Expressing non-volitional causality in English

Jet Hoek, Radboud University Nijmegen &

Merel C.J. Scholman, Saarland University

Abstract

English *because* is assumed to be polysemous in that it can be used to mark causal relations in all domains. The current study examines this claim and explores the suitability of *because* to mark non-volitional content relations. In a parallel corpus study, we investigate how causal relations translated into Dutch using *doordat* (prototypically marking non-volitional causal relations), *omdat* (marking content relations), and *want* (marking epistemic and speech act relations) were originally expressed in English. The results show that while *omdat* and *want* are indeed typically translations of *because* in English, non-volitional *doordat* is not. A qualitative analysis reveals that non-volitional causality is more often expressed in English in a single discourse unit or using a connective restricted to the content domain. These findings have important consequences for the presumed domain generality of English *because* and call for a reconsideration of English translation recommendations for *doordat*.

Key words: volitionality, causality, Coherence relations, parallel corpus

1 Introduction

Sweetser (1990) classified causal relations into a seminal trichotomy: the content domain, the epistemic domain and the speech-act domain. Content causality is based on cause-and-effect relationships in the real world, as in (1), epistemic causality involves the speaker's reasoning, as in (2), and speech act causality expresses the motivation for a speaker's performing a particular speech act, as in (3).

(1) The delivery guy quit, because the restaurant got rid of their employee discount.

(2) The chef must love parsley, because it makes an appearance on every single dish.

(3) What are you doing tomorrow night, because I have an extra movie ticket.

Other researchers have further cut up the content causal relations into non-volitional and volitional content relations (e.g., Mann and Thompson, 1988; Pander Maat and Degand, 2001; Pander Maat and Sanders, 2000; Sanders et al., 1992; Stukker et al., 2008). Volitional content relations involve a thinking actor who is responsible for the action in the consequent of the causal relation. In (1), for instance, the quitting is a volitional action by the delivery guy; the reason for that action is given in the antecedent. In the non-volitional causal relation in (4), on the other hand, the cause-effect relation does not involve a volitional action by a thinking actor. In this case, the storm weakening the roof leads to the roof collapsing.

(4) The roof caved in, because the storm had severely weakened its structural integrity.

Languages seem to differ in the extent to which their causal connectives specialize in marking causal relations in specific domains. Spanish, for example, appears to have few causal connectives that are prototypically confined to a specific domain (Santana Covarrubias, 2019). Many causal connectives in Dutch (e.g., Canestrelli et al., 2013), German (e.g., Pit, 2003), and Mandarin Chinese (e.g., Li et al., 2016), however, do seem to have specific domain specializations. Dutch *omdat*, for example, has a slight preference for content relations and cannot be used in speech act relations; *want* is often used to express epistemic and speech act relations (Degand and Pander Maat, 2003; Pit, 2003). In addition, Dutch has a causal connective that is considered non-felicitous in all but non-volitional content relations: *doordat* (Degand and Pander Maat, 2003; Rooij, 1982).

Sweetser (1990) noted that *because*, like many other English connectives, seems to be polysemous in that it can be used in all three domains, as is illustrated by (1-4). However, English *because* being able to mark all these causal relations does not necessarily mean that it is also the preferred means with which to express causal relations from all domains. This study uses parallel corpus data to explore how non-volitional content relations are typically marked in English, departing from the observation that Dutch *doordat* is a causal connective specialized in non-volitionality.

We compare how causal relations translated into Dutch using *doordat*, *omdat*, and *want* were originally expressed in English. The primary suggested translation equivalent of all three of these Dutch connectives in translation dictionaries is *because* (e.g., Van Dale, 2020) and, as explained above, *because* is in theory compatible with all types of causal relations that *doordat*, *omdat*, and *want* prototypically mark. If *because* is equally preferred in non-volitional content, volitional content, and epistemic/speech act relations, we would expect the proportion of *because* as the source text equivalent of the Dutch connectives from our data set to be approximately the same for each of the three connectives. However, we find big differences, with the proportion of *because* being the source text equivalent of *doordat* in only a minority of cases. This suggests that *because* is not the preferred way of expressing non-volitional causality in English. By qualitatively analyzing the English source text equivalents of Dutch *doordat*, we create an inventory of alternative means English has to express non-volitional causality. Our findings imply that (i) English *because* might not be as domain-general as is generally presumed, (ii) non-volitional causality is often expressed through prepositional structures in English, and (iii) a reconsideration of English translation recommendations for *doordat* is called for. Section 4 discusses the implications of our chosen methodology and highlights directions for follow-up research.

2 Method: parallel corpus study

In a parallel corpus study, we investigate how causal relations translated into Dutch using *doordat*, *omdat*, and *want* were originally expressed in English. Note that *doordat* and *omdat* are both subordinating conjunctions, while *want* is a coordinating conjunction. If *because* is not used in non-volitional relations as frequently as commonly assumed, we expect to see a different pattern of original marking compared to *omdat* and *want*; *doordat* will instead be a translation of alternative signals or sentence structures in English. A qualitative analysis will reveal exactly which types of alternative signals and sentence structures are commonly used to express non-volitional relations in English.

We depart from Dutch translations and determine what marking occurs in the English source text to ensure we are looking at discourse originally uttered in English (see Levshina and Degand, 2017, for an approach based on the same principle). If we were to look at English translations of Dutch coherence relations, our results would be heavily influenced by the translation process; translations can differ from texts originally produced in a language in many ways, including in the marking of coherence relations (Cartoni et al., 2011). Note that these differences are mainly quantitative, not qualitative, in nature: a specific connective or construction can for instance be more frequent in translation (for example because of the influence of the source text language), but (semi-)professional translations do not tend to use connectives in ways that are unattested in original texts in that language (e.g., *since* to mark contrast relations).

2.1 Corpus data

We extracted translation data from two parallel corpora; the Europarl Direct corpus (Cartoni et al., 2013; Koehn, 2005) and the Ted Corpus Search Engine (TCSE: Hasebe, 2015). Both corpora contain semi-prepared, structured spoken data. The Europarl Direct corpus consists of the proceedings of the European Parliament from 1996 to 2012. We use only data from before 2004, since that is when the European Parliament starting making use of pivot languages, which means that after 2003 a direct translation from one language into the other can no longer be guaranteed. We use the directional version of the Europarl corpus for the same reason: all English fragments in the Europarl Direct corpus were originally uttered in English. The TCSE contains the transcripts of TED talks, which are highly structured speeches that are often minutely prepared and meant to provide targeted information on various topics or ideas.

In our data set, we included English data that were translated into Dutch, with

Table 1: Overview of final data set

	doordat	*omdat*	*want*	Total
TCSE	172	250	250	672
Europarl	236	248	239	723
Total	408	498	489	1395

the Dutch translation containing *doordat* (289 instances in the Europarl Direct corpus; 264 instances in the TCSE), *omdat* (250 instances per subcorpus) or *want* (250 instances per subcorpus) from both Europarl and TCSE. From this initial dataset, we removed any instances where the Dutch target connective occurred in a longer, fixed expression (e.g., 'dit komt doordat' *this is because*) and any non-connective uses of the search tokens (e.g., noun 'want' *mitten*). An overview of the final data set used for analysis is given in Table 1.[1]

The use of two corpora leaves us with a larger data set than only one corpus would have. In addition, it accounts for the slight genre difference between the two corpora: TCSE is less formal than Europarl. The degree of formality might have an effect on connective use in the source text and the translations to the target text (e.g., connectives such as 'consequently' might be more common in formal genres than informal genres). Including data from a more and a less formal corpus contributes to the generalizability of both the quantitative and qualitative results, since translations and, as a consequence, translation corpora can differ in several potentially relevant ways: there can for instance be differences in how free or literal translations are (e.g., Leppihalme, 1997) or in the overall degree of implicitation or explicitation of coherence relations (i.e., making an explicit relation implicit in translation or vice versa, see e.g., Hoek et al., 2017).

[1] The full annotated data set can be accessed at `https://tinyurl.com/yylfquqy`.

2.2 Annotating the English source text equivalents of *doordat*

The aim of the annotation effort was to categorize the way in which translated coherence relations expressed by *doordat* in Dutch were originally expressed in English. We conducted an inductive analysis of the data, identifying for each instance how the causality in the English source text was expressed. This analysis led to the creation of the following nine categories to describe what *doordat* is a translation of:[2]

- ***Because***: the source text contains *because*, and the Dutch target connective *doordat* is a direct translation of this English connective.

  EN We in Europe are lagging behind the US because at present the EU market is fragmented as a result of linguistic and cultural diversity. {ep-00-12-13}

  NL Wij blijven in Europa achter bij de VS doordat de EU-markt momenteel als gevolg van de diversiteit op het gebied van taal en cultuur versnipperd is.

- **Connective or cue phrase other than *because*:** *doordat* is a translation of a connective or cue phrase other than *because*, for instance *when, as,* or *while*.

  EN The single market is violated as the circulation of goods is impeded. {ep-02-04-11}

  NL De interne markt wordt ernstig verstoord doordat de distributie van goederen belemmerd wordt.

[2] Examples with ep-numbers (ep-year-month-day) were taken from the Europarl Direct corpus. Examples taken from the TCSE corpus are accompanied by their TCSE transcript ID.

- **Causal verb:** *doordat* is a translation of an English causal verb, such as *cause, result from,* or *create*.

  EN We want to avoid that, but we believe that such collisions could result from the failure of the Member States to implement the habitats directive and the birds directive also. {ep-01-05-16}

  NL Wij willen dergelijke situaties vermijden, maar wij weten tegelijkertijd dat die conflicten wel degelijk kunnen ontstaan doordat de lidstaten de richtlijnen inzake habitats en in het wild levende vogels niet ten uitvoer leggen.

- **Preposition:** the English original contained a preposition (e.g., *by, through, in, with*) either followed by a nominalization (e.g., "by *the use* of too many different bottle types") or by a gerund (e.g., "in *continuing* to raise this issue"). This includes complex causal phrases with prepositions, such as *because of, as the result of,* or *due to*.

  EN They are often hampered by the use of too many different bottle types. {ep-02-09-02}

  NL Het functioneren van deze systemen wordt vaak bemoeilijkt doordat teveel verschillende soorten flessen worden gebruikt.

- **Nominalization / gerund**: one or both of the segments were nominalized in the English original, but there is no preposition or causal verb that explicitly indicates the causal relation expressed by *doordat* in Dutch.

  EN It is truly an achievement of disabled people working and campaigning together. {ep-03-12-18}

NL Dit succes is echt bereikt doordat gehandicapten er samen de schouders onder hebben gezet. *'This achievement has been reached because disabled people have worked and campaigned together'*

- **Free adjunct:** one of the clauses from the Dutch target relation originally appeared as a free adjunct in English, with the causal relation not being explicitly marked. Unlike the previous category, the free adjunct is structured as a stand-alone clause.

  EN Being so few in number, they do not pose a road safety threat. {ep-96-10-23}

  NL Doordat ze weinig talrijk zijn, vormen ze hoe dan ook geen gevaar voor de verkeersveiligheid. *'Because they are so few in number, ... '*

- **Relative clause:** the two clauses related to each other by *doordat* in Dutch originally made up a relative clause construction in English, with the causal relation not being explicitly marked.

  EN All we have actually seen is a delay of three months from the report by Mr Elles, which was sent back to committee in December despite the Socialist wishes. {ep-99-03-22}

  NL Het enige wat we hebben zien gebeuren, is dat het verslag van de heer Elles drie maanden vertraging opliep doordat het in december werd terugverwezen naar de commissie, in weerwil van de socialisten. ' *..., because it was sent back in December ... '*

- **Implicit:** the causal relation has been made explicit in the Dutch translation, but was implicit in the English original and appeared as two independent, juxtaposed clauses.

EN The coffee is hot, the liquid is sterile. {TCSE-845}

NL Doordat de koffie heet is, is de vloeistof ook steriel. *'Because the coffee is hot, . . . '*

2.3 Annotation procedure

For every Dutch connective instance in the data set, the authors annotated the English source text equivalent using the qualitative coding scheme. A random subset of the data (50 Europarl and 50 TCSE relations containing *doordat*) was double-coded by both annotators to determine inter-annotator reliability. Inter-annotator agreement was high: 90%; κ=.88. Disagreements were discussed and resolved, and the remainder of the data was then single-coded, with one author annotating instances from Europarl and the other annotating instances from TCSE.

2.4 Data analysis

We present the data both per corpus and for the whole data set combined and analyze the results both quantitatively and qualitatively. The quantitative analysis tests whether there is a difference in the proportion of instances in which the Dutch connectives were translations of *because*. This is done to establish whether *because* is indeed the preferred marker to express non-volitional content causality in English, similar to its usage in volitional content relations. The data were modeled using linear regression models in the statistical software R (R Core Team, 2020). We used likelihood ratio tests to determine the significance of fixed effects, comparing the fit of the model to that of a model without the fixed effect. Categorical predictor variables were deviation coded and follow-up pairwise comparisons were obtained using a subset of the data that contained only the relevant conditions (if necessary, predictor variables were re-centered).

The purpose of the qualitative analysis is to get an overview of the different ways to express non-volitional causality in English. Since we do not have hypotheses about the distribution of the different source text equivalents or about any differences between the two corpora, we do not statistically test these data.

3 Results

3.1 Quantitative analysis: *because* as source text equivalent of *doordat*, *omdat*, and *want*

Table 2 shows the proportion of instances in which the Dutch connectives were translations of *because*.

We modeled the binary dependent variable of *because* as source text equivalent or not in a linear model, with corpus and NL-connective as predictor variables. The interaction between corpus and NL-connective was significant ($p < .001$).

Table 2: Raw count and percentage of *doordat*, *omdat*, and *want* relations with *because* as source text equivalent.

	doordat	*omdat*	*want*
TCSE	44 (26%)	209 (84%)	227 (91%)
Europarl	33 (14%)	162 (65%)	137 (57%)
Total	78 (19%)	371 (75%)	364 (74%)

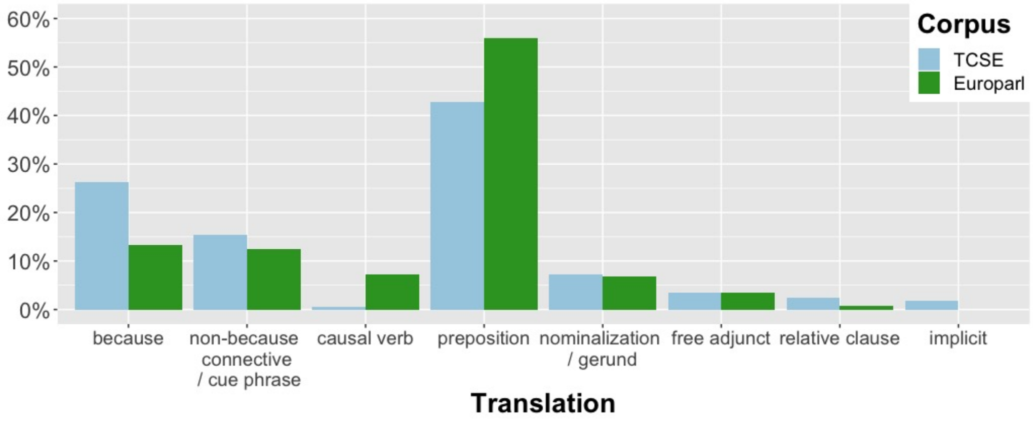

Figure 1: Proportion of English source text equivalents of *doordat*, per corpus.

To interpret this interaction, we performed follow-up analyses on the data for each corpus separately. In both corpora, there was a main effect of condition ($p < .001$ in both). Pair-wise comparisons reveal that in the Europarl data, the proportion of *because* equivalents was smaller for *doordat* than for both *omdat* ($\beta = -0.51, SE = 0.04, t = -13.48, p < .001$) and *want* ($\beta = -0.43, SE = 0.04, t = -11.02, p < .001$), but the difference between *omdat* and *want* was not significant ($\beta = 0.08, SE = 0.04, t = 1.82, p = .070$). In the TCSE data, the proportion of *because* equivalents for *doordat* was also smaller than for both *omdat* ($\beta = -0.57, SE = 0.04, t = -14.46, p < .001$) and *want* ($\beta = -0.65, SE = 0.04, t = -18.18, p < .001$), but in addition, *want* was more often the translation of *because* than *omdat* ($\beta = 0.07, SE = 0.03, t = 2.42, p = .016$). This difference will be reflected on in the Discussion.

3.2 Qualitative analysis: source text equivalents of *doordat*

Figure 1 and Table 3 show the English source text equivalents of *doordat* in the two corpora, for all instances not marked by *because*.

Table 3: English source text equivalents of *doordat*, per corpus. Raw counts and percentages.

	Because	Non-*because* connective	Causal verb	Preposition	Nominal./ gerund	Free adjunct	Relative clause	Implicit	Total
TCSE	44 (26%)	27 (16%)	1 (1%)	72 (42%)	12 (7%)	6 (4%)	4 (2%)	3 (2%)	169
Europarl	33 (14%)	29 (12%)	17 (7%)	131 (56%)	16 (7%)	8 (3%)	2 (1%)	0 (0%)	236
Total	77 (19%)	56 (14%)	18 (4%)	203 (50%)	28 (7%)	14 (3%)	6 (2%)	3 (1%)	405

In both corpora, *doordat* was most often the translation of a prepositional phrase construction. In the majority of cases (58%), the preposition used was *by*, as in (5).

(5) Bats are also threatened in the U.S. **by** their attraction to wind farms. {TCSE-1605}

The fragment in (5) states that bats' innate attraction to wind farms (for reasons apparently still largely unknown) results in them being threatened (because the wind turbines kill them). Bats' attraction to wind farms is not volitional, and the construction used underlines this. It should be noted that the antecedent in the English original of this causal relation, unlike in the Dutch *doordat* translation, is nominalized. Other reoccurring prepositions were *with* (10%), as in (6), *in* (7%), as in (7), and *through* (6%), as in (8).

(6) **With** data almost doubling every year, within the next two decades, we may even reach the point for the first time in history where we've discovered the majority of the galaxies within the universe. {TCSE-2069}

(7) The fishing industry has paid too high a price **in** having to share its abundant fishing grounds off the west coast of Ireland with greedy neighbours. {ep-02-01-17}

(8) Iceland and Norway, **through** not being Member States of the European

Union, will be excluded from discussions on the future development of the free movement area if it were to be incorporated into the Union treaties. {ep-96-06-19}

The second most frequent category was the connective *because*, accounting for 19% of all *doordat*-translations. The fact that *because* ranks second validates the prior observation that it is possible for *because* to mark non-volitional causal relations. It also indicates that *because* is a more frequent source for *doordat* than a non-*because* connective. Nevertheless, *because* is a less preferred means to mark non-volitionality than a prepositional phrase construction.

The third most frequent category was a connective or cue phrase other than *because*. Within this category, *as* was used most often (34%), see (9) for an example.

(9) *That meant it [the oil] went down the drain.* That has resulted in blocked sewers and other environmental problems **as** congealed fats and oils interfere with the infrastructure below our streets taking away the waste water. {ep-02-09-23}

English *as* is often used in relations where two events coincide (e.g., Webber et al., 2019). This can result in a relation that is mainly temporal in nature, but when two events overlap or follow each other in close succession, causality is often inferred. This also seems to be the case in (9). *When*, another connective found to be translated by *doordat* in our data set relatively often (11%), also signals simultaneity. Temporal relations, unlike causal relations, exclusively occur in the content domain (e.g., Crible, 2018; Hoek et al., 2019; Evers-Vermeul et al., 2017; Sanders et al., 1992). A third reoccurring cue phrase was *in that* (13%), see (10).

(10) Your freedom has been enlarged **in that** now you have a free area of movement throughout the Schengen area. {ep-01-04-03}

Although *in that* does not appear to have been included in many studies on the marking of coherence relations, this cue phrase also seems to be restricted to use in the content domain.

None of the other individual categories were particularly frequent in our data set, but what is noteworthy is that the majority of English source text fragments do not contain a coherence relation consisting of two separate discourse segments (typically clauses). The categories *preposition*, *causal verb*, and *nominalization / gerund* together make up 57% of all English source text equivalents of relations marked by *doordat*, and 74% of the English source text fragments for *doordat* relations that did not use *because*. Typically, these constructions would be considered single discourse units (see Hoek et al. (2018) for a discussion).

4 Discussion & conclusion

The current study investigated how causal relations translated into Dutch using *doordat*, *omdat*, and *want* were originally expressed in English, with a specific focus on exploring the suitability of *because* to mark non-volitional content relations. The quantitative analysis of our data, which contained instances from two semi-prepared spoken corpora, revealed an interaction effect indicating a difference between the two corpora in the proportion of *because* as source text equivalent of the three Dutch connectives. This effect was driven by *want* and *omdat*: in the Europarl Direct corpus, *want* and *omdat* were translations of *because* equally often, but in the TCSE data, *want* was more often the translation of *because* than *omdat* was. We had not predicted such an interaction. This difference between Europarl Direct and TCSE may be due to the degree of formality of the two corpora. While both contain spoken discourse that has to a large extent been prepared, TED talks, which are aimed at telling a story about a specialized topic to a broad audience, constitute a more informal genre than the proceedings of the European parliament, which consist of politicians talking to other politicians about legislation in a official setting. Dutch *want* is more frequent in informal than in formal registers (Sanders and Spooren, 2015). This may have resulted in a stronger 1:1 relationship between *because* and *want* than between *because* and *omdat*. In addition, there appears to be more variation in the Europarl translations than in the TED talk translations, which seem to stick somewhat more closely to the source text structure. This can, for instance, be seen from Table 2, which shows that the proportion of *because* as source text equivalents was lower for all three connectives in the Europarl data than in the TCSE data. The difference in translation variation between the two corpora might be due to the translation style and technique used by the translators. This difference may also have contributed to the observed interaction.

More central to the question of this study, however, the quantitative analysis also showed that in both corpora, *doordat* was much less often the translation of English *because* than *omdat* and *want*. Such a difference would not be expected if *because* is equally favored to mark non-volitional causal relations as it is to mark volitional causal relations or causal relations from the epistemic or speech act domain. Our results therefore contest a complete domain generality of English *because*.

A qualitative analysis of the English source text equivalents of *doordat* provided an overview of alternative ways in which non-volitional causality can be expressed in English. Two main observations can be made. First of all, non-volitional relations were often expressed in English using a connective or cue phrase that specifies the content aspect of the coherence relation, but leaves the causal aspect of the relation underspecified. *When*, for example, prototypically marks content temporal relations, but often also allows for a causal reading. Because they clearly mark the relation as holding in the content domain, the non-*because* connectives found in the English part of our data set might be very suitable to mark non-volitional causal relations despite being ambiguous in terms of causality: language users readily interpret relations as causal if this is a plausible option (Sanders, 2005). This suggests that a connective such as *when* (which underspecifies the causal aspect, but specifies the content dimension of a relation) does not underspecify non-volitional causality more than a causal connective such as *because* (which can also mark volitional content, epistemic, and speech act relations). In fact, using a non-causal connective specific to the content domain might even be a better cue for non-volitional causal relations if language users have a preference for inferring volitional content, epistemic, or speech act causality over non-volitional causality (see also Crible et al., 2019, for related work on underspecification). While to our knowledge no linguistic studies have investigated this question, there is evidence from psychological studies on causal inference that suggests that people prefer to attribute causal responsibility to human agents over other types of causes (e.g., Alicke, 1992; Johnson and Keil, 2014; Lagnado and Channon, 2008). Determining whether language users have a preference for inferring other types of causal coherence relations over non-volitional content relations might be a fruitful topic for future research.

The second observation that can be made based on the quantitative analysis is that the English fragments often expressed the non-volitional causal relationship in a construction consisting of only a single discourse segment. In the most frequent category *preposition*, for example, one of the parts of the causal relation, typically the antecedent, is often nominalized and the causal relationship is then expressed in a single grammatical clause. In their discussion of the linguistic realization of causality, Stukker et al. (2008) argue that Sweetser's trichotomy is relevant not just to causal relationships expressed in coherence relations, but also to causality expressed using different linguistic means, for instance causal verbs or prepositions (see also Degand, 2000). They make the observation, however, that epistemic causal relations cannot be expressed using an inter-clausal construction, whereas other types of causal relations can be expressed in an inter-clausal relation. Attempting to express an epistemic causal relation using a causal verb, for example, results in ungrammaticality or in a sentence that can only be interpreted as a causal content relation. This is illustrated in Examples (11)-(14). The volitional causal relation between two full clauses in (11) can be reformulated as the volitional inter-clausal causal construction expressed in (12). The same cannot be done for the epistemic causal relation in (13); expressing it as a inter-clausal causal construction, as in (14), forces a causal content reading (instead of an epistemic one).

(11) The restaurant got rid of their employee discount. As a result the delivery guy quit.

(12) The restaurant getting rid of their employee discount caused the delivery guy to quit.

(13) The restaurant is always full, so the food must be good.

(14) ?The restaurant always being full caused the food to be good.

Stukker et al. (2008) point out that epistemic relations consist of two independent propositions that the speaker then relates to each other in their utterance (note that the same holds for speech act relations). Since discourse segments generally correspond to individual propositions (e.g., Hoek et al., 2018), it makes sense that epistemic and speech act relations are difficult to express within a single discourse segment. Stukker et al. (2008) frame this as an instantiation of grammatical differences corresponding to conceptual differences, following a key insight from cognitive linguistics (e.g., Langacker, 1987). As explained in the introduction, non-volitional causality refers to causal relations that happen without the interference of any thinking actor, which places non-volitional content relations on the other side of the spectrum from epistemic and speech act relations. In the absence of a specialized non-volitional causal connective, expressing causality in a single clause might thus be an alternative signal for non-volitionality.

The findings reported in this paper give rise to several follow-up questions. First, we consider the status of prepositions as discourse markers. Prepositions and prepositional phrases are commonly considered potential discourse connectives, with complex prepositional phrases (*because of, as the result of, due to*) typically being strong explicit discourse markers. Other types of prepositions, such as *by* and *in*, are generally only considered discourse markers when they take clausal complements (Carlson and Marcu, 2001; Webber et al., 2019). Some complex prepositions, such as *as the result of* and *due to*, are unambiguously non-volitional. However, many other prepositions, including the ones that were frequently the source text equivalent of *doordat*, are underspecified with regard to volitionality. More work is needed to further understand the function of prepositions in coherence relations, the semantic nature of these prepositions, and the distribution of their usage in volitional versus non-volitional relations. When are prepositions preferred over other discourse markers to express non-volitional causality specifically, and coherence relations in general? (See Degand, 2000, for an investigation of this question in Dutch) The findings reported in the current contribution also highlight the need for experimental work investigating the role of prepositions in processing, a point raised by Degand (2000): do prepositional phrases play an important role in text processing similar to other discourse markers, or does their intra-clausal status imply that they are less important in terms of processing?

Second, the finding that non-volitional causality is often expressed in a single discourse unit in English raises the question of how non-volitional causality is typically expressed in other languages, including Dutch. The results might seem to indicate that there is a difference between English and Dutch, in that non-volitional causality is expressed more often in a single discourse unit in English but in segment-pairs in Dutch. However, this cannot be concluded based on the data available in the current study, because we specifically selected the data to include only segment-pairs in Dutch. As Degand (2000) and Stukker et al. (2008) have shown, non-volitional causality can also be expressed in a single discourse unit in Dutch. The question remains whether these languages differ with respect to the distribution of inter-clausal and intra-clausal non-volitional relations. Given that Dutch has a specialized connective to express non-volitional causality whereas English does not, it is possible that this type of causality is expressed in segment-pairs more often in Dutch than in English.

A note should be made regarding the correspondence between *because* and *omdat*. *Omdat* is prototypically a content connective, but it is underspecified for its volitionality. In other words, like *because*, *omdat* can be used to express non-volitional relations as well as volitional relations. It is possible that our dataset of *omdat* relations therefore contains non-volitional relations as well. In cases where the causal relation is marked by *because* and is underspecified in its volitionality, *omdat* might provide a better translation alternative than *doordat*, as it preserves the ambiguity of the original utterance. In such cases, *doordat* would involve a semantic-pragmatic narrowing that translators might not want to commit to. To address this, it would be interesting to study the translation choices of *because* relations that have been annotated for their volitionality.

The generalizability of our results to other types of discourse also deserves some consideration. In our study, we made use of parallel corpora: TCSE and Europarl Direct, both of which contain spoken discourse. Spoken and written texts are produced differently, which impacts various discourse factors, including discourse marking. For example, coherence relations are more often marked explicitly in spoken data compared to written data (Rehbein et al., 2016). Nevertheless, we note that both TCSE and Europarl contain semi-prepared, structure discourse. They can, therefore, be considered to have characteristics resembling both regular spoken data and written data. We hypothesize that the findings discussed in the current paper are generalizable to written text as well; however, this would need to be further investigated in follow-up research.

Finally, we consider the implications of our chosen methodology. The aim of the current study was to explore the marking of non-volitionality in English, which was done by analyzing translations of English data into Dutch. There are two potential pitfalls to this approach. First, *doordat* is typically considered a non-volitional connective, but its usage in our corpus was not always unambiguously non-volitional. Consider the following example.

(15) EN I was the guy beaten up bloody every week in the boys' room, until one teacher saved my life. She saved my life by letting me go to the bathroom in the teachers' lounge.

NL Ik was de kerel die elke week in de jongenskamer bloedig in elkaar geslagen werd, totdat een leerkracht mijn leven gered heeft. Ze heeft mijn leven gered doordat ik naar het toilet mocht gaan in de lounge van de leerkrachten.

It could be argued that the teacher did not necessarily intend to save the speaker's life by letting him go to the teachers' bathroom, but she likely did grant permission as a volitional action to protect the speaker from bullies. The English source therefore does not seem unambiguously non-volitional. The occurrence of such ambiguous *doordat*-items was not very frequent in our corpus, and they could therefore merely be deviances from the prototypical usage of *doordat* as a result of the translation process (possibly as a result of 'by'-constructions relatively frequently being translated into 'doordat'-relations). However, this does deserve further consideration, for example by testing whether such ambiguous cases are more frequent in translated text compared to original Dutch text or by looking at how causal 'by'-constructions are typically translated.

A second potential pitfall of this approach is that there might be different types of non-volitional causal relations (not translated using *doordat*) that are not included in this study. To address this, as well as the genre generalizability point raised above, a follow-up study could focus on annotating English causal relations for the volitionality/non-volitionality distinction, for example by adding an annotation layer to causal instances from a large discourse-annotated corpus such as the Penn Discourse Treebank (Webber et al., 2019).

5 Acknowledgments

The second author was supported by by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding”. We are very grateful to four anonymous reviewers for their valuable comments.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63(3):368–378.

Canestrelli, A. R., Mak, W. M., and Sanders, T. J. M. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive processes*, 28(9):1394–1413.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:1–56.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27(1):23–42.

Cartoni, B., Zufferey, S., Meyer, T., and Popescu-Belis, A. (2011). How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 78–86. Portland, OR, USA.

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*, volume 286. John Benjamins Publishing Company.

Crible, L., Abuczki, Á., Burkšaitienė, N., Furkó, P., Nedoluzhko, A., Rackevičienė, S., Oleškevičienė, G. V., and Zikánová, Š. (2019). Functions and translations of discourse markers in TED Talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

Degand, L. (2000). Causal connectives or causal prepositions? Discursive constraints. *Journal of Pragmatics*, 32(6):687–707.

Degand, L. and Pander Maat, H. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. *LOT Occasional Series*, 1:175–199.

Evers-Vermeul, J., Hoek, J., and Scholman, M. C. J. (2017). On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse*, 8(2):1–20.

Hasebe, Y. (2015). Design and implementation of an online corpus of presentation transcripts of TED talks. *Procedia: Social and Behavioral Sciences*, 24:174–182.

Hoek, J., Evers-Vermeul, J., and Sanders, T. J. M. (2018). Segmenting discourse:

Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.

Hoek, J., Evers-Vermeul, J., and Sanders, T. J. M. (2019). Using the Cognitive approach to Coherence Relations for discourse annotation. *Dialogue & Discourse*, 10(2):1–33.

Hoek, J., Zufferey, S., Evers-Vermeul, J., and Sanders, T. J. M. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131.

Johnson, S. G. B. and Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, 143(6):2223–2241.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86. Phuket, Thailand.

Lagnado, D. A. and Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Leppihalme, R. (1997). *Culture bumps: An empirical approach to the translation of allusions*, volume 10. Multilingual Matters.

Levshina, N. and Degand, L. (2017). Just because: In search of objective criteria of subjectivity expressed by causal connectives. *Dialogue & Discourse*, 8(1):132–150.

Li, F., Sanders, T., and Evers-Vermeul, J. (2016). On the subjectivity of mandarin reason connectives: Robust profiles or genre-sensitivity. *Genre in language, discourse and cognition*, 33:15.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Pander Maat, H. and Degand, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics*, 12(3):211–246.

Pander Maat, H. and Sanders, T. J. (2000). Domains of use or subjectivity? the distribution of three Dutch causal connectives explained. *Topics in English Linguistics*, 33:57–82.

Pit, M. (2003). *How to express yourself with a causal connective: Subjectivity and causal connectives in Dutch, German and French*, volume 17. Rodopi.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rehbein, I., Scholman, M. C. J., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 23–28, Portoroz, Slovenia.

Rooij, J. d. (1982). 'omdat' en 'doordat' in het Nederlands. *Nieuwe Taalgids*, 75(4):329–342.

Sanders, T. J. M. (2005). Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, pages 105–114, Toulouse, France.

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

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

Santana Covarrubias, A. (2019). *Is* porque *more like* because *or like* omdat*?: An exploration of causality and subjectivity in Spanish backward causal connectives*. PhD thesis, Utrecht University.

Stukker, N., Sanders, T. J., and Verhagen, A. (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.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, volume 54. Cambridge University Press.

Van Dale (2020). Van Dale NL-EN dictionary. `https://www.vandale.nl`, accessed 30-10-2020.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). *The Penn Discourse*

Treebank 3.0 annotation manual. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.