Dialogue & Discourse 8(1) 132–150 doi: 10.5087/dad.2017.105

Just *because*: In search of objective criteria of subjectivity expressed by causal connectives

Natalia Levshina NATALIA.LEVSHINA@UNI-LEIPZIG.DE
Leipzig University
Institute of British Studies

Liesbeth Degand LIESBETH.DEGAND@UCLOUVAIN.BE
Université catholique de Louvain
Institute for Language and Communication

Editor: Raquel Fernández

Submitted 04/2016; Accepted 01/2017; Published online 02/2017

Abstract

The connective *because* can express both highly objective and highly subjective causal relations. In this, it differs from its counterparts in other languages, e.g. Dutch, where two conjunctions *omdat* and *want* express more objective and more subjective causal relations, respectively. The present study investigates whether it is possible to anchor the different uses of *because* in context, examining a large number of syntactic, morphological and semantic cues with a minimal cost of manual annotation. We propose an innovative method of distinguishing between subjective and objective uses of *because* with the help of information available from an English/Dutch segment of a parallel corpus, which is accompanied by a distributional analysis of contextual features. On the basis of automatic syntactic and morphological annotation of approximately 1500 examples of *because*, every English sentence is coded semi-automatically for more than twenty contextual variables, such as the part of speech, number, person, semantic class of the subject, modality, etc. We employ logistic regression to determine whether these contextual variables help predict which of the two causal connectives is used in the corresponding Dutch sentences. Our results indicate that a set of semantic and syntactic features that include modality, semantics of referents (subjects), semantic class of the verbal predicate, tense (past vs. non-past) and the presence of evaluative adjectives, are reliable predictors of the more subjective and objective uses of *because*, demonstrating that this distinction can indeed be anchored in the immediate linguistic context. The proposed method and relevant contextual cues can be used for identification of objective and subjective relationships in discourse.

Keywords: *because*, causal connectives, objective and subjective causality, parallel corpus

1 Theoretical background and aims of this study

This paper deals with the distinction between subjective and objective uses of the English causal connective *because*, and proposes a method to distinguish between these two uses in discourse. The issue of subjective versus objective (causal) connectives has received a lot of attention in linguistics, going back to seminal studies like Rutherford (1970), Van Dijk (1979) and Sweetser

(1990). These studies demonstrate that connectives like *because* seem to have systematically different patterns of meaning and use. More recently, the distinction between objective and subjective causal relations has been put forward in studies of coherence relations and their linguistic markers in different languages (Canestrelli, 2013; Degand & Fagard, 2012; Pander Maat & Sanders, 2001; Sanders & Sweetser, 2009; Stukker & Sanders, 2012; Zufferey & Cartoni, 2012). Objective causal relations express causality between events in the real world, as in (1), whereas subjective causal relations express the speaker's motivation for mental conclusions or speech acts, as in (2)[1].

(1) *The tennis match was cancelled because it had been raining too much.*
(2) *Apparently it has been freezing, because the geraniums are dead.*

This distinction is grounded in Sweetser's (1990) seminal trichotomy establishing (causal) relations in the content domain, epistemic domain and speech-act domain, illustrated with her examples (1990: 77–78) in (3-5), respectively. Thus, CONTENT use is based on the cause-and-effect relationships in the real world; EPISTEMIC use introduces the speaker's reason for making a conclusion, and SPEECH ACT use expresses the motivation for the speaker's performing a particular speech act, e.g. asking a question in (5).

(3) *John came back because he loved her.*
(4) *John loved her, because he came back.*
(5) *Since you are so smart, when was George Washington born?*

An alternative classification of causal relations and connectives is in terms of a Speaker Involvement scale, where speaker involvement "refers to the degree to which the present speaker is implicitly involved in the construal of the causal relation … speaker involvement increases with the degree to which both the causal relation and the related units are constituted by the assumptions and actions of the present speaker." (Pander Maat & Degand, 2001: 214). The proposed scale for causal relations is given in (6) (based on Pander Maat & Degand 2001: Table 1):

(6) non-volitional < volitional < causal epistemic < noncausal epistemic < speech act

Broadly speaking, non-volitional and volitional causal relations correspond to the content domain[2]; causal and noncausal epistemic to the epistemic domain[3], and speech act to the speech act domain, obviously. For more details on the nature of these finer-grained distinctions, we refer the reader to the original publications (Pander Maat & Degand, 2001; Degand & Pander Maat, 2003). More important for our present purposes is the distinct operationalization of subjectivity in the approaches presented, namely a categorical (based on Sweetser's trichotomy) *vs.* a scalar one

[1] Degand and Fagard (2012) distinguish between objective, subjective and intersubjective causal relations, thus broadly mapping Sweetser's (1990) original domain distinctions. Here we propose to follow the bipartite classification in objective content-based causal relations, on the one hand, and subjective relations, on the other hand, the latter including both reasoning-based and speech-act-based causal relations (see also, Pander Maat & Sanders, 2001; Sanders & Spooren, 2015).

[2] The distinction between volitional and non-volitional relations was introduced in Rhetorical Structure Theory (Mann & Thompson, 1988), and further developed in later work on causal connectives (Degand, 2001; Pit, 2006). A non-volitional (causal) relation comes about without the intervention of a volitional actor, while a volitional relation results from the will of a conscious, active, participant (Sanders & Pander Maat, 2001).

[3] In Pander Maat and Degand's (2001) proposal, causal epistemic relations are grounded in deductive reasoning, while noncausal epistemic relations are grounded in abductive reasoning. This distinction is not followed here.

(Pander Maat & Degand, 2001). It is the categorical approach that we will follow here based on the typical mapping between connective use and (causal) domain.

In Pander Maat and Degand's words, a (causal) "connective encodes a certain speaker-involvement level, which it contributes to the interpretation of its discourse environment. When this level is too low or too high to be combined with the level allowed for by the discourse environment, the use of the connective is inappropriate." (p. 230). In other words, there are (semantic) constraints on the relational context in which a given connective can appear. These constraints are of course language-specific (cf. Sanders & Sweetser, 2009). While the English connective *because* can be used to express both subjective and objective causal (backward) relations,[4] as illustrated in examples (1-4) (see also Couper-Kuhlen, 1996; Ford, 1994; Kac, 1972; Knott & Dale, 1994; Knott & Sanders, 1998; Schleppegrell, 1991), other languages have specialized connectives to express different types of causal relations (Pit, 2007; Stukker & Sanders, 2012; Zufferey & Cartoni, 2012), and Dutch is a case in point. Several corpus-based studies have established a systematic relationship between the different types of causal relations and the connectives used to express these relations. More specifically, adopting a categorical point of view, the connective *want* is a typical marker of subjective relations, and the connective *omdat* typically expresses objective relations (e.g. Degand & Pander Maat, 2003; Pit, 2006; Sanders & Spooren, 2009).[5] These are strong tendencies which have been confirmed both for spoken, spoken-like and written data (Spooren, Sanders, Huiskes & Degand, 2010; Sanders & Spooren, 2015), even if there is no one-to-one relation between the connective and the type of causal relation (see Sanders & Spooren, 2013).

While the distinction between subjective and objective causal relations is conceptually fairly straightforward to grasp, categorizing authentic data in terms of one or the other appears to be a lot less straightforward, not the least because the criteria used to classify less prototypical examples vary (cf. Sanders, Vis & Broeder 2012 on the reusability of causal connective corpus studies). Spooren and Degand (2010) report kappa agreement values of .60 or lower for variables that are central to determining the degree of subjectivity of a causal relation marked by *want*. In their discussion of the sources of disagreements, they distinguish two distinct cases: (i) disagreements resulting from "real ambiguities", i.e. segments which can receive several interpretations; such disagreements are probably inevitable, and (ii) disagreements that are in fact coding errors resulting from a misinterpretation of the categorization variables (pp. 250-251). It is the latter type of disagreements, which are avoidable, which we would like to tackle in this study.

The present paper thus pursues the following goals. From a theoretical perspective, we want to find contextual cues that are strongly associated with objective and subjective uses of causal connectives, focusing on the connective *because*, which can express all of these meanings. This also has a practical use: such cues will help annotators in the future to code the subtle semantic distinctions and make the annotation more reliable.

To reach these goals, we propose a semi-automatic approach based on an innovative combination of the data from a parallel corpus and the methodology of distributional semantics. The cross-linguistic data enables one to use a more specific language to make predictions about another language that does not make certain distinctions. In this study, we disambiguate the functions of the English *because* with the help of Dutch, a language which makes the distinction

[4] In a backward causal relation, the causal segment is the host of the causal connective and is in general preceded by the consequence segment, as in (1): Segment1 = the match was cancelled (*effect* or *consequence)* because Segment2 = it has been raining (*cause*).

[5] Dutch also has a highly specific connective *doordat* (and its forward counterpart *daardoor*) specialized to express non-volitional (highly objective) causal relations. This connective is however a lot less frequent (Degand, 2001; Pit, 2003). With only 62 occurrences, it was also relatively rare in our data (see Section 2.1).

between more objective and more subjective causal relations by means of different connectives, most importantly, *omdat* (more objective) and *want* (more subjective). For this purpose, we use aligned English and Dutch sentences from the OPUS version of the Europarl corpus (Koehn, 2005; Tiedemann, 2012), which contains proceedings of the European Parliament. Examples (7-9) illustrate the principle. In (7), the English because segment is aligned with Dutch *omdat*, expressing objective causal relations (non-volitional in 7a and volitional in 7b).

(7) a. OPUS Europarl 33356150

*Listen to and see the agony of a doctor who has to tell a non-smoker that she has cancer **because** she breathed the smoke of someone who thought that smoking looked cool on a screen or on a page.*

*Stelt u zich de pijn voor die een arts voelt wanneer hij een nietrookster moet meedelen dat zij kanker heeft **omdat** zij de rook heeft ingeademd van iemand die vond dat roken "cool" stond op televisie of in de pers.*

b. OPUS Europarl 2458129

*I voted for the Atkins report **because** it is extremely important for pensioners…*

*Ik heb voor het verslag-Atkins gestemd **omdat** het voor de gepensioneerden (...) zeer belangrijk is...*

In contrast, *want* expresses epistemic subjective causality, as in (8a), where the subordinate clause provides an explanation for the speaker's inference presented in the main clause, and subjective speech-act relations, as in (8b), where the subordinate clause contains the reason for asking the question in the main clause. Again, both segments are aligned with English *because*:

(8) a. OPUS Europarl 33885879

*Mr Henderson knows that **because** he is on the Council.*

*Dat weet de heer Henderson **want** die zit in de Raad...*

b. OPUS Europarl 2439113

*Is it available, **because** I have noticed that many MEPs have not seen this text?*

*Is hij ter beschikking, **want** ik merk dat heel veel collega's deze tekst niet gezien hebben.*

The second component of our method involves the use of distributional semantics (Wittgenstein, 1953; Harris, 1954; Firth, 1957, etc.). According to this approach, semantic properties of words and constructions are closely linked with contextual environments where these words or constructions occur. Moreover, different senses of a word or construction will be observed in different types of contexts. In other words, these senses will have different distributional properties. This idea goes back to structuralist semantics (e.g. Apresjan, 1966) and has been more recently implemented in automatic algorithms of word sense disambiguation (e.g. Pedersen, 2006). We will use a bottom-up approach and employ contextual variables, which can be coded semi-

automatically with the help of syntactic and morphological information about the English sentences with *because*. These variables will be investigated with the help of logistic regression analysis in order to select those contextual features that can be used for distinguishing between objective and subjective uses of *because*, which correspond to *omdat* and *want*, respectively.

The idea of using distributional clues for subjectivity analysis is not new. In fact, there has been some work on subjectivity word sense disambiguation (SWSD) in computational linguistics (e.g. Akkaya et al., 2009). Consider (9) as an illustration. The noun *alarm* has an objective meaning in (9a) and a subjective one in (9b):

(9) a. *The **alarm** went off.*

b. *His **alarm** grew*. (Akkaya et al., 2009: 191)

The task for a SWSD algorithm is to determine whether the word is used subjectively or objectively on the basis of contextual clues. Although the type of subjectivity and objectivity in the lexicon is different from the one established at the clausal level, this task is essentially rather similar to ours. However, there are important differences. First, the purpose of SWSD is to classify words, sentences or texts as correctly as possible. The process of deciding for subjective or objective meaning is ultimately a black box. In our study, we develop a method of automatic disambiguation, too. Our primary goal, however, is to learn which contextual features can help us discriminate between the subjective and objective uses of *because*. Second, the contextual cues in our study are at a higher level of abstraction than those in SWSD. We employ diverse syntactic and morphological information about the clauses that are connected by the conjunction, as well as the semantic classes of the subjects and predicates, since these features have been demonstrated to work well for disambiguation of discourse relationships expressed by connectives (e.g. Pitler & Nenkova, 2009). In contrast, the SWSD approach is usually based on more lexically specific clues (surrounding words).

The rest of the paper is organized as follows. Section 2 presents the data (parallel corpus) and contextual variables. Section 3 reports the results of the statistical analysis. Finally, Section 4 summarizes the findings and suggests some directions for further research.

2 Data and contextual variables

This section describes the data and variables that were used in this study.

2.1 Data source and extraction procedure

The data come from Europarl, a collection of the European Parliament proceedings in 21 languages of the European Union (Koehn, 2005), which constitutes a part of the OPUS corpus (Tiedemann, 2012). This corpus was chosen because the proceedings contain many uses of causal clauses in various functions, such as defending one's political position, justification of requests, explaining why one chooses a particular wording, or introducing the causes and consequences of some socially relevant events.

The OPUS query engine was used to extract 5,000 examples of contexts that contained *because* in the English version.[6] These contexts were manually checked. We kept only those sentences where *because* was used as a subordinate conjunction (thus excluding the preposition *because of*) and corresponded to *omdat* or *want* in the Dutch version. The direction of translation was not taken into account. The initial large sample contained a large number of repetitions. This is why we had to remove all repeated sentences. We also discarded the contexts in the following cases:

[6] See http://opus.lingfil.uu.se (last access 01.02.2017).

- causal connectives with adverbial modifiers in Dutch, e.g. P, *juist omdat* "just because" Q; P, *niet in de laatste plaats omdat* "not least because" Q;
- paired conjunctions in Dutch, e.g. P, *niet (alleen) omdat* "not (only) because" *Q, maar (ook) omdat* "but (also) because" R;
- the subordinate clause with the causal connective in the preposed position in Dutch (*omdat* Q, P).

In all these cases, only *omdat* can be used, *want* being excluded for syntactic reasons. This is why it would not make sense to include these contexts in the sample. As a result of this cleaning procedure, the final sample contained 1521 examples in total, 798 with *omdat* and 723 with *want*. These sentences (in the English version) were first analysed manually, so that the clauses that represented the cause and effect were extracted. After that, the clauses were parsed automatically with the help of the Stanford Parser (Klein & Manning, 2003). As a result, we obtained the syntactic dependencies and morphological information (part of speech) about every word in a clause. With the help of a Python script written specifically for this purpose, this morphological and syntactic information was used for data annotation. The variables are presented in Section 2.2. The annotation was manually checked.

2.2 Contextual variables

This section describes the contextual variables that were used for distinguishing between the uses of *because* that correspond to *omdat* and *want* in the Dutch segment of Europarl. All of these variables, except for the one describing the Dutch connectives, represent the structural and semantic properties of the English sentences. First, we describe the variables related to the entire sentence. Many of these variables were inspired by corpus analytic work on the expression of subjectivity and objectivity in language, mainly Pit (2003, 2006) and Torres Cacoullos and Schwenter (2005). The motivation for most of the variables is specified below. Note however, that a decisive criterion for including a variable was whether it was possible to code it automatically on the basis of the available syntactic and morphological information from the parser. An exception is the semantic coding of the subjects, but in the future, we hope to be able to perform semantic classification automatically, too, with the help of the state-of-the-art word sense disambiguation methods and semantic resources, such as WordNet.

- Dutch connective: whether the Dutch equivalent of *because* was *omdat* or *want*.
- The presence or absence of direct address in the sentence with *because*. For example, (10) contains direct addresses *Mr President* and *ladies and gentlemen*:

  (10) OPUS Europarl 37571406

  Mr President, ladies and gentlemen, I voted against the Thyssen report, ***because*** *it is detrimental in many respects to certain business sectors and of course to people starting up in business.*

  If the speaker uses a direct address, he or she highlights the interactive character of discourse, which makes it more subjective.

- The presence or absence of evaluative adjectives (positive or negative). This variable was added because manual analysis has shown that subjective causal relations tend to co-occur with evaluative adjectives and/or adverbs, while objective causal relations do not (Pander Maat & Degand, 2001; Pit, 2006; Sanders & Spooren, 2015). The variable was

coded with the help of SentiWordNet (Baccianella, Esuli & Sebastiani, 2010). In this data base, which has the structure similar to one of WordNet (Fellbaum, 1998), words have scores along the positive, negative and objective dimensions. We considered the word emotionally charged if its scores either on the negative or positive scales were greater than zero (the scores were averaged across the meanings). Consider (11):

(11) OPUS Europarl 94416

*Colleagues, I am in a very difficult position **because** I cannot change the agenda.*

This sentence contains the adjective *difficult*, which has negative scores 0.75 (in the sense 'hard') and 0.625 (in the sense 'unmanageable', as in *difficult child*). We expect the more emotionally charged contexts to be more subjective. The coding was done automatically; no word sense disambiguation was performed. The aim was to see how far we can go with a fully automatic approach.

- The presence or absence of evaluative adverbs. The logic and the procedure were the same as above.
- explicit references to conceptualizer and his or her conceptualization of causal relationships. In other words, the mental process of establishing a causal connection becomes explicit. This information could be present in structures like *I/We/X think/believe/consider/feel/trust… that P because Q*, where Q gives a reason for *my/our/X's* thinking, believing, etc. that P. According to Langacker (1985), an explicit reference to the conceptualizer objectifies him or her. Usually, *omdat* is used in these cases, as in (11), whereas *want* is used with implicit conceptualizers (cf. Pit, 2003; Sanders & Spooren, 2009). Compare the examples in (12) and (13):

(12) OPUS Europarl 2347145

(…) *I think that it is imperative to discuss plans for extending it and to look at pricing again, **because** we simply have to get a hold on the situation.*

(…) *ik denk dat het dringend noodzakelijk is opnieuw over uitbreidingsplannen en prijzen te praten **omdat** we de situatie gewoon onder controle moeten krijgen.*

(13) OPUS Europarl 5709553

*This is necessary **because** it is the only way to achieve an increasingly balanced labour market.*

*Dat evenwicht is ook nodig, **want** alleen op die manier kunnen we zorgen voor een evenwichtige groei van de werkgelegenheid.*

The variables that are listed below were coded for the main and subordinate clauses separately:

- Part of speech of the subject of the clause. The values are 'Noun' (including nominalizations, e.g. the rich), 'Pronoun' (all possible pronouns) or 'No Subject' (in case there are no subjects). The subject of a clause very often corresponds to the causally primary participant (or CP, Pit, 2006) which plays an important role in the causal conceptualization of the state or event. According to Pit (2006: 162) "CPs referred to by a nominal are more objective (less deeply perspectivized) than CPs referred to by a pronominal" (cf. Langacker, 1985: 126–127). Furthermore, subject coreferentiality (morphologically coded by pronouns) has been shown to reflect subjective use (Torres Cacoullos & Schwenter, 2005);
- Grammatical person of the subject: '1st', '2nd', '3rd' or 'No Subject'. This variable is motivated by the assumption that first and second person participants are more subjective than third person participants;
- Grammatical number of the subject: 'Singular', 'Plural' or 'No Subject'. One can expect subjective contexts to be more associated with singular cognizers;
- Semantic class of the subject: 'Animate' (including people, animals and organizations), 'Inanimate' (all the rest) or 'No Subject'. The assumption is that subjective contexts are associated with animate subjects more than with inanimate ones;
- Tense of the finite predicate: 'Present', 'Past', 'Future', 'Other' (modals and imperatives), 'No Predicate' (in case there is no finite predicate), where present tense is assumed to mark more subjective relations (in the "here and now") (Pit, 2003; Pander Maat & Degand, 2001);
- Voice of the finite predicate: 'Active', 'Passive' or 'No Predicate'. One would expect passive forms to be more associated with objective contexts (Biber 1988);
- The presence of a modal verb in the predicate: 'Yes', 'No' or 'No Predicate' (see above), where modal verbs would signal greater subjectivity;
- Polarity: 'Positive' or 'Negative' (when the clause contains a negation). We expect negative polarity contexts to be more subjective because negation is central in political debates, where speakers correct or reject different proposals;
- Semantic class of the verbal predicate: 'Mental' (verbs of perception, desire, thinking, resolution, etc.), 'Social' (verbs of communication), 'Other' (all other verbs) or 'No Predicate'. Mental and social verbs are expected to be more frequently present in subjective contexts.

The semantic class annotation of the subjects was tested first on a small sample of 200 observations by the co-authors. For the main clauses, Cohen's kappa was 0.89, and for the subordinate clauses, it was 0.79. From that we concluded that the annotation schema was reliable enough and coded the entire sample. Due to much lower kappa scores for the verb classes in the pilot study, it was decided to code the verbs following a closed list approach. The list was based on Levin's (1993) classes, where the mental verbs included the classes "declare", "conjecture", "see", "sight", "peer", "feel", "admire", "marvel", "want", "long", positive and negative judgement verbs, "assess", "investigate", whereas the communication verbs included the classes "tell", "snap/cackle", "cable", "talk", "chitchat", "say", "complain" and "advise".

For an illustration of the coding schema, consider the example in (14):

(14) OPUS Europarl 18558329

*We want a new framework agreement **because** without the ability to scrutinise your Commission effectively, we cannot do our job properly.*

We willen een nieuwe kaderovereenkomst, ***want*** *zonder het vermogen uw Commissie doeltreffend te controleren, kunnen wij ons werk niet goed doen.*

The sentence has the following values:

1. Dutch conjunction: *want*

2. Variables coded for the entire English sentence:
- Direct address: No
- Evaluative adjective: Yes (*new*)
- Evaluative adverbs: No
- Explicit reference to conceptualization of causation: No

3. Variables coded for the main and subordinate clauses separately

3.1. The properties of the main clause
- Part of speech of the subject of the main clause: Pronoun
- Person of the subject of the main clause: 1st
- Number of the subject of the main clause: Plural
- Semantic class of the subject of the main clause: Animate
- Tense of the finite predicate of the main clause: Present
- Voice of the finite predicate of the main clause: Active
- Modality of the main clause: No
- Polarity of the main clause: Positive
- Semantic class of the verbal predicate of the main clause: Mental

3.2. The properties of the subordinate clause
- Part of speech of the subject of the subordinate clause: Pronoun
- Person of the subject of the subordinate clause: 1st
- Number of the subject of the subordinate clause: Plural
- Semantic class of the subject of the subordinate clause: Animate
- Tense of the finite predicate of the subordinate clause: Present
- Voice of the finite predicate of the subordinate clause: Active
- Modality of the subordinate clause: Yes
- Polarity of the subordinate clause: Negative
- Semantic class of the verbal predicate of the subordinate clause: Other

The relevance of these variables for the disambiguation of subjective and objective causal relations was investigated in the statistical analyses that are presented in Section 3.

3 Quantitative analyses: data transformation and logistic regression

This section first describes the data transformation procedures and next reports the results of our logistic regression modelling.

3.1 Data transformation

Regression analysis is sensitive to low-frequency values and strong associations between predictors. These factors can seriously undermine the quality of a model. This is why some of the

initial variables were conflated, as well as some values of the variables, on the basis of standard regression model diagnostics (see Levshina, 2015: Ch. 12). The transformed variables were the following:

- Semantic class, part of speech and person of the subject, both in the main and subordinate clause. The resulting variables were called simply 'Subject' and contained the values 'Speech Act Participants (SAP)' (i.e. 1st and 2nd person pronouns), 'Animate' (all other animate subjects) and 'Inanimate'.
- Tense of the verbal predicate, both in the main and subordinate clauses. To optimize the analyses, we recoded the variable as 'Past' and 'Non-past' (present, future, other, no predicate) on the basis of the model diagnostics.

All cases with incomplete sentences without verbal predicates or subjects were removed, because they produced data sparseness and made the model suboptimal. In total, we had 1512 observations left.

3.2 Logistic regression model

We performed a binary logistic regression analysis. This is a method for modelling the binary outcome (here, *omdat* or *want*) that can be predicted from other variables, which are usually called predictors. Here, the predictors were the variables that describe the English contexts. We fitted a multiple regression model, where the effect of each predictor of interest was measured while controlling for the other predictors. We used the free statistical software R (R Core Team, 2015) with an add-on package *rms* (Harrell, 2015).

The model structure was determined by using the following procedure. First, a full model with all pairwise interactions was defined on the basis of bidirectional (backward and forward) stepwise selection based on all predictors and all possible pairwise interactions between them. Interactions are observed when the effect of two or more variables on the outcome is non-additive. Stepwise selection means that the algorithm tries to add (forward selection) or remove (backward selection) a variable or interaction term one by one, until no further improvement can be made. The criterion for improvement was AIC (Akaike Information Criterion), which shows how well a model fits the data, while at the same time giving advantage to more parsimonious models with a smaller number of predictors. After that, usual diagnostic tests were performed. Unfortunately, a bootstrap validation revealed that the model suffered from severe overfitting. In particular, the optimism, which is commonly used as a diagnostic statistic, was about 0.38 in the slope and 0.11 in R^2 (cf. Harrell, 2001). These levels were too high to be tolerated. All that means that the model would be useless when applied to new data and has thus little scientific value. For a detailed explanation of the statistical procedures, see Levshina (2015: Ch. 12).

To solve that problem, we tried to minimize the number of interactions by excluding all non-significant ones and those that do not change the direction of the predictors' effects. This did not help to solve the problem of overfitting, even when a penalty was applied for shrinkage (*Ibid.*). The only model with acceptable optimism was the one with main effects only, selected on the basis of backward elimination with AIC as the criterion (see Table 1). Still, we had to apply a penalty to shrink the estimates (with the penalty factor of 6.8) in order to correct for the undue optimism. We also tested possible interactions between the predictors manually and inspected them graphically, but they did not reveal significant cross-over effects and therefore were not included in the final model.

Although some variables are clearly related (e.g. only animate subjects can take mental verbs as predicates), logistic regression is known to be robust with regard to some correlations between predictors. Moreover, there were no symptoms of strong multicollinearity, since all VIF (Variance

Inflation Factor) scores, which are used traditionally for diagnostics, were below 5. All this means that we do not have reasons for concern regarding the quality of the model.

The predictive power of the final model (i.e. how well it could discriminate between the contexts that corresponded to *want* and *omdat* in Dutch), was modest, with the concordance index $C = 0.67$ and Nagelkerke's $R^2 = 0.13$. As a rule of thumb, the *C* value should be at least 0.7 for the model to be considered good. The accuracy, defined as the proportion of correct predictions made by the model, was 0.62. This is greater than the baseline level of 0.52, which would be the probability of making a correct prediction if one always selected the more frequent response, i.e. *omdat*. In our view, this result can still be regarded as satisfactory because we try to predict the use of a Dutch connective from its equivalent contexts in English, where numerous other factors play a role (see discussion in Section 4).

Parameter	**Coefficient**	**SE**	***p*-value**
Intercept	-1.56	0.23	< 0.001
SAP subject in main clause (in contrast with animate)	0.81	0.18	< 0.001
Inanimate subject in main clause (in contrast with animate)	0.85	0.18	< 0.001
Singular subject in main clause	0.23	0.13	0.089
Mental verb in main clause	0.35	0.14	0.012
Verb of communication in main clause	0.05	0.21	0.826
Non-past tense in main clause	1.03	0.19	< 0.001
Passive predicate in main clause	-0.30	0.22	0.17
Modal verbs in main clause	0.55	0.14	< 0.001
SAP subject in subordinate clause (in contrast with animate)	0.43	0.16	0.009
Inanimate subject in subordinate clause (in contrast with animate)	0.48	0.15	< 0.001
Singular subject in subordinate clause	0.28	0.13	0.029
Passive predicate in subordinate clause	-0.27	0.17	0.112
Modal verb in subordinate clause	0.48	0.16	0.003
Evaluative adjective(s)	0.19	0.11	0.076

Table 1. Logistic regression coefficients and their standard errors and *p*-values.

The final regression estimates are presented in Table 1. The coefficients in the second column of the table are log-odds ratios. An exception is the intercept, which represents logarithmically transformed odds of *want* against *omdat* in the reference level context, i.e. when all variables have the default values, or the values opposite to those displayed in the table. Positive coefficients (log-odds ratios) show that this value of a variable increases the odds of *want* in the Dutch sentence in comparison with *omdat*. Negative coefficients, in contrast, indicate that the value increases the likelihood of *omdat* in comparison with *want*. The other columns in the table contain the standard errors, which give an idea of how variable the estimates may be, as well as the *p*-values, which are used in frequentist statistics to determine if the effect is statistically significant. A *p*-value below the conventional level of 0.05 serves as an indication that the effect is not due to chance alone. The values between 0.05 and 0.1 are often considered as marginally significant. In a pilot study, it makes sense to try to interpret the marginally significant variables (here, between 0.05 and 0.2), as well, so that they can be more carefully investigated in larger-scale follow-up studies.

The estimates suggest the following. Both in the main and subordinate clauses, the preferences are rather similar. First, the presence of modal predicates, singular subjects (marginally significant

in the main clause), SAP subjects (or first and second person subjects) and inanimate subjects in comparison with animate 3rd person subjects increase the odds of *want* in the Dutch version. Moreover, there are marginally significant effects of passive predicates, both in the main and subordinate clauses. The passive forms tend to increase the chances of *omdat* to be found in the Dutch sentence. Non-past tense forms and mental verbs (as opposed to verbs of communication and all other verbs) in the main clause increase the odds of *want*. There is also a marginally significant effect of the presence of evaluative adjectives, which boost the chances of *want*, too.

The effects of SAPs as subjects, evaluative adjectives, mental verbs, non-past and non-passive verbs are theoretically interpretable. Most of these features are typical of involved non-abstract, non-technical communication (Biber, 1988), which is characterized by a high degree of subjectivity. This kind of communication is contrasted with informative and abstract, technical types of discourse. The distinction between involved and informational and abstract communication closely corresponds, in our opinion, to the distinction between subjectivity and objectivity of discourse. In involved communication, the speaker and the hearer, their beliefs, attitudes and intentions are in the centre of attention. In contrast, abstract technical communication distances from the speaker and hearer's personal experiences, focusing instead on the properties and events of the external world. In addition, the use of modals is a typical feature of explicit marking of the speaker's own point of view or, alternatively, of argumentative discourse designed to persuade the addressee (*Idem.:* 111). Therefore, the use of modals is indicative of subjective communication. It is more difficult to explain, however, why the plural and 3rd person animate subjects disfavour *want*. A close inspection of the individual contexts suggests that these are often the names of social and political groups and entities, including members of a party (15a) and representatives of countries (15b).

(15)a. OPUS Europarl 912705

Mr President, let me repeat what I said yesterday, namely that ***the French Socialists*** *will not take part in the vote because they believe that this is not the correct procedure, since the only environmental directives referred to are those on wild birds and on Natura 2000 and a more balanced approach should have been taken.*

b. OPUS Europarl 689846

That really is a brave decision, because some ***Member States*** *will clearly find themselves towards the bottom of the league table, which is something nobody likes.*

In these contexts, the people are conceptualized as political groups, rather than as individual subjects, which explains why the more objective causal connective is used.

The regression model data also enable us to compute the so-called fitted values of every example in the data set, i.e. the predicted probabilities of *want* and *omdat* based on the values of the predictors and their coefficients in the regression model. The two sentences with top highest predicted values of *want* are given in (16a), where the predicted probability of *want* was 80%, and (16b), where the predicted probability of the connective was 78%. This means that the sentences contain many features that boost the probability of *want*. Not surprisingly, the corresponding Dutch sentences contained the predicted connective.

(16)a. OPUS Europarl 25188486

*Well, Mr. Belder and colleagues from the PPE-DE and ALDE groups, either your homework has not been done properly, or we must congratulate the magical powers of the Commission, **because** two years ago it must have eaten some Chinese fortune cookie which said that in September 2006 Parliament would make such a call to initiate a structured dialogue.*

*Welnu, geachte heer Belder en collega's van de PPE-DE-Fractie en de ALDE-Fractie, of u heeft uw huiswerk niet goed gedaan, of we moeten dankbaar zijn voor de magische krachten van de Commissie, **want** twee jaar geleden moet de Commissie een of ander Chinees gelukskoekje hebben gegeten waarop stond dat het Parlement haar in september 2006 zou vragen het initiatief te nemen voor een gestructureerde dialoog.*

b. OPUS Europarl 23706492

*There must be a typing error in the Commission's speech, **because** I would have thought you would very happily have looked forward to countries introducing more stringent legislation in order to achieve the Kyoto objectives.*

*Verder vraag ik mij af of er geen tikfout in de getallen van de Commissie zit, **want** u zou het toch met vreugde hebben begroet als landen striktere wetgeving invoeren om de doelstellingen van Kyoto te bereiken?*

The examples in (16a) and particularly in (16b) are cases of an epistemic use in English. Interestingly, the Dutch version of (15b) contains a reported question in the main clause *Ik vraag mij af of…* "I ask myself whether…" and a question in the subordinate clause with the modal particle *toch*, which reflects the speaker's desire to be reassured or confirmed.

For *omdat*, the sentences with the highest predicted scores (96% and 95%, respectively), are shown in (17a) and (17b). Again, the Dutch versions contain the predicted conjunction *omdat*.

(17)a. OPUS Europarl 20477436

*These countries were deprived of their sovereignty for many decades **because** they had no partner prepared to perform the duties of an ally without hesitation.*

*...die gedurende decennia verschrikkelijk hebben geleden en hun soevereiniteit kwijt waren, **omdat** zij niet over een partner beschikten die er niet voor terugschrok zijn verplichtingen als bondgenoot na te komen.*

b. OPUS Europarl 16019572

*A constituent of mine bought a flight online with the notorious Ryanair but when they went to collect the ticket they were denied access **because** they had an international student ID card and were refused boarding on the grounds that it was out of date.*

*… Toen deze persoon het ticket ophaalde mocht hij niet aan boord **omdat** hij een internationale studentenkaart had die verlopen was.*

In both cases, the subjects of the main clauses are non-agentive, since they are affected by someone else's actions. The causes are either historical consequences, as in (16a), or impersonal rules (16b). Thus, the causal relationships are construed as objective.

3.3. Cases of mismatches

It is also instructive to study the cases of mismatches, where the Dutch connective used in a particular context has a low predicted probability. This could help us identify the reasons why the prediction is far from being perfect. e.g. whether there are additional contextual variables that need to be taken into account, or whether there are some structural differences between English and Dutch that distort the picture. In order to identify the cases where the predictions made by the model differed the most from the observed Dutch connective, we performed the following. First, we computed the predicted scores, as was shown in Section 3.2. The greater the score, the higher the probability of *want*. Next, we binarized the observed categories, with *omdat* having the value 0 and *want* corresponding to 1. After that, we computed the differences between the binarized outcome and the predicted scores. The observations with the greatest absolute differences between the observed and predicted scores were examined. The overwhelming majority of the examples with the greatest mismatch scores contain *want*, but the model predicts *omdat* with a very high probability. Example (18) shows the observation with the greatest mismatch, where *omdat* was predicted with the probability of almost 90%.

(18) OPUS Europarl 24326485

*I would like to stress, however, that the decisions leading to such a situation were probably not taken by women, **because** there are practically no women in the places where decisions are taken on security policy or at negotiation tables.*

*Naar alle waarschijnlijkheid waren het echter geen vrouwen die de beslissingen namen die tot die situatie hebben geleid, **want** op plekken waar wordt besloten over veiligheidsbeleid en aan de onderhandelingstafels is vrijwel geen vrouw te vinden.*

In spite of the fact that the English sentence contains such *omdat*-favouring features, as the passive predicate, inanimate subject in the main clause and past tense, it expresses, however, the speaker's conjecture. The sentence contains the adverb *probably*, which expresses epistemic modality. The subordinate clause provides the speaker's grounds for making this conjecture. This marker corresponds to the phrase *naar alle waarschijnlijkheid* "by all odds" in Dutch. Thus, epistemic modality markers may be a new variable that should be added to the list of markers.

Another example of a mismatch is provided in (19). Again, the sentence has a high probability of *omdat* (almost 87%), but the Dutch translation contains *want*. This sentence expresses the speaker's evaluation of someone else's actions ('X was right to do Y because…'), and provides the reason for this evaluation. In Dutch, the evaluation is present, as well, although the structure is somewhat different ('X rightly did Y because…'), with the adverb *terecht* "rightly".

(19) OPUS Europarl 18486909

*Mr Schulz was right to draw attention to Commissioner Vitorino s important role, **because** the excellent result has been achieved partly thanks to his input and influence.*

De heer Schulz heeft terecht gewezen op de belangrijke rol van commissaris Vitorino, ***want*** *mede door zijn inbreng en invloed is een heel goed resultaat geboekt.*

The other examples of mismatch where the subjectivity of a sentence is not captured by the model are similar to the ones provided above. They show that subjectivity can be expressed by a rich variety of linguistic strategies. A full account of such strategies and their automatic extraction remains a task for future research, however. Although some of these subjectivity indicators are lexical and easy to list, e.g. English *probably* and Dutch *terecht* "rightly", many other expressions are periphrastic and more difficult to capture automatically. For example, the phrase 'X was right to do Y because…', allows for a large number of possible modifications, such as *X was completely right not to openly criticize this idea* or *Nobody can accuse me of having been wrong to do so*.

Another possible reason of mismatches is a non-perfect equivalence between the English and Dutch versions. Consider (20), where the speech act (directive) in English becomes a modal clause with *moeten* "must" mitigated by *Ik vrees dat…* "I'm afraid that…" in Dutch.

(20) OPUS Europarl 38368559

Let us try to redirect the situation ***because*** *your reply has also been very general.*

Ik vrees dat we iets dieper op de zaak zullen moeten ingaan, mijnheer de commissaris, ***want*** *uw antwoord blijft nogal algemeen.*
"I'm afraid we must go somewhat deeper into the matter, Mr. Commissioner, because your reply is still very broad."

However, in the overwhelming majority of our examples, the features of the Dutch sentences are faithfully reflected in the English contexts, and vice versa. The lack of full structural and semantic correspondence is thus not the main factor that can explain the mismatches.

4. Conclusions and outlook

Using the data from a parallel corpus, we have managed to predict the choices between the more objective Dutch causal connective *omdat* and the more subjective *want* on the basis of the contextual properties of the English sentences with *because*. We found that such semantic and syntactic features as modality, semantics of referents (subjects), semantic class of the verbal predicate, tense (past vs. non-past) and the presence of evaluative adjectives are significantly associated with the use of *omdat* or *want* in the Dutch sentence. This means that our pilot study has shown promising results in disambiguation between objective and subjective uses of causal connectives. In view of the high cost of manual annotation of discourse connectives, we would like to suggest that the semantic and syntactic features we identified as predictors of subjective and objective contexts may be used to automatically annotate (or rather pre-annotate) the data. In a second step, this automatic annotation would then be verified or corrected by manual analysis thus gaining time and analysis effort. We hope that these features will be applied in new research on other discourse markers. If we aim at facilitating the annotation process, its reliability should also be taken into account. In addition to being work-intensive, manual annotation of discourse connectives is known to give rise to fairly low interrater agreement (Spooren & Degand, 2010). An

interesting question is therefore whether the automatic annotation compares to the manual one, or would even outrank it. A way to find out is to manually code part of the data in order to calculate agreement between the automatic and the manual annotation. However, it is worthwhile to question whether the manual annotation is really the more reliable than the automatic one. This is an issue which will have to be left for future research.

Another area in which our approach could also be useful is objective measuring of the degree of subjectification and intersubjectification in grammaticalization, for instance by uncovering the contextual linguistic features that progressively anchor emerging subjective meanings (see Torres Cacoullos & Schwenter, 2005).

Yet, judging from the modest discriminating power of the model, the model is far from being perfect. The most important reason, as suggested by our analysis of mismatches between the observed connectives and the ones predicted by the model, is our operationalization of the subtle pragmatic and semantic functions with the help of very coarse-grained and discrete contextual features. The strategies of expressing subjectivity in discourse are very diverse and present a challenge for automatic identification. A comprehensive list of such markers and constructions is not available, to the best of our knowledge, and their grammatical structures vary greatly cross-linguistically.

Moreover, the mismatches can be explained by the usage itself. The semantics of causal connectives, similar to that of many other linguistic categories, has a prototypical structure, with the core functions and periphery (Stukker, 2005). When a causal connective is used non-prototypically, this may be due to the speaker's intention to change the construal of a causal relationships for rhetorical purposes (Stukker & Sanders, 2012). Such modulations can be detected only on the basis of a careful contextual analysis, as it is done, for example, in Sanders and Spooren (2013). Yet, we are convinced that the gains of our approach outweigh its limitations. Namely, our approach provides objective criteria for subjectivity and thus helps the linguist to avoid circularity in his or her semantic and pragmatic analyses.

Notably, the best predictors in our analysis are the usual suspects in analysis of register variation when it comes to the dimensions of involved and non-abstract non-technical communication as opposed to informational and abstract, technical discourse (Biber, 1988). One might wonder if the other features associated with involved communication in general, such as the use of general emphatics and discourse particles, hedges and amplifiers, or the features associated with abstract, technical discourse (e.g. conjuncts, past participial clauses and adverbial subordinators) might be relevant predictors in distinguishing between more subjective and objective uses of discourse markers. These interesting questions are left for future research.

Acknowledgements

The present study was a part of the project "Mapping the causative continuum: A multivariate typological investigation of causative constructions based on a multilingual parallel corpus", which was carried out by the first author at the University of Louvain in Louvain-la-Neuve (2013–2015) and which was funded by the Belgian research foundation F.R.S – FNRS. The first author is grateful to these two institutions for the generous financial, administrative and scientific support. All usual disclaimers apply.

References

Cem Akkaya, Janyce Wiebe and Rada Michalcea (2009). Subjectivity Word Sense Disambiguation. Proceedings of the 2009 Conference on Empirical Methods in Natural

Language Processing, pages 190–199, Singapore, 6-7 August 2009. Available online at http://www.aclweb.org/anthology/D09-1000 (last access 01.02.2017).

Juri Apresjan (1966) Analyse distributionelle des significations et champs sémantiques structures. *Langage*, 1(1):44–74.

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), 17–23 May, Malta.* European Language Resources Association (ELRA).

Douglas Biber (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511621024.

Anneloes Canestrelli (2013). *Small Words, Big Effects? Subjective versus Objective Causal Connectives in Discourse Processing*. LOT, Utrecht.

Elisabeth Couper-Kuhlen (1996). Intonation and clause-combining in discourse: The case of *because*. *Pragmatics*, 6(3):389-426.

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

Liesbeth Degand and Benjamin Fagard (2012). Competing connectives in the causal domain: French *car* and *parce que*. *Journal of Pragmatics*, Causal connectives in discourse: a cross-linguistic perspective, 44(2):154-68.

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

Christiane Fellbaum (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

John R. Firth (1957). A synopsis of linguistic theory 1930–1955. In: *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, Oxford.

Cecilia E. Ford (1994). Dialogic aspects of talk and writing: *because* on the interactive-edited continuum. *Text - Interdisciplinary Journal for the Study of Discourse*, 14:531-554.

Frank E. Harrell, Jr. (2001). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

Frank E. Harrell, Jr. (2015). rms: Regression Modeling Strategies. R package version 4.4-0. Available online at http://CRAN.R-project.org/package=rms. (last access 01.02.2017).

Zelig Harris (1954). Distributional structure. *Word*, 10(23):146–162.

Michael B. Kac (1972). Clauses of saying and the interpretation of because. *Language*, 48(3): 626-632.

Dan Klein and Christopher D. Manning (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 423-430. Available online at http://nlp.stanford.edu/manning/papers/unlexicalized-parsing.pdf (last access 01.02.2017).

Alistair Knott and Robert Dale (1994). Using linguistic phenomena to motivate a set of rhetorical relations. *Discourse Processes*, 18(1):35-62.

Alistair Knott and Ted J.M. Sanders (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135-175.

Phillip Koehn (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit, September 12–16, 2005, Phuket, Thailand*. Available online at http://www.statmt.org/europarl/ (last access 01.02.2017)

Ronald W. Langacker (1985). Observations and speculations on subjectivity. In John Haiman (ed.) *Iconicity in Syntax*, pages 109-150. John Benjamins, Amsterdam.

Beth Levin (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago Press, Chicago.

Natalia Levshina (2015). *How to Do Linguistics with R: Data exploration and statistical analysis*. John Benjamins, Amsterdam.

William Mann and Sandra Thompson (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text - Interdisciplinary Journal for the Study of Discourse* 8(3): 243–281.

Henk Pander Maat and Ted J.M. Sanders (2001). Subjectivity in causal connectives: An empirical study of language in use. *Cognitive Linguistics*, 12(3):247-273.

Ted Pedersen (2006). Unsupervised corpus-based methods for WSD. In Eneko Agirre & Philip Edmonds (eds.), *Word Sense Disambiguation: Algorithms and Applications*, pp. 133–166. Springer, New York.

Mirna Pit (2003). *How to express yourself with a causal connective. Subjectivity and causal connectives in Dutch, German and French.* Amsterdam, John Benjamins.

Mirna Pit (2006). Determining Subjectivity in Text: The Case of Backward Causal Connectives in Dutch. *Discourse Processes*, 41(2):151-74.

Mirna Pit (2007). Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast*, 7:53-82.

Emily Pitler and Ani Nenkova (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (ACLShort '09), pages 13-16. Association for Computational Linguistics, Stroudsburg, Pennsylvania.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/. (last access 01.02.2017)

William E. Rutherford (1970). Some observations concerning subordinate clauses in English. *Language*, 46(1):97-115.

Ted J.M. Sanders and Wilbert P.M. Spooren (2009). Causal categories in discourse – Converging evidence from language use. In Ted J.M. Sanders & Eve Sweetser (eds.) *Causal Categories in Discourse and Cognition*, pages 205–246. Mouton de Gruyter, Berlin.

Ted J.M. Sanders and Wilbert P.M. Spooren (2013). Exceptions to rules: a qualitative analysis of backward causal connectives in Dutch naturalistic discourse. *Text and Talk*, 33(3):377–398.

Ted J.M. Sanders & Wilbert 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):53-92.

Ted J.M. Sanders and Eve Sweetser (2009). Introduction: Causality in language and cognition – what causal connectives and causal verbs reveal about the way we think. In Ted Sanders & Eve Sweetser (eds.) *Causal Categories in Discourse and Cognition*, pages 1–18. Mouton de Gruyter, Berlin.

Ted J.M. Sanders, Kirsten Vis and Daan Broeder (2012). Project notes on the Dutch project DiscAn. *Eighth Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Pisa.

Mary J. Schleppegrell (1991). Paratactic because. *Journal of Pragmatics*, 16:323-367.

Wilbert P.M. Spooren and Liesbeth Degand (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241-66.

Wilbert P.M. Spooren, Ted J.M. Sanders, Mike Huiskes and Liesbeth Degand (2010). Subjectivity and Causality: A Corpus Study of Spoken Language. In John Newman & Sally Rice (eds.) *Empirical and Experimental Methods in Cognitive/Functional Research*, pages 256-270. CSLI Publications, Stanford, California.

Ninke Stukker (2005). Causality marking across levels of language structure. A cognitive semantic analysis of causal verbs and causal connectives in Dutch. PhD Diss. Universiteit Utrecht.

Ninke Stukker and Ted J.M. Sanders (2012). Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics*, 44(2):169-190.

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

Jörg Tiedemann (2012). Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012), Istanbul, Turkey*, pages 2214 – 2218. European Language Resources Association (ELRA). Available online at http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf (last access 01.02.17).

Rena Torres Cacoullos and Scott A. Schwenter (2005). Towards an operational notion of subjectification. In Rebecca T. Cover et Yuni Kim (eds.). *Proceedings of the 31st Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Prosodic Variation and Change*, pages 347-358. Berkeley Linguistics Society, Berkeley, California.

Teun A. Van Dijk (1979). Pragmatic connectives. *Journal of Pragmatics*, 3:447-456.

Ludwig Wittgenstein (1953). *Philosophical Investigations*. Blackwell, Oxford.

Sandrine Zufferey and Bruno Cartoni (2012). English and French causal connectives in contrast. *Languages in Contrast*, 12(2):232–250.