is why,' *dus* 'so,' *omdat* 'because,' and *want* 'because/since.' Most research on the basis of manual annotations has also focused on these connectives, which allows us to compare the results from

our automatic analysis to results from previous, smaller-scale corpus studies. With the analysis in Section 4 focusing on *daarom*, *dus*, *omdat*, and *want*, we also assess the quality of the automatic segmentation for relations signaled by these four connectives (see Section 3.3).

3.1 Identifying causal connectives

The category of connectives includes words from different syntactic categories: subordinate conjunctions (e.g., *omdat* 'because'), coordinating conjunctions (e.g., *want* 'because'), prepositions (e.g., *om* 'to'), and adverbs (e.g., *immers* 'because after all'). This is why some words can be a connective in some contexts, but not in others; *om*, for instance, qualifies as a connective if it has a clausal complement, but not when its complement is an NP. In addition, we have to account for homographs of entries from our connective list, for instance *want* as an NP meaning 'glove,' or *als* 'if' as a comparative (*even groot als* 'as big as', *verkleed als* 'dressed up as'). In most cases, connectives cannot be reliably identified on the basis of the string search, but also require the POS information provided by the syntactic annotations of the corpora. Some connectives, however, do not need such a POS condition; *omdat*, for instance, is always connective, which is why a simple string search suffices.

3.2 Identifying the segments

The segments that the connectives relate to each other are identified on the basis of the parse tree. First, it is determined whether the connective is positioned before both segments (mrk-s1-s2), between the two segments (s1-mrk-s2), or, in case of adverbials, inside the second segment. The positioning of the connective is determined on the basis of the connective and its position in the sentence. Coordinating conjunctions such as *want* can only appear between S1 and S2, so determining its position in the relation can be done on the basis of just the connective. Subordinating conjunctions, such as *omdat*, however, can appear either before or between S1 and S2. In those cases, the position of the connective has to be determined on the basis of the parse tree. For subordinating conjunctions, it is calculated whether the connective is the first item in the sentence or whether it is preceded by other elements. If it is the first item in a sentence, the positioning of the connective is marked as mrk-s1-s2; if it is not, the order of the relation is determined to be s1-mrk-s2.

The discourse segments are then determined on the basis of the position of the connective. For s1-mrk-s2 relations, the S1 is extracted from the text preceding the connective and the S2 from the text following the connective. If "mrk" is sentence initial in s1-mrk-s2, S1 will be found in the preceding sentence. Segments are found on the basis of grammatical nodes indicative of a finite clause or a discourse unit, or on the basis of the sentence boundary. Everything between the node or sentence boundary and the connective is extracted for S1; the contents of the node following the connective or all text until the sentence boundary is extracted for S2. For mrk-s1-s2 relations, the contents of two nodes indicative of a finite clause or discourse unit following the connective are extracted for S1 and S2. In case the connective appears in the middle of S2, the contents of the node indicative of a finite clause or discourse unit under which the connective is situated is extracted as S2; the contents of the preceding node is extracted as S1.

While this baseline system for identifying connectives and their discourse segments works in most cases, exceptions have to be accounted for. Against prescriptive conventions,

omdat can be used between the segments with each segment in a different sentence, as in (2).[5] If only one is detected in the sentence headed by *omdat*, the preceding clause is taken as the s1 of the relation. Similarly, we have to account for the possibility of connectives being modified, as in (3) 'especially because,' to make sure that the modifier is not identified as S1 and to get the correct positioning of the connective within the relation in case of a modifier-connective-s1-s2 construction. Finally, embedded coherence relations, as in (4) "but because…" are tackled by specifying that if a subordinating connective, such as *omdat*, is preceded by a coordinating conjunction, such as *maar* 'but,' the two clauses following the subordinating connective have to be identified as the discourse segments.

(2) Maar er is in de voorbije jaren een grote verschuiving in de populariteit van de types geweest. **Omdat** [lang niet iedereen born to be wild blijkt.]$_{S1}$ (dpc-rou-000480-nl-sen.0004)
But in past years there has been a huge shift in the popularity of the types. Because not nearly everyone turned out to be born to be wild.

(3) En toch is de 21e eeuw niet de 19de, [vooral niet]$_{S1}$ **omdat** [de globalisering zowel handel als communicatie gebracht heeft.]$_{S2}$ (dpc-ind-001642-nl-sen.0049)
And yet the 21st century is not the 19th, especially because globalization brought about both commerce and communication.

(4) Dat is op zich geen beperking van het systeem, maar **omdat** [het loon meestal hoger is op het einde van de carrière dan in het begin, betekent dit in de realiteit dat …]$_{S2}$ (dpc-fsz-000551-nl-sen.0066)
That is not necessarily a limitation of the system, but because wages are usually higher toward the end of a career than in the beginning, it means that in reality …

3.3 Assessing the quality of the segmentation

We manually assessed the quality of the segmentation for the four most common Dutch causal connectives (*daarom*, *dus*, *omdat*, and *want*) across three genres (newspaper, spoken, and chat). We judged the quality of S1 on one set of relations and the quality of S2 on another to make sure the judgments were independent; it is highly likely that when S1 is wrongly identified, S2 is also incorrect, and vice versa. We checked two hundred segments per connective, per genre, per S1/S2, which totals to 4800 segments (200*4*3*2). All items were checked by Coder1 (first author). Coder2 (a trained Linguistics student) checked instances 1-100 for each set of two hundred; Coder3 (third author) checked instances 101-200. The inter-annotator agreement scores are given in Table 2. To obtain the overall agreement score and the agreement per genre, we treated Coder2 and Coder3 as a single coder. In addition, we report the agreement between Coder1 and Coder2 and between Coder1 and Coder3.

[5] The source of all corpus-based examples is given in parenthesis after the example; the fragment can be located by typing in the corpus code in the 'Browse' section of the CESAR portal.

Table 2

Inter-annotator agreement scores for the task of judging whether the segments of a relation have been correctly identified.

Dataset	Cohen's κ
Overall	.62
News	.70
Spoken	.54
Chat	.59
Coder1-Coder2	.61
Coder1-Coder3	.66

The inter-annotator agreement on the whole dataset is κ=.62 (Cohen's Kappa), which, especially considering the degree of subjectivity involved in discourse-annotation tasks and inter-annotator agreement range common for discourse annotation tasks (Spooren & Degand, 2010), we take to be satisfactory for the current purposes. Agreement is better for the relations taken from newspaper texts than for relations taken from spoken or chat corpora. This could be due to a number of reasons. First of all, a lot of corpus-based research on discourse coherence uses newspaper texts, so there may be more consensus on what constitutes a discourse segment in this genre than there is for spoken or chat discourse. Newspaper articles are also monologues and more edited than spoken or chat language, which tend to be dialogic and more spontaneous; as a result, there tend to be fewer discontinued segments and coherence relations between speakers in newspaper texts than in spoken or chat discourse. This makes the identification of discourse segments a more straightforward task in newspaper texts than in spoken or chat conversations. Finally, the differences in inter-annotator agreement scores between newspaper, spoken, and chat texts may in part be due to differences in the quality of the automatic segment identification. As can be seen in Table 3, the automatic text segmentation was best for newspaper texts. The dataset for this genre thus contained many straightforward correctly identified cases, leading to a better agreement than for the spoken or chat relations.

Table 3
Percentage of correctly identified segments

Dataset	%
Overall	77
News	88
Spoken	79
Chat	63
S1	70
S2	83
Daarom	77 (S1: 70 S2: 85)
Dus	61 (S1: 49 S2: 73)
Omdat	85 (S1: 87 S2: 82)
Want	84 (S1: 76 S2: 91)
Consequents	80

The percentage of correctly identified segments is given in Table 3. The percentages are based on the judgments of Coder1. Overall, 77% of the segments were identified correctly by our automatic approach. Segmentation was better for newspaper data than for spoken data, the segmentation of which was in turn better than for chat data. As already mentioned above, the automatic segmentation of the newspaper texts is most likely best because of the fact that they constitute a highly edited form of monologic discourse. While spoken data and chat data seem comparable in terms of spontaneity and the amount of dialogue, the spoken data (taken from the CGN, Oostdijk 2002) has been professionally transcribed, while the chat data has been left as is (Verheijen & Stoop 2016). This seems crucial for the quality of the syntactic annotations (see also Section 3.4), on which identification of the segments depend; syntactic parsing, POS tagging, and lemmatization are all dependent on words being spelled correctly/conventionally.

S2 was more often correctly identified than S1. This may in large part be due to the fact that the connective is always adjacent to S2, but not necessarily to S1. Especially for dialogic discourse, S1 may not be the clause immediately preceding the connective, since there may be an intervening utterance by another speaker, see examples (11) and (12) in Section 3.4.2.

The automatic segmentation method could less reliably identify segments related to each other by *dus* than segments connected by *daarom*, *omdat*, or *want*. On closer inspection, it is especially the S1 for *dus* that is often not identified correctly (49%). The identification of S1 for *dus* relations is also heavily influenced by genre: 68% correct for news, 46% for spoken, and 31% for chat. The main source of segmentation errors with *dus* seems to be the non-adjacency of discourse segments, with intervening linguistic material from either the same speaker or a different speaker, see also Section 3.4.2, examples (11) and (12).

Finally, we report the percentage of correctly identified segments that express the consequent of the causal relation, since these segments will be the focus of the analysis in Section 4. For *daarom* and *dus*, S2 expresses the consequent; for *omdat* and *want*, S1. Overall, slightly more segments expressing consequent were identified correctly than segments expressing antecedents (80% versus 74%), which seems in large part due to the bad quality of the S1s for *dus*.

While the quality of the segmentation thus differs between subsets of the data, the overall quality of the automatic segmentation approach we developed using the CESAR tool seems reasonable. The quality could be further improved by fine-tuning the current general segmentation approach to a specific connective or to a specific genre. On the other hand, we identified several segmentation problems that do not seem to have an easy fix and may be permanent sources of mistakes in the current approach to automatically segmenting coherence relations; these are discussed in the next section.

3.4 Problems for the current automatic discourse analysis approach

In assessing the segmentation quality of our automatic discourse analysis, we encountered many instances of connective identification or segmentation mistakes that seem difficult to fix under the current approach. These include mistakes due to errors in the syntactic annotation and mistakes due to discourse structural ambiguity.

3.4.1 Mistakes in the syntactic annotations

As was already mentioned in Section 2.3, the search project relies heavily on the quality of the syntactic annotations. Many problems with identifying connective uses of the search tokens that do not have to do with the identification of the segments seem to be caused by mistakes in the POS tagging or the syntax tree. An example of this is reliably distinguishing between connective uses of *daarom* (also *daardoor*, *hierom*, and *hierdoor*), as in (5), and cases in which *daar* (or *hier*) is used anaphorically and has been merged with a preposition stemming from the verb, as in (6).

(5) De verhaallijn is op zich wel sterk, en de volgende twee seizoenen, die niet meer zijn opgenomen zijn **daarom** als Avatar 1 & 2 in boekvorm verschenen. (WR-P-E-I-0000027197)
'The plot is fairly good, and the next two seasons, which have not been recorded, have ***therefore*** *been published as books as Avatar 1 & 2.'*

(6) De Nederlanders die niet gewend waren een legitimatiebewijs bij zich te dragen en te tonen als **daarom** gevraagd werd, voelden zich diep gekwetst en in hun vrijheid aangetast. (WR-P-E-I-0000050381)
'Dutch citizens who were not used to carrying an ID and showing it if asked ***for it****, felt deeply hurt and compromised in their freedom.'*

In cases where this problem arises, it seems that the issue can be attributed to the parse tree. In some cases, the second type of *daarom* is labeled as 'BW-PC' (adverbial prepositional object), in which case they can be filtered out, but in both (5) and (6) *daarom* is labeled as 'BW-MOD' (adverbial modifier), which is correct in (5) but incorrect in (6). In addition, the parse trees present a highly similar structure.

Other examples where the segments were incorrectly identified can be found in (7) and (8). In both examples, the mistakes in the automatic discourse analysis seem to be caused by mistakes in the syntactic annotations.

(7) Philips kon hier rekenen op de morele steun van de overheid omdat die er "y voyait un cours supplémentaire de physique et d'hygiène." (WR-P-E-I-0000049645)
In this, Philips had the government's moral support because it "y voyait un cours supplémentaire de physique et d'hygiène."

(8) De aspirine voorkomt de werking van Cyclooxygenase en voorkomt daarmee de vorming van prostaglandine, waardoor een groot gedeelte van de pijn verdwijnt, en ook de koorts en de ontsteking geremd worden **omdat** dat de prostaglandine deze reacties niet meer kan veroorzaken. (WR-P-E-I-0000000014)
The aspirine blocks the Cyclooxygenase and in doing so prevents prostaglandine from being formed, which is why a lot of the pain disappear and the fever and infection are reduced, because that prostaglandine can no longer cause these reactions.

In (7), there is a problem with the syntactic parsing that appears to arise from the fact that half the sentence is in French. In (8) the problem can be plausibly attributed to the typo right after *omdat* (*dat* 'that' instead of *dan* 'then').

The automatic identification of connectives and their segments using the CESAR interface is thus dependent on the quality of the syntactic annotations in the available corpora. Mistakes in the syntactic parsing, POS tagging, or lemmatization, either reoccurring, as in (6), or novel, as in (7) and (8), may all lead to a reduced quality of the discourse analysis.

3.4.2 Segmentation and the meaning of a fragment

Other common problems for the automatic segmentation approach used in this paper, as well as for automatic segmentation approaches in general, are discourse-structural ambiguities that can only be resolved using the interpretation of the fragment. Especially prone to discourse-structural ambiguities are fragments that contain embedded clauses (Hoek, Evers-Vermeul, & Sanders, 2017). This problem is illustrated by (9) and (10). In (9), the *omdat*-clause provides a reason for why company Inovara claims that they have been charged too much for the use of 'wijkcentrales,' central places where internet and phone cables enter a neighborhood. As such, everything before *omdat* should be included in S1. In (10), on the other hand, the *omdat*-clause provides a reason for Morsi not taking sufficient measures; this entire causal relation forms the complement of the matrix verb *verwijten* 'accuse.'

(9) Inovara stelt dat KPN hiervoor een te hoog bedrag heeft gerekend omdat het bedrijf van minder wijkcentrales gebruik heeft gemaakt dan aanvankelijk was gepland. (BAec2)
Inovara claims that phone company KPN has charged too much for this because Inovara made use of fewer 'wijkcentrales' *than initially planned.*

(10) Het leger verweet hem dat hij [Morsi] niet hard genoeg optrad tegen de extremisten in de Sinaï omdat hij hen als bondgenoot zag. (NRC_Handelsblad_egypte066)
The army made the accusation that he [Morsi] did not take sufficient measures against the extremists in the Sinai because he say them as his allies.

While (9) and (10) thus have a distinct hierarchical discourse structure and should be segmented differently, their surface structure is virtually identical: a complement-taking

predicate, a clause, *omdat*, and another clause. The relations in (9) and (10) can be segmented correctly by taking into account the meaning of the sentences (Hoek, Evers-Vermeul, & Sanders, 2017), but this is not something an automatic segmentation approach can do. Indeed, in our dataset, (9) and (10) have been segmented in the same way; in both fragments, the most local relation has been segmented (i.e., the first segments do not contain the complement-taking predicates *Inovara stelt dat* and *Het leger verweet him dat*). This is the correct segmentation option for (10), but not for (9).

In addition, coherence relations can hold between single sentences or clauses, but they can also hold between larger text spans, in which case segments consist of multiple sentences. This presents a segmentation problem similar to the problem in (9)-(10); human coders can determine which parts of the text are involved in a specific coherence relation, but this is much more difficult in automatic segmentation approaches.

Finally, coherence relations in a dialogue may hold between two segments uttered by the same speaker, but speakers can also relate their own utterance to an utterance from another speaker, as in (11), where speaker B draws a conclusion on the basis of the information provided by speaker A. In addition, there may be intervening utterances between the segments of a coherence relation, as in (12).

(11) A: Ik heb ’t maar meegenomen.
B: Dus je hebt er twee meegenomen. (fn000595)
A: I decided to take it. B: So you took two.

(12) A: Ik hoefde ook niet hier te eten.
B: Nee.
A: Dus dat scheelt. (fn008058)
A: I also did not have to eat here. B: No. A: So that is a plus.

The example in (11) was segmented correctly by our segmentation approach, but in (12), the segments were indicated to be not the two utterances from speaker A, but between the utterance from speaker B and the second utterance from speaker A.[6]

Ambiguity in discourse structure thus appears to pose a substantial problem for automatic discourse segmentation. In our automatic segmentation approach, we opted to always segment the most local relation, which for fragments such as (10) or (12) leads to segmentation options that would most likely not be produced by human coders. In evaluating the quality of our automatic segmentation approach (see Section 3.3), we were lenient toward mistakes that were due to discourse structural ambiguity, counting partly correct cases, for instance (9), as correct, but counting completely incorrect cases, for instance (12), as incorrect.

As was explained in Section 2, the automatic segmentation performed in the CESAR interface works on the basis of functions, making it a rule-based approach to discourse segmentation. While this method is accessible to linguists without extensive programming skills and does not require any discourse-annotated input, it may not be the approach best

[6] We allowed *ja* ‘yes’ and *nee* ‘no’ (and variations) as discourse segments, since they imply full propositions, e.g., answering ‘yes’ to the question ‘Will you come to my party’ implies ‘I will come to your party.’ Example: *ja want dat had ik de tweede dag* ‘yes because I had that on the second day” (fn007109).

equipped to tackle an intricate problem such as ambiguity in discourse structure; for this, a statistical learning approach (or a combination of rule based and statistical learning) seems more promising (e.g., Muller, Afantenos, Denis, & Asher, 2012). However, the frequency of the segmentation problems described above is not very high. For our current purposes, the restrictions of a rule-based approach are therefore far outweighed by its benefits: it is accessible without programming skills with the CESAR interface, it does not require a large set of discourse-annotated training data, and it is flexible and easily adaptable to other purposes and research topics.

4 Automatic subjectivity analysis of Dutch coherence relations

Manual analyses that have investigated coherence relations, including those that have focused on the subjectivity profiles of connectives, typically make use of relatively small samples of corpora. For example, Sanders and Spooren (2015) used 100 instances of each connective (*omdat, want*) in each of the three genres under investigation (newspaper articles, spontaneous conversations and chat interactions).[7] The CESAR tool allows us to scale up that analysis, provided that we have a proxy for the subjectivity features of causal segments used in manual analyses. We will first outline the subjectivity hypothesis of Dutch causal connectives in Section 4.1, after which we automatically analyze the subjectivity of a large number of Dutch causal relations in three different genres and compare the results to the subjectivity hypothesis in Section 4.2.

4.1 The subjectivity hypothesis

Coherence relations differ in their *source of coherence*; they can be either *objective* or *subjective* (e.g., Sanders, Spooren, & Noordman, 1992; Sweetser, 1990). In text-linguistic and cognitive linguistic work on causal connectives, an utterance is considered to be subjective when its interpretation requires an active *Subject of Consciousness* (SoC: Pander Maat & Sanders, 2000). An SoC crucially involves an animate subject, typically a person, whose intentionality is conceptualized as the ultimate source of reasoning, evaluating, describing or acting 'in the real-world.' An utterance is subjective when there is some thinking entity in the discourse who evaluates or concludes. For instance, *He thought California was nice* is subjective because it involves an evaluation by a character in the discourse. Compare this with an utterance like *California is in the USA*, which is presented as a fact in the world that does not depend on the evaluation by an SoC. To be more precise, in the utterance *He thought California was nice*, the validity of the proposition "California is nice" depends on the SoC *He*, whereas in the utterance *California is in the USA* the proposition "California is in the USA" can be verified directly in the non-linguistic reality.

Of course, each utterance in a discourse comes from a speaker or author, and therefore each utterance is dependent on an SoC. However, in some utterances, the SoC is manifest because the sequence cannot be interpreted without reference to an SoC. Such cases – typically feelings, conclusions, or evaluations of all kinds – are considered subjective; they simply

[7] In case of chat conversations, the corpus used by Sanders and Spooren (2015) did not contain 100 instances of *omdat*, which is why the analysis was based on only 51 instances of *omdat*.

cannot be interpreted without making reference to the SoC's thoughts and feelings. In contrast, utterances that do not depend on such a manifest reference to the SoC are considered objective.

Causality can be expressed using backward or forward causal connectives. In a forward causal construction, the first segment introduces a cause or an argument, and the second segment expresses a consequence or a claim. In backward constructions, the first segment expresses a claim or a consequence, and the second segment expresses the argument or the cause. In backward constructions, the connective typically occurs at the beginning of the second segment. In Dutch, the backward connective *want* 'for/since' is typically used to express subjective relations, whereas the backward connective *omdat* 'because' is typically used to express objective relations. Several studies have shown that these characteristics are robust and vary from strong preferences to clear restrictions on the relations they can express. Taken together, these observations show how the Dutch language "cuts up" backward causality (Degand, 2001; Degand & Pander Maat, 2003; Pit, 2006).[8]

In the Dutch lexicon of causal connectives, the same division of labour in terms of subjective and objective relations accounts for the most frequent connectives expressing forward causality: *daarom* (that's why) and *dus* (So/therefore), as in examples like *It has rained a lot this week. Daarom / That's why the fields are wet* versus *It has rained a lot this week. Therefore/So all soccer games will surely be cancelled*. In forward relations the subjective 'conclusion'-relations are more often expressed by the connective *dus* 'so', while the connective *daarom* 'that's why' has a preference for objective relations (Pander Maat & Sanders, 2000; Pander Maat & Sanders, 2001; Stukker et al., 2008; Stukker et al., 2009).

Sanders & Spooren (2015) have also argued in favour of the subjectivity hypothesis to explain the systematic differences between causal connectives. They report a corpus study on the meaning and use of the Dutch connective pair *want* and *omdat*, using an integrative empirical approach in which the complex construct of subjectivity was decomposed into several characteristics, which were analyzed separately. A corpus of *omdat* and *want* relations from written, spoken, and chat discourse was manually analyzed by two annotators. Subjectivity was expected to go across the modalities of written, spoken and chat language. The main hypothesis was that *want* occurs in more subjective contexts than *omdat*, irrespective of the genre. They formulated specific hypotheses on the way in which the connectives *want* and *omdat* would show differences in terms of subjectivity and all four were corroborated, across all media. The two of them most relevant for the current paper are:

- *Want* is used more often to express subjective relations than *omdat*
- *Want* is used more often to support a judgment than *omdat*

[8] It should be noted that while there are prototypical examples of objective and subjective coherence relations, classifying naturally occurring coherence relations as either objective or subjective, as would happen in manual corpus annotation, can be very difficult (e.g., Spooren & Degand 2010).

Prototypical cases from their corpus include fragment (13), from the spoken corpus, which illustrates a judgment in S1in a *want*-connection.

(13) [dat is gewoon krankzinnig.]$_{S1}$ want [als hij uhm mensen goed inschat moet ie ook weten dat ik m'n uiterste best doen om dat zo snel mogelijk voor elkaar te krijgen.]$_{S2}$
that is simply insane WANT if he uhm is such a good judge of character then he should also know that I am doing my very best to take care of that as soon as possible

The fragment in (14) shows the typical *omdat* pattern: the propositional attitude in S1 is something other than judgment.

(14) [Drie vrouwen van middelbare leeftijd worden achterna gezeten]$_{S1}$ omdat [ze het waagden te protesteren]$_{S2}$
Three middle-aged women are chased OMDAT they dared to protest.

Although Sanders and Spooren (2015) find differences in the degree of subjectivity of the three genres they investigate (with the least number of subjective relations found in written discourse and most subjective relations in chat), *want* is more subjective than *omdat* in all three genres. They thus conclude that the subjectivity hypothesis is supported by their data and that the subjectivity patterns of connectives appear to be robust across genres.

4.2 Automatically testing the subjectivity hypothesis

We used the CESAR portal and our automatically identified and segmented causal coherence relations to analyze the subjectivity patterns of four of the most frequent Dutch causal connectives, *daarom*, *dus*, *omdat*, and *want*, on a large scale. For each relation, we determined which segment was the consequent (Section 4.2.1) and whether or not the consequent was subjective (Section 4.2.2). To make the relations maximally comparable, we only included relations in which the connective was positioned between the two segments (or in the middle of S2) in our dataset; out of the four connectives, only *omdat* can be positioned either between the segments or at the head of the relation. We aim to replicate the following findings from the literature:

1. *Want* is more often used in subjective relations than *omdat*
2. *Dus* is more often used in subjective relations than *daarom*
3. The subjectivity patterns of the causal connectives are constant across genres
4. Formal (edited) written data contains fewer subjective relations than spoken data, which in turn contains fewer subjective relations than spontaneously and interactively produced written data, for instance chat or WhatsApp conversations.

4.2.1 Determining the direction of the causality and identifying the consequent

In some cases, the direction of the causality (forward: S1 presents the antecedent, S2 presents the consequent) is straightforward. *Want* always signals a backward causality and *dus* and *daarom* are always forward. However, in case of subordinating conjunctions the directionality depends on the position of the connective: does it occur in initial position (*omdat S1, S2*) or in

medial position (*S1 omdat S2*)? In some cases, most frequently in informal genres, the *omdat*-clause is presented as an independent utterance, in which case the S1 is the utterance preceding the current utterance, which is why the directionality of *omdat* cannot be reliably determined by only checking whether *omdat* occurs sentence-initially. In the current project, we also take into account the number of clauses in the sentence headed by *omdat*; if there is only one clause, S1 is taken from the previous sentence and *omdat* is determined to be positioned between the two segments.

4.2.2 Determining the subjectivity of the consequent

In this study, we use the occurrence of subjective words as a proxy for subjectivity. A common characteristic of subjective causal relations is that they include subjective words in their segments, specifically in their consequents (*John is a terrible person*). We exploit this characteristic to automatically determine whether a causal relation is objective or subjective; subjective if the relation includes one or more subjective words, objective if it does not.

We use an existing subjectivity lexicon to define which adverbs and adjectives qualify as subjective: the gold1000 list which was established by De Smedt & Daelemans (2012) by having participants rate the subjectivity of 1012 adjectives on a scale from 0 to 1. As in Spooren and Hendrickx (2015), who successfully applied the gold1000 list to automatically classify coherence relations as subjective or objective, we categorized entries that had a score of 0.7 or higher for each of its meanings as subjective.[9] This approach to automatically determining the subjectivity of coherence relations is similar to the approach taken by Bestgen, Degand, and Spooren (2006), who used a so-called thematic analysis to investigate the subjectivity of different connectives automatically, departing from the assumption that subjective words are more likely to occur in the context of a subjective connective than that of a more objective connective.

In terms of the CESAR search interface, the set of subjective adjectives and adverbs are included as global variables, specified in the 'fixed' section of the interface (Figure 1). Global variables are typically independent of the specific search being carried out. The adjectival and adverbial use of these words was established using the POS tags available in the corpora. Whenever a word from the consequent of a coherence relation carried a POS tag corresponding to adjective or adverb and matched one of the subjective words from the subjectivity lexicon, the relation was classified as subjective; if no adverbs or adjectives from the consequent of a relation matched the entries from the lexicon, the relation was marked as objective.

A potential disadvantage of using the gold1000 subjectivity lexicon is that it contains many fairly formal words (such as *geestdriftig* 'impassioned,' *magistraal* 'masterful,' or *ondoorgrondelijk* 'inscrutable'). In order to validate the analysis, we also looked at another measure of subjectivity, namely whether or not the segment contains a verb of cognition or modal verb (e.g., Biber & Conrad, 2009). These are words such as *say, claim, feel, can,* and *must*, which express a mental activity of the speaker or indicate the speaker's attitude towards the content of the utterance. This approach seems more robust against register influence, as the verbs are relatively high frequent, and can occur in any genre. For each segment it was

[9] 650 subjective adjectives in total. For example: *overweldigend* ('overwhelming'), *afschuwelijk* ('horrible').

established whether it contained one or more verbs of cognition/modal verbs (similar to the subjective adjectives and adverbs, we included a list of verbs of cognition/modal verbs as global variables in our search project in CESAR). If it did, the relation was determined to be subjective. If not, it was marked objective.

To find the causal connectives in newspaper data we made use of the SoNaR newspaper corpus. This corpus contains a large number of texts, which would require a very long search time. We therefore made use of a feature of CESAR that allows researchers to restrict the search to a specific number of texts that are randomly chosen from the corpus. Here we set the search to 10,000 newspaper texts.

We analyzed the results using binary logistic regression (logit) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2016, version 3.2.4), modelling the outcome of *want* versus *omdat* or *dus* versus *daarom*. We determined the significance of fixed effects by performing likelihood ratio tests to compare the fit of the model to that of a model without the fixed effect. In case of fixed effects with more than two levels (genre), we obtained pairwise comparisons by performing Tukey tests using the multcomp package (Hothorn, Bretz, & Westfall 2008). We carried out the analyses separately for backward connectives (*omdat* and *want*, which have the consequent of the causal relation in S1) and forward connectives (*daarom* and *dus*), which have the consequent in S2. To assess Hypothesis 4, we also performed an analysis with the forward and backward causal relations grouped, in which we model the proportion of subjective versus non-subjective consequents per genre.

4.3 Results

This section will first present the results of the subjectivity analysis on the basis of the presence of subjective adverbs and adjectives (Section 4.3.1), after which it will present the results of the analysis using the presence of modal verbs and verbs of cognition (Section 4.3.2).

4.3.1 Subjective adverbs and adjectives

Backward causality

Table 4 gives an overview of the occurrence of *omdat* and *want* in relations with subjective and non-subjective consequents in our corpus search, per genre. Consequents are considered subjective if they contain at least one subjective adjective or adverb; non-subjective if they do not.

Table 4

Raw count and percentages (per row) of omdat *and* want *in relations with consequents containing subjective adjectives or adverbs (subj) and in relations without (non-subj), per genre.*

	subj		non-subj		Total
	omdat	*want*	*omdat*	*want*	
Newspaper	147 (12.7%)	106 (9.1%)	670 (57.8%)	236 (20.4%)	1159
Spoken	124 (5.7%)	491 (22.7%)	391 (18.0%)	1161 (53.6%)	2167
WhatsApp	47 (4.7)	162 (16.2%)	216 (21.6%)	574 (57.5%)	999
Total	318	759	1277	1971	4325

The logit analysis shows that the data are best described with a model containing main effects of subjectivity and genre (the interaction subjectivity*genre was only marginally significant at p=0.0502). The main effect of subjectivity indicates that *want* has more consequents with a subjective adjective/adverb than *omdat* (B=0.41, z=4.96, p<.001). For genre, Tukey comparisons show that *want* is less often used in newspaper data than in spoken (B=2.02, z=24.66, p<.001) and WhatsApp data (B=1.92, z=19.77, p<.001). There is no difference in the frequency of *want* and *omdat* between spoken and WhatsApp data (B=-0.11, z=1.22, p=0.44).

The analysis shows that *want* is more often used in relations with subjective consequent segments than *omdat*. The analysis also shows that this difference is not dependent on genre. The results are thus in line with the subjectivity hypothesis. We also found that *want* occurred more often in spoken and WhatsApp data than in newspaper data, which is what hypothesis 4 predicts.

Forward causality

An overview of the occurrence of *daarom* and *dus* in relations with subjective and non-subjective consequents in our corpus search, per genre is given in Table 5.

Table 5

Raw counts and percentages (per row) of omdat *and* want *in relations with consequents containing subjective adjectives or adverbs (subj) and in relations without (non-subj), per genre.*

	subj		non-subj		Total
	daarom	*dus*	*daarom*	*dus*	
Newspaper	85 (7.6%)	174 (15.6%)	291 (26.1%)	566 (50.7%)	1116
Spoken	41 (0.7%)	1053 (16.9%)	239 (3.8%)	4882 (78.6%)	6215
WhatsApp	22 (0.9%)	390 (15.9%)	121 (4.9%)	1920 (78.3%)	2453
Total	148	1617	651	7368	9784

The logit analysis resulted in a model containing a main effect of genre (the interaction subjectivity*genre was not significant at p=.74; the main effect of subjectivity was not significant at p=.21). For genre, Tukey comparisons show that *daarom* is used more often in newspaper data than in spoken (B=2.38, z=27.00, p<.001) and WhatsApp data (B=2.11, z=19.68, p<.001). In addition, *daarom* is used more often in WhatsApp data than in spoken data (B=0.27, z=2.57, p<.05).

These results are not entirely in line with the subjectivity hypothesis, since we do not find that *dus* occurs more often in relations with subjective consequents than *daarom* (although the pattern we do find is consistent across genres). While hypothesis 4 does predict that *daarom* is more frequent in newspaper data than in spoken or WhatsApp data, the finding that *daarom* is more frequent in WhatsApp data than in spoken data goes against the predictions of this hypothesis.

Additional analysis: proportion of subjective relations per genre

Table 6 shows the proportion of subjective and non-subjective consequents per genre, over the data for the backward and forward connectives combined.

Table 6

Raw counts and percentages (per row) of consequents containing subjective adjectives or adverbs (subj) and in relations without (non-subj), per genre.

	subj	non-subj	Total
Newspaper	512 (22.5%)	1763 (77.5%)	2275
Spoken	1709 (20.4%)	6673 (79.6%)	8382
WhatsApp	621 (18.0%)	2831 (82.0%)	3452
Total	2842	11267	14109

The logit analysis shows a main effect of genre ($p<.001$). Tukey comparisons reveal that there are fewer relations with subjective consequents in WhatsApp data than in spoken ($B=-0.15$, $z=-2.98$, $p<.01$) or newspaper data ($B=-.28$, $z=-4.19$, $p<.001$). The difference in the number of relations with subjective consequents between spoken and newspaper data is only marginally significant ($B=0.13$, $z=2.20$, $p=.07$). These results are only partly in line with the subjectivity hypothesis.

Discussion subjective adverbs and adjectives

The results are only partly in line with the subjectivity hypothesis. While we do find that *want* is used more often in subjective relations than *omdat*, we do not find a difference between *daarom* and *dus*. Moreover, we find an effect of genre, in that the newspaper corpus has the most subjective segments. Given that we expected the newspaper corpus to be less subjective than spoken and WhatsApp data, this is a surprising result. This could, however, be due to the nature of the subjectivity lexicon we used. As mentioned in Section 4.2, the lexicon contains relatively many formal words, but fewer informal, colloquial, or slang words. It could therefore very well be the case that the spoken and WhatsApp data contain many adverbs or adjectives that would be judged as overtly subjective by human coders, but that are not included in the subjectivity lexicon we used. Indeed, a quick run-through of the spoken corpus reveals examples such as the ones in (15) and (16), which contain words that are not included in the subjectivity lexicon but that are clearly subjective nonetheless.

(15) [ik ben een beetje <u>chagrijnig</u> op rogier]$_{S1}$ want [hij is iets aan t ophangen en t hangt scheef]$_{S2}$ (fn000646)
I am a little <u>grumpy</u> at Rogier WANT he is hanging something and it is crooked

(16) [en wat nu opvalt is dat t zo <u>brandschoon</u> is hier ondanks die apparaten om me heen]$_{S1}$ want [dat zijn apparaten waar toch heleboel smeer en olie aan te pas komt]$_{S2}$ (fn007245)
what is remarkable is that it is so <u>sparkling clean</u> here despite these machines around me WANT they are machines that involve a whole lot of grease and oil

We therefore also assess the subjectivity of *omdat*, *want*, *daarom*, and *dus* across genres using another measure of subjectivity: the presence of a verb of cognition or a modal verb in the consequent of a causal relation.

4.3.2 Verbs of cognition and modality

Backward causality

Table 7 gives an overview of *omdat* and *want* relations with and without subjective consequents in our corpus search, per genre. Now, consequents are considered subjective if they contain at least one modal verb or verb of cognition, and non-subjective if they do not.

Table 7

Raw counts and percentages (per row) of omdat *and* want *in relations with consequents containing modal verbs/verbs of cognition (modal/voC) and in relations without (no modal/voC), per genre.*

	modal/voC		no modal/voC		Total
	omdat	*want*	*omdat*	want	
Newspaper	148 (12.8%)	122 (10.5%)	669 (57.7%)	220 (19.0%)	1159
Spoken	137 (6.3%)	659 (30.4%)	378 (17.4%)	993 (45.8%)	2167
WhatsApp	75 (7.5%)	277 (27.7%)	188 (18.8%)	459 (45.9%)	999
Total	360	1058	1235	1672	4325

The analysis resulted in a model containing main effects of connective and genre (the interaction subjectivity*genre was only marginally significant at p=0.054). As for connective, *want* was used more often in consequents with a modal verb/verb of cognition than *omdat* (B=0.65, z=8.33, p<.001), in line with the subjectivity hypothesis. The main effect of genre is the same as in the analysis with subjective adjectives/adverbs: *want* occurs less often in newspaper data than in spoken (B=-1.99, z=-24.11, p<.001) or WhatsApp data (B=-1.86, z=-19.11, p<.001); there is no difference between spoken and WhatsApp data (B=-0.13, z=-1.46, p=.31).[10] These results are mostly in line with our predictions: *want* is more often used in subjective relations than *omdat*, and this is not influenced by genre.

Forward causality

An overview of consequents with and without modal verbs or verbs of cognition in *daarom* and *dus* relations in our corpus search is presented in Table 8.

Table 8

Raw counts and percentages (per row) of daarom *and* dus *in relations with consequents containing modal verbs/verbs of cognition (modal/voC) and in relations without (no modal/voC), per genre.*

	modal/voC		no modal/voC		Total
	daarom	*dus*	*daarom*	dus	

[10] Because the proportion of *want* and *omdat* in each genre is the same in this analysis as in the analysis accompanying Table 4, the effect of genre is expected to be the same. However, this need not necessarily be the case, since the effect of genre in the two analyses is modeled alongside different main effects: the presence of subjective adjectives/adverbs for Tables 3 and 4, and the presence of modal verbs/verbs of cognition for Tables 7 and 8. The same holds for the main effect of genre for the analyses of the data in Tables 5 and 9.

Newspaper	80 (7.2%)	171 (15.3%)	296 (26.5%)	569 (51.0%)	1116
Spoken	61 (1.0%)	1369 (22.0%)	219 (3.5%)	4566 (73.5%)	6215
WhatsApp	26 (1.1%)	597 (24.3%)	117 (4.8%)	1713 (69.8%)	2453
Total	167	2137	605	6848	9784

The logit analysis resulted in a model containing only a main effect of genre. The interaction subjective*genre was not significant at p=.32; the main effect of connective was only marginally significant at p=.09. Tukey comparisons revealed that the main effect of genre is due to a higher proportion of *daarom* in newspaper data than in spoken (B=2.38, z=26.99, p<.001) or WhatsApp data (B=2.10, z=19.65, p<.001), in line with the subjectivity hypothesis. In contrast with the subjectivity hypothesis –though in line with the analysis using subjective adjectives/adverbs– there was also a higher proportion of *daarom* in spoken data than in WhatsApp data (B=0.28, z=2.61, p<.05).

As was the case with the adverb and adjective analysis (see Section 4.2.3), these results are not in line with the subjectivity hypothesis, since we did not find a difference in subjectivity between *daarom* and *dus* (although again the pattern we did find was consistent across genres).

Additional analysis: proportion of subjective relations per genre

Table 9 shows the proportion of subjective and non-subjective consequents per genre, over the data for the backward and forward connectives combined. Remember that subjectivity is operationalized differently than in Section 4.3.1: consequents are considered subjective if they contain at least one modal verb or verb of cognition, and non-subjective if they do not.

Table 9

Raw counts and percentages (per row) of consequents containing subjective adjectives or adverbs (modal/voC) and in relations without (no modal/voCj), per genre.

	modal/voC	no modal/voC	Total
Newspaper	521 (22.9%)	1754 (77.1%)	2275
Spoken	2226 (26.6%)	6156 (73.4%)	8382
WhatsApp	975 (28.2%)	2477 (71.8%)	3452
Total	3722	10387	14109

The logit analysis shows a main effect of genre (p<.001). Tukey comparisons reveal that there are fewer relations with subjective consequents in newspaper data than in spoken (B=-0.20, z=-3.5, p<.01) or Whatsapp data (B=-.28, z=-4.50, p<.001). The difference in the number of relations with subjective consequents between spoken and Whatsapp data is not significant (B=0.08, z=1.88, p=.14). These results are only partly in line with the subjectivity hypothesis.

5 General discussion and conclusion

In this paper, we used the CESAR web interface to automatically analyze coherence relations. Automatically analyzing coherence relations makes it possible to consider much more data than can be feasibly analyzed using manual annotation. Using a rule-based approach, we identified coherence relations on the basis of the presence of a connective, after which we identified the segments of each relation using a set of functions from the CESAR portal. The

overall quality of the segmentation was decent with 77% of all segments identified correctly, although segmentation quality was found to differ considerably between connectives and genres. A closer inspection of the segmented relations revealed that many of the segmentation errors were due to mistakes in the syntactic annotations upon which our automatic analysis depends. Most other segmentation problems were due to ambiguities in discourse structure. This latter problem may be difficult to solve using a rule-based approach; future discourse segmentation endeavors may therefore benefit from implementing statistical learning. Still, the quality of segmentation reached, especially on the more edited genre of newspaper texts, demonstrates that a rule-based approach to segmenting coherence relations can function as a promising starting point for the automatic analysis of coherence relations. In addition, our automatic segmentation method is publicly available in CESAR and can be directly used by other researchers interested in quantitatively analyzing Dutch discourse.

Departing from the segmented coherence relations, we then analyzed the subjectivity patterns of four frequent Dutch causal connectives, *want*, *omdat*, *daarom*, and *dus*, in three different genres. In manual annotation, the subjectivity of coherence relations is often established by determining whether a Subject of Consciousness is involved in the construal of the relation. Our automatic subjectivity analysis required using concrete linguistic features to approximate this criterion. We analyzed our data using two different linguistic features associated with subjective relations: subjective adverbs/adjectives, and modal verbs/verbs of cognition. Our results differed depending on the type of subjectivity measure we used, especially when it came to the degree of subjectivity of the genres. The reason for this could be that the subjectivity lexicon we used is more equipped to be used on relatively formal texts than on informal discourse, which would lead to a higher number of subjective adverbs and adjectives detected in the newspaper texts than in our other genres. This scenario is supported by the fact that the analyses on the basis of the modal verbs and verbs of cognition are more in line with observations from prior studies. In order to be better able to analyze the subjectivity of coherence relations across different genres, as well as to improve other applications of subjectivity lexicons such as sentiment analysis, it seems worthwhile to develop a subjectivity lexicon for colloquial discourse.

Although the two types of features resulted in slightly different results, our study largely corroborated most of the subjectivity hypothesis. When taking the results from both subjectivity indicators together (see Tables 5 and 9 and their corresponding analyses), our results also suggest that spontaneously written discourse (chat) contains more subjective relations than spoken discourse, which in turn contains more subjective relations than newspaper texts. In addition, subjectivity patterns of Dutch causal connectives were not found to differ between genres. The subjectivity profiles of the backward causal connectives *omdat* and *want* are also largely corroborated by our data: *want* appears to be more subjective than *omdat*. However, our results for *daarom* and *dus*, both forward causal connectives, were not in line with the subjectivity hypothesis. This could indicate that the subjectivity profiles of *omdat* and *want* are more pronounced than those of *daarom* and *dus*. This result warrants further investigation in future research: is the non-replication of the subjectivity profiles of *daarom* and *dus* due to the methods used in the current automatic approach, or has the small sample size in studies involving manual annotation resulted in conclusions about *daarom* and *dus* that are not supported in larger datasets?

Overall, our study has demonstrated the potential of the automatic analysis of coherence relations using a rule-based approach. Continuing our efforts to corroborate observations made on the basis of small, manually annotated datasets in large-scale corpus studies is an invaluable step in research on discourse coherence, including the subjectivity of causal connectives.

References

Bates, D., Maechler, M., Bolker, B., and S. Walker. 2015. Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* 67(1), pp. 1-48.

Bestgen, Y., Degand, L. and W.P.M.S. Spooren. 2006. Toward automatic determination of the semantics of connectvies in large newspaper corpora, *Discourse Processes* 41(2), pp. 175-193.

Biber, D. and S. Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Carlson, L., Okurowski, M., and D. Marcu. 2002. *RST Discourse Treebank*. Web Download. Philadelphia: Linguistic Data Consortium.

De Smedt, T. and W. Daelemans. 2012. "Vreselijk mooi" (terribly beautiful): A subjectivity lexicon for Dutch adjective, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 3568-3572. Istanbul, Turkey.

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

Degand, L. and H.L.W. Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale in A. Verhagen and J. van de Weijer (eds.), *Usage Based Approaches to* Dutch, pp. 175-199. Utrecht: LOT.

Hoek, J., Evers-Vermeul, J., and T.J.M. Sanders. 2018. Segmenting discourse: Incorporating interpretation into segmentation?, *Corpus Linguistics and Linguistic Theory*, 14(2), 357-386.

Hothorn, T., Bretz, F., and P. Westfall. 2008. Simultaneous inference in general parametric models, *Biometrical Journal* 50(3), pp. 346-363.

Komen, E.R. and J. Hoek. submitted. *Automatic coherence analysis for non-programmers*.

Mann, W.C. and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization, *Text* 8(3), pp. 243-281.

Muller, P., Afantenos, S., Denis, P., and N. Asher. 2012. Constrained decoding for text-level discourse parsing, *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pp. 1883-1900. Mumbai, India.

Oostdijk, N. 2002. The design of the Spoken Dutch Corpus in P. Peters, P. Collins, and A. Smith (eds.), *New Frontiers of Corpus Research*, pp. 105-112. Amsterdam: Rodopi.

Oostdijk, N., Reynaert, M., Hoste, V., and I. Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch in P. Spyns and J. Odijk (eds.), *Essential speech and language technology for Dutch: Theory and applications of natural language processing*, pp. 219-247. Berlin: Springer.

Pander Maat, H.L.W. and T.J.M. Sanders. 2000. Domains of use or subjectivity? The distribution of three Dutch causal connectives explained in E. Couper-Kuhlen and B. Kortmann (eds.), *Cause, Condition, Concession and Contrast: Cognitive and Discourse Perspectives*, pp. 59-81. Berlin/New York: Mouton de Gruyter.

Pander Maat, H.L.W. and T.J.M. Sanders. 2001. Subjectivity in clausal connectives: an empirical study of language in use, *Cognitive Linguistics* 12(3), pp. 247-273.

Pit, M. 2006. Determining subjectivity in text: The case of backward causal connectives in Dutch, *Discourse Processes* 41(2), pp. 151-174.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and B.L. Webber. 2008. The Penn Discourse Treebank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2961-2968. Marrakech, Morocco.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Sanders, T.J.M., Spooren, W.P.M.S., and L.G.M. Noordman, 1992. Toward a taxonomy of coherence relations, *Discourse Processes* 15(1), pp. 1-35.

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

Spooren, W.P.M.S., Berntzen, M., Hulsbosch, M.A., Komen, E.R., and van den Heuvel, H. 2018a. *Whatsapp corpus Berntzen*. DANS EASY [Dataset].

Spooren, W.P.M.S., Hoek, J., Komen, E.R., Hulsbosch, M.A., and van den Heuvel, H. 2018b. *NRC2011*. DANS EASY [Dataset].

Spooren, W.P.M.S., Verheijen, L., Hulsbosch, M.A., Komen, E.R., and van den Heuvel, H. 2018c. *Whatsapp corpus Berntzen*. DANS EASY [Dataset].

Spooren, W.P.M.S. and I. Hendrikx. 2015. *Beyond manual analyses of discourse coherence*. Unpublished manuscript. Centre for Language Studies. Radboud University Nijmegen.

Spooren, W.P.M.S. and L. Degand, L. 2010. Coding coherence relations: Reliability and validity, *Corpus Linguistics and Linguistic Theory* 6(2), pp. 241-266.

Stukker, N.M., Sanders, T.J.M., and A. 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.

Stukker, N.M., Sanders, T.J.M., and A. Verhagen. 2009. Categories of subjectivity in Dutch causal connectives: a usage-based analysis in T.J.M. Sanders and E.E. Sweetser (eds.), *Causal categories in discourse and cognition*, pp. 119-171. Berlin: Mouton de Gruyter.

Sweetser, E.E. (1990). *From Etymology to Pragmatics: The Mind-body Metaphor in Semantic Structure and Semantic Change*. Cambridge: Cambridge University Press.

Van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Tjong Kim Sang, E., and V. Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy in P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch: Theory and Applications of Natural Llanguage Processing*, pp. 147-164. Berlin: Springer.

Verheijen, L., & W. Stoop (2016). Collecting Facebook posts and WhatsApp chats: Corpus compilation of private social media messages in P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.), *Text, Speech and Dialogue: 19th International Conference, TSD 2016, LNAI 9924*, pp. 249–258. Berlin: Springer.

Vis, K., Sanders, J., & Spooren, W.P.M.S. 2012. Diachronic changes in subjectivity and stance - a corpus linguistic study of Dutch news texts. *Discourse, Context & Media* 1.