Jet Hoek*, Jacqueline Evers-Vermeul and Ted J.M. Sanders

Segmenting discourse: Incorporating interpretation into segmentation?

DOI 10.1515/cllt-2016-0042

Abstract: Discourse segmentation is an important step in the process of annotating coherence relations. Ideally, implementing segmentation rules results in text segments that correspond to the units of thought related to each other. This paper demonstrates that accurate segmentation is in part dependent on the propositional content of text fragments, and that completely separating segmentation and annotation does not always yield text segments that correspond to the text units between which a conceptual relationship holds. In addition, it argues that elements belonging to the propositional content of the discourse should necessarily be included in the segmentation, but that inclusion of other text elements, for instance stance markers, should be optional.

Keywords: segmentation, discourse structure, coherence relations, corpus annotation, stance marking

1 Introduction

Annotated corpora have become increasingly valuable resources for the study of language. They allow us to investigate the functions of linguistic forms, to study the linguistic realization of particular functions, to test linguistic theories, and to develop new ones. Many annotated corpora contain annotations at the levels of syntax, semantics, and morphology, as well as the annotation of lexical features. In addition, the last two decennia have seen the rise of corpora annotated at the level of discourse. At the discourse level, one of the things that are annotated is the coherence within a text. By annotating the *coherence relations* within a discourse, it becomes apparent how idea units in a text are related to each other, e. g., are they causally related, contrasted, part of an enumeration, etc.?

A coherence relation can be defined as "an aspect of meaning of two or more discourse segments that cannot be described in terms of the meaning of

***Corresponding author: Jet Hoek,** Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK, Utrecht, The Netherlands, E-mail: j.hoek@uu.nl
Jacqueline Evers-Vermeul: E-mail: j.evers@uu.nl, **Ted J.M. Sanders:** E-mail: t.j.m.sanders@uu.nl, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK, Utrecht, The Netherlands

the segments in isolation," or, in other words, the meaning of a coherence relation is "more than the sum of its parts" (Sanders et al. 1992: 2). In line with Sanders et al. (1992), we consider coherence relations to be a feature of the cognitive representation of a text, rather than a feature of its linguistic realization (see also Hobbs 1979; Kehler 2002; Kehler et al. 2008). This definition assumes coherence relations to hold between the idea units that readers or listeners construct on the basis of the linguistic input. If we want to make claims about the nature of such coherence relations, it seems important that the text segments indicated to feature in a coherence relation correspond to the idea units that are related to each other in the cognitive representation of a discourse.

The notion of *idea unit* is not a clearly delineated linguistic category, unlike for instance the notions of *subject* and *object* in syntactic annotation. While a lot of attention is paid to creating relation inventories specifying the types of relations that can hold between segments (cf. Asher and Lascarides 2005; Carlson and Marcu 2001; Hobbs 1990; Kehler 2002; PDTB Research Group 2007; Reese et al. 2007; Sanders et al. 1992; Wolf and Gibson 2005), much less theoretical consideration has been given to the exact characteristics of the segments and the way in which they are structured in a discourse (notable exceptions are Matthiessen and Thompson 1988; Polanyi 1988; Schilperoord and Verhagen 1998; Verhagen 2001). Many approaches to discourse annotation have taken the clause as the basis for identifying segments, although annotation frameworks are not uniform in this respect and exceptions or addenda to the clause as basic unit also differ between approaches. The variability between annotation approaches in their operationalization of *idea units* and the (syntactic) rules on the basis of which they identify discourse segments has consequences for the eventual annotation of the coherence relations that are annotated in a corpus, and can consequently affect theories and conclusions formulated on the basis of the data.

Taking a syntactic structure as the basis for segmentation rules makes the segmentation process relatively objective and enables annotators to treat segmentation and annotation as separate steps. However, it appears that strict application of these segmentation rules does not always result in segmentation that does justice to the interpretation of a fragment. As will be illustrated in this paper, applying conventional segmentation rules may produce segments that are too small, in which case it does not include an entire unit of thought, or too big, in which case only part of the segment connects to the adjacent segment. Alternatively, certain inferred coherence relations may not be segmented at all, as is often the case for coherence relations that are embedded in syntactic constructions such as complement clauses or restrictive relative clauses.

The current paper will theoretically approach discourse segmentation and focus on two issues concerning segmentation that were proposed by Mann and Thompson (1988) in their introduction of Rhetorical Structure Theory, but that have been implemented in many other discourse annotation approaches as well: the treatment of segmentation and annotation as a two-step process, which prevents the circularity of a process in which annotation and segmentation are intertwined (Taboada and Mann 2006), and the completeness constraint, which poses that the segmentation of a text has to include all elements of that text. In this paper, we consider segmentation to be accurate when the segments correspond to the idea units that are related to each other. We propose that accurate segmentation is at least in part dependent on the propositional content of text fragments, and that completely separating segmentation and annotation, as well as adhering to the completeness constraint, can be at the expense of the quality of the segmentation.

After establishing the clause as the syntactic basis for the identification of discourse segments, we discuss fragments, mainly from the Europarl Direct corpus (Cartoni et al. 2013; Koehn 2005, all fragments were originally uttered in English), that present segmentation difficulties.[1] We focus specifically on fragments with complement structures, sentential adverbs, restrictive relative clauses, and stance markers. Building on a proposal for discourse segmentation by Schilperoord and Verhagen (1998), we present an approach to segmentation that results in text segments that correspond to the text units between which a conceptual relation is presumed to hold. As a means of determining whether segments actually represent the units of thought related to each other, we will make use of paraphrases (see e. g. Sanders 1997) throughout this paper. Comparing different paraphrases of the same relation can help determine between which idea units a coherence relation holds. The idea units that feature in the best paraphrase should be represented by the text segments.

2 The clause as the basis for identifying discourse segments

The smallest unit that can function as a discourse segment is often taken to be the grammatical clause (cf. Evers-Vermeul 2005; Mann and Thompson 1988; Sanders and van Wijk 1996; Wolf and Gibson 2005), which can be defined as a

1 The ep-number (ep-year-month-day) following each fragment refers to the corpus file from which the fragment was retrieved.

unit headed by a verb. This rule was introduced by Mann and Thompson (1988) as a theory-neutral approach to the classification of a text into segments. Considering the definition of coherence relations we employ, selecting the clause as the minimal unit for discourse segments seems appropriate, since the clause is the smallest grammatical unit that can function meaningfully in isolation.

Requiring discourse segments to be minimally clauses eliminates prepositional phrases as discourse units: (1a) is considered to be a single discourse segment, even though its meaning is similar to (1b), which consists of two segments between which a causal relation holds.

(1a) *Their fears and uncertainties have been compounded because of their belief that immigrants will pose a threat to future employment.* {ep-01-02-14}

(1b) [*Their fears and uncertainties have been compounded*] *because* [*they believe that immigrants will pose a threat to future employment*.]

(1c) *Their fears and uncertainties have been compounded because of their beliefs*.

Although sentences with prepositional phrases can be very similar to coherence relations, as in (1a), this is often not the case. It is, for example, not possible to paraphrase (1c) in a way that resembles a coherence relation, since the prepositional phrase contains only a simple noun phrase.

In addition, employing the criterion that discourse segments have to be clauses eliminates the possibility of considering fragments such as (2a), in which a verb, in this case *cause*, signals causality, as coherence relations. Even though (2a) resembles the causal relation in (2b) in meaning, it is only one clause and does therefore not contain a coherence relation.

(2a) *In the year 2000 smuggling of tobacco caused losses of GBP 3.8 million to the British Exchequer*. {ep-02-02-05}

(2b) [*In the year 2000 the British Exchequer lost GBP 3.8 million*,] *because* [*tobacco was smuggled into the country*.]

One of the advantages of taking the clause as the basis for identifying units and not considering prepositional phrases and the objects of causal verbs to be independent discourse units is that it allows us to systematically distinguish between intra- and interclausal ways of expressing something, for instance the causality in the above examples (cf. Degand 1996; Stukker et al. 2008).

In theory, identifying clauses should be fairly straightforward. However, clauses need not be complete, and although it is commonly agreed upon that clauses with ellipted elements can be discourse segments, there is less consensus on when exactly a clause should no longer be considered to be a discourse segment. Two types of approaches for assigning discourse segment status can be identified: defining what can still be considered a clause, or defining what cannot be considered a clause.

Both Sanders and van Wijk (1996) and Carlson and Marcu (2001) provide guidelines for what can still be considered a clause. Sanders and van Wijk (1996: 126), for example, allow only one "major constituent" to be contracted. Carlson and Marcu (2001: 12) allow the subject, auxiliary verb, and adverb of a clause to be ellipted, and even the main verb, provided that "there are strong rhetorical cues marking the discourse structure." Neither approach would allow the segmentation in (3).

(3) *The virus harms cold-blooded animals. It does not replicate at temperatures above 25° centigrade and* [*would*,]$_{S2a}$ *if* [*present in fish for human consumption*,]$_{S1}$ [*be inactivated when ingested*.]$_{S2b}$ {ep 00-03-01}

In the first segment (S_1) of the coherence relation in (3), both the subject and the main verb have been left out, without there being any "strong rhetorical cues."[2] If we were to adhere to the segmentation guidelines provided by Sanders and van Wijk (1996) or Carlson and Marcu (2001), we would not be able to segment the conditional relation in (3). Not segmenting this relation seems overly conservative, since the segmentation in (3) seems very plausible and exactly captures the two segments related by the connective *if*. Not segmenting the conditional relation would lead to a crucial coherence relation missing from the final annotation of the fragment.

Pander Maat (2002: 41), on the other hand, proposes that multiple elements can be contracted in a sentence, as long as in addition to a connective there is also another phrase present between the non-contracted elements. Although it is not entirely clear how this guideline applies to (3), Pander Maat's segmentation rule appears to be primarily aimed at excluding the possibility of segmenting

2 Carlson and Marcu (2001: 12) do not give a concrete definition of a 'strong rhetorical cue,' but do provide the following example (in bold): "Back then, Mr. Pinter was **not only** the angry young playwright, **but also** the first to use silence and sentence fragments and menacing stares, almost to the exclusion of what we previously understood to be theatrical dialog." It not clear whether *if* is a strong enough rhetorical cue, since it only marks one of the discourse segments and is not as prominent as *not only … but also*.

coordinated nouns, which is not the case in (3). Wolf and Gibson (2005) also seem to prioritize excluding coordinated elements, since they state that they do not consider conjoined nouns in a noun phrase or conjoined verbs in a verb phrase to be separate discourse segments. If we were to follow Wolf and Gibson's (2005) guidelines, (3) could be segmented, since there is no coordination within a phrase.

The type of elision that is illustrated in (3) is not exclusive to conditional relations, but can for instance also be found in segments preceded by *although* or *but*.

(4) *Although* [*no expert*,]$_{S1}$ [*I would certainly support the calls for all prisoners of conscience to be freed, in Syria and elsewhere*.]$_{S2}$ {ep-02-06-13}

(5) … [*parties can choose their own contract law in relation to these particular contracts*,] $_{S1}$ *but* [*not their own winding-up proceedings law*.] $_{S2}$ {ep-01-01-05}

As in (3), both the subject and the finite verb have been left out of the clauses following the connective in (4) and (5). Strikingly, in all three fragments, the elided verb is a copula verb. The elements following *although* in (4), *but* in (5), and *if* in (3), are therefore all subject complements and, as such, part of the predicate. If we slightly adjust our definition of a clause from "a structure headed by a verb" to "a structure containing a predicate," we could formulate the tentative segmentation rule that structures can be discourse segments if they contain (at least part of) a predicate.

If we use the presence of a predicate, or parts of a predicate, as the criterion for discourse segment status, we automatically include non-finite clauses as potential discourse segments. Most discourse annotation approaches seem to indeed allow segments of coherence relations to be non-finite: this is explicitly stated in some manuals (e. g., Carlson and Marcu 2001: 6–7) or it can be concluded on the basis of provided definitions and examples (e. g., Mann and Thompson 1988; PDTB Research Group 2007; Reese et al. 2007). At the same time, using the presence of a predicate for assigning discourse segment status excludes structures such as prepositional phrases and non-clausal adverbials or modifiers from receiving discourse segment status, which is also in line with most discourse annotation manuals (e. g., Carlson and Marcu 2001; Mann and Thompson 1988; PDTB Research Group 2007, but not Reese et al. 2007: 3).

Taking the predicate instead of the verb as the basis for assigning discourse segment status, prevents compound subjects from being segmented, since subjects are not part of the predicate. Segmenting coordinated nouns in subject position seems indeed something to avoid if discourse segments have to

correspond to a unit of thought. Fragment (6), for instance, expresses only one unit of thought, even though it contains a compound subject. The segmentation indicated in (6) does therefore not seem appropriate, which is signaled by the hashtag in front of the fragment.

(6) # [*The Commissioner*] *and* [*Mr Hatzidakis said that regional disparities will become twice as great*.] {ep-01-01-31}

Using the presence of a predicate as the basis for segmentation would, however, allow objects to be individual discourse segments. This appears to be too liberal, since it would also allow segmentations like the one in (7), despite the fact that the fragment expresses only one unit of thought: one group of people is being thanked for the same thing.

(7) # [*I want to thank the rapporteur*,] [*the Commissioner*] *and* [*other colleagues who are here tonight*.] {ep-97-11-18}

It seems therefore necessary to also include a rule resembling Pander Maat's (2002) or Wolf and Gibson's (2005) in order to prevent segmentation of coordinated structures within a single phrase. By adding this, we exclude segmentation of for instance coordinated nouns, as in (7), or coordinated verbs, as in (8). Amending our predicate-based segmentation rule with the rule that coordinated structures within a single phrase cannot be segmented would exclude segmentations like the ones in (7) and (8), but potentially allow the segmentation in (9).

(8) # [*I, therefore, would ask*] *and* [*request that this House support Amendment No 4*.] {ep-97-03-11}

(9) [*I want to congratulate Mrs van den Burg for an enormously well done job*] *and* [*the Commissioner for introducing this directive*.] {ep-02-11-20}

The fragment in (9) contains two direct objects, but they are, arguably, not coordinated within a single noun phrase. This results in the segmentation indicated in (9). Unlike (7) or (8), (9) appears to contain two separate idea units, which in this case are explicit speech acts: thanking Mrs. van den Burg for her great output on the one hand, and thanking the Commissioner for coming up with the initiative on the other. The segmentation in (9) seems therefore a more appropriate representation of the discourse structure than the

segmentations in (7) or (8). Whether or not two elements are conjoined within a single phrase, and, consequently, whether they should be considered to be independent discourse segments, can be left to the judgment of the annotators.

This section outlined the essential structural properties of discourse segments. The next section will establish another criterion a clause has to satisfy in order to have the status of discourse segment: *conceptual dependency*, which entails that if a clause is an integral part of another clause, it cannot be an independent discourse segment. After introducing the concept of conceptual dependency, we will discuss the consequences the conceptual dependency criterion has for the process of attributing discourse segment status to clauses, and, in turn, for discourse segmentation.

3 Conceptual dependency and the segmentation of embedded clauses

Clauses may satisfy all structural criteria outlined in Section 2 and still be excluded from having discourse segment status. The general rule that clauses can be discourse segments is often amended by a few exceptions. The clause types listed in (10) are for instance often denied the status of discourse segments:

(10)
i **Clausal complements** (*We saw that people wanted to dance*)
(Carlson and Marcu 2001; Evers-Vermeul 2005; Mann and Thompson 1988; Sanders and van Wijk 1996)
ii **Clausal subjects** (*Dancing is my favorite thing to do*)
(Carlson and Marcu 2001; Evers-Vermeul 2005; Mann and Thompson 1988; Sanders and van Wijk 1996)
iii **Restrictive relative clauses** (*Susan likes men who can dance*)
(Evers-Vermeul 2005; Mann and Thompson 1988; Reese et al. 2007; Sanders and van Wijk 1996; Schilperoord and Verhagen 1998; Verhagen 2001)
iv **Restrictive adverbial clauses** (*I am going to dance until the music stops*)
(Evers-Vermeul 2005; Pander Maat 2002; Renkema 2009; Schilperoord and Verhagen 1998)

Although Reese et al. (2007) do not specifically list clause types excluded from receiving discourse segment status, with the exception of restrictive relative clauses (p. 4), they do state that they do not allow segmentation of embedded

structures (p. 3). In practice, this means that at least clausal subjects and clausal complements are also not viewed as discourse segments in their annotation method.

Several approaches to discourse annotation include *attribution relations* in their relation inventory (cf. Carlson and Marcu 2001; Reese et al. 2007; Versley and Gastel 2013; Wolf and Gibson 2005). Attribution relations indicate who is responsible for the information in a fragment (cf. Pareti 2012), as in (11).

(11) *You also said that the budget should have the same discipline as national budgets*. {ep-99-09-14}

Attribution relations inherently assign discourse segment status to clausal complements. In (11), for instance, *you also said that* would be S_1 of the attribution, while *the budget should have the same discipline as national budgets*, a clausal complement, would be S_2. In order to be able to consider attribution relations as coherence relations, some annotation approaches have included exceptions to their segmentation rules (or rules for attributing discourse segment status) for fragments that contain communication verbs. Carlson and Marcu (2001: 7) for instance state that "normally, clausal complements are not considered to be EDUs [elementary discourse units – discourse segments]. We make exception to this in the case of clausal complements of *attribution verbs*" (original emphasis). However, neither Carlson and Marcu (2001) nor any of the other annotation approaches provide a comprehensive explanation for making exceptions to segmentation rules on the basis of verb semantics. The definition of coherence relations employed in this paper seems to exclude attribution relations as coherence relations: the meaning of an attribution construction as a whole is not *more* than the sum of its parts, and only one of the two "segments" of attribution relations can function meaningfully in isolation, namely the embedded clause. The importance of segments being able to function meaningfully in isolation for their status as discourse segments will be further elaborated upon in Section 3.1, in which we introduce the notion of conceptual dependency to explain why clausal complements and the other clause types listed in (10) are often excluded from being independent discourse segments.

3.1 Clausal complements

Schilperoord and Verhagen (1998) introduce the notion of *conceptual dependency* to explain why embedded clauses are often excluded from being

independent discourse segments, something they themselves do not strictly agree with:

> If a constituent of clause A is conceptually dependent on a clause B, B is an integral part of the conceptualization of A, and therefore not available as a separate discourse segment (cannot enter into a discourse coherence relation with A, or any other part of the discourse). (p. 150)

Matrix clauses that contain a clausal complement or a clausal subject are not complete without the complement or the subject and are therefore not conceptually independent. Noun phrases that are followed by a restrictive relative clause are also, for their conceptualization, dependent on the restrictive relative clauses: without the restrictive relative clause, the concept to which the noun phrase refers is usually underspecified. Since coherence relations are defined to hold between segments that can potentially be independent (Sanders et al. 1992), there can be no coherence relation between clausal complements, clausal subjects, or restrictive relative clauses, and their host clauses. Crucially, this definition of conceptual dependency assumes that it is the main clause that is dependent on the subordinate clause, instead of the other way around. The subordinate clause from (11), for example, could by itself be an independent discourse unit, as is illustrated by (12). (12) is a full clause, from which no essential elements are missing.

(12) *The budget should have the same discipline as national budgets*.

Schilperoord and Verhagen (1998) point out that not treating the clause types listed in (10) as discourse segments can at times be problematic. They provide the Dutch example in (13), in which dashes are used to indicate clause boundaries, to illustrate that not segmenting embedded clauses can result in a segmentation that underestimates the number of discourse segments in a fragment.

(13) Daarbij komt //dat zijn vrouw ernstig gehandicapt is //en dat hij een gezin heeft te onderhouden.
Thereby comes //that his wife severely disabled is //and that he a family has to take care of.
To this it can be added that his wife is severely disabled and that he has to take care of his family.
(Schilperoord and Verhagen 1998: 145)

(13) contains three clauses, but since two of them are coordinated clausal complements and therefore integrated parts of the main clause, applying the

clause criterion results in segmenting (13) as one discourse segment. However, Schilperoord and Verhagen (1998) point out that this goes against the intuition that two idea units are contained in the fragment: *his wife is severely disabled* and *he has a family to take care of*. They propose that after the first complement, the main clause has been completed, and is therefore not conceptually dependent on the second complement clause. The second complement clause can then be treated as a separate discourse segment.

(13a) [*Daarbij komt dat zijn vrouw ernstig gehandicapt is*] *en* [*dat hij een gezin heeft te onderhouden*.]

Although this seems like an adequate solution for this particular fragment, problems arise when trying to apply this same line of reasoning to relations such as (14).

(14) (*Mr President, I should like to take Commissioner Bolkestein back to the last part-session here when we discussed sales promotion*.)
He may remember that //I complimented him //because he had written an article in a journal //complimenting Parliament on //rescuing the internal market. {ep-02-09-25}

(14) contains five clauses, indicated by dashes, but the main clause, *he may remember that*, is conceptually dependent on a complement. If complement clauses are not allowed to be segmented, (14) would be a single discourse segment, since everything is embedded under the matrix structure *He may remember that*, or, in case of the fourth and fifth clause, embedded under the matrix structure and one or two other structures (as a reduced relative clause modifying the NP *an article in a journal*, and as a complement of the prepositional verb *compliment on* within the reduced relative clause, respectively). Following Schilperoord and Verhagen's (1998) reasoning, we arrive at the segmentation in (14a). Considering only the second clause, *I complimented him* as the complement embedded in the main clause suffices to make the main clause a conceptually independent unit. The clause following *because* can then be considered an independent discourse segment, which means it can enter into a coherence relation with other parts of the discourse. In (14), the coherence relation is explicitly signaled by means of *because*, indicating that the third clause *he had written an article in a journal complimenting Parliament on rescuing the internal market* is a reason for the content of the preceding discourse segment. It seems, however, inaccurate to state that the fact that Commissioner Bolkestein once wrote an article is the reason for him

remembering that the speaker once complimented him. Instead, it is more plausible that Commissioner Bolkestein's article was the reason for the speaker to compliment him. If the objective behind discourse segmentation is to represent the units of thought that are related to each other, the segmentation in (14a) seems undesirable, while the segmentation in (14b) more accurately captures the discourse structure.

(14a) [*He may remember that I complimented him*]$_{S1}$ *because* [*he had written an article in a journal complimenting Parliament on rescuing the internal market*.]$_{S2}$

(14b) *He may remember that* [*I complimented him*]$_{S1}$ *because* [*he had written an article in a journal complimenting Parliament on rescuing the internal market*.]$_{S2}$

Although the segmentation in (14b) may be appealing on the basis of the propositional content of the segments between which the causal relation is indicated to hold, it does leave the main clause of the sentence stranded. We want to propose that even if a complement is segmented as in (14b), the coherence relation as a whole can function as the complement of the main clause, as in (14c). This makes it structurally identical to a simple complement construction such as (15).

(14c) [*He may remember that* [*I complimented him*]$_{S1a}$ *because* [*he had written an article in a journal complimenting Parliament on rescuing the internal market*.]$_{S1b}$]$_{S1}$

(15) *He may remember that I complimented him*.

Schematically, this can be represented as in (16). X can be a single clause or a bigger chunk of text composed of multiple clauses.

(16) *He may remember that X*.

The relation in (17) differs from the relation in (14), even though both fragments have identical surface structures. In (17), the second segment in the coherence relation, *Air France cancelled my flight at 2.10 p.m.*, is not a reason for the content of the complement, the fact that the speaker is present in the meeting, but rather an argument supporting the content of the main clause (including its complement), the statement that it is an achievement that she is present.

(17) *Madam President, it is in itself an achievement that we are having this debate on the new URBAN Community initiative and* [*it is an achievement that I am here tonight*]$_{S1}$ *because* [*Air France cancelled my flight at 2.10 p.m.*]$_{S2}$ *but I am here!* {ep-00-02-14}

Adapting the segmentation rules to acknowledge that there can be other idea units expressed in a fragment in addition to the main clause will result in a more complete and accurate description of the discourse as a whole, since no information is lost because of the embeddedness of a clause. At the same time, allowing for the possibility of segmenting embedded clauses enables us to distinguish fragments in which a coherence relation holds between two clauses within a complement, as in (14), from fragments in which a clause is related to a main clause that contains a complement, as in (17). This difference is not only relevant to the organization of the discourse structure, but also helps us differentiate between two distinct meanings. It has been proposed that an important function of object complement constructions is to assign a proposition to the mental space of a subject (cf. Givón 1993; Verhagen 2001, 2005). In relations like (14), a causal relation is embedded in a subject's mental space. In (17), on the other hand, a reason is given for a mental space plus its content. This difference is illustrated by Figures 1 and 2. Determining whether a relation holds between two clauses within a clausal complement or between one segment containing a clausal complement and another segment can be done by considering the

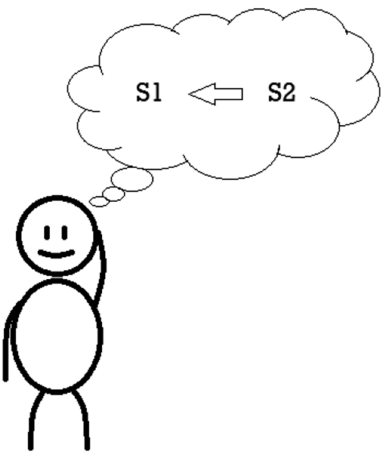


Figure 1: Coherence relation embedded in a mental space.

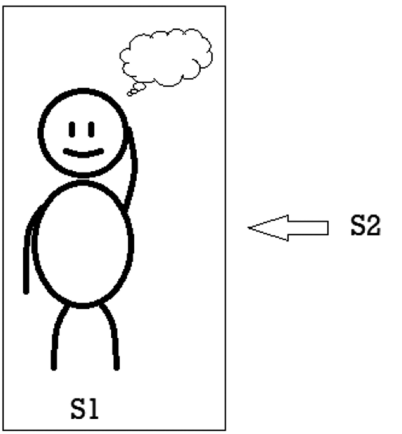


Figure 2: Coherence relation between a proposition and a proposition embedded in a mental space.

mental representation of the discourse and determining between which units of thought the relation holds. This is an interpretation process, in which annotation and segmentation are mixed.

If we allow the segmentation of clauses within a complement, the entire relation can be treated as part of the main clause when considering the larger discourse structure. In (18), for instance, a causal relation holds between the two clauses of the complement of the verb *see*, as indicated in (18a). The next clause, *in committee I proposed some form of business impact assessment* appears to be a result, explicitly signaled by *that is why*, of the preceding main clause including its complements: because the speaker did not want people to lose their jobs over social protection costs, he proposed investigating the effects the social protection plans would have on businesses. This relation is segmented in (18b).

(18) *I am in favour of social protection, I am in favour of the original Commission document, but I do not want to see people priced out of jobs because social protection costs become unrealistically high. That is why in committee I proposed some form of business impact assessment, so that costs and risks to jobs could be taken into account, and the EPP-DE Group supported this amendment*. {ep-00-02-15}

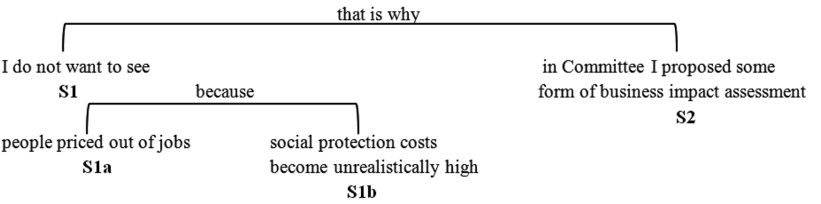
(18a) ... *I do not want to see* [*people priced out of* jobs]$_{S1}$ *because* [*social protection costs become unrealistically high.*]$_{S2}$

(18b) ... [*I do not want to see people priced out of jobs because social protection costs become unrealistically high.*]$_{S1}$ *That is why* [*in committee I proposed some form of business impact assessment*]$_{S2}$ ...

(18c) ... [*I do not want to see* [*people priced out of jobs*]$_{S1a}$ *because* [*social protection costs become unrealistically high.*]$_{S1b}$]$_{S1}$ *That is why* [*in committee I proposed some form of business impact assessment*]$_{S2}$ ...

The fact that the main clause is stranded and not included in the first segment of the relation in (18a), as would be done when following Schilperoord and Verhagen's (1998) method, has the advantage that it can be connected to other parts of the discourse without giving the impression that it features as the first segment of multiple relations. In addition, by segmenting the two relations as in (18a) and (18b), or as in the merged version in (18c), the segments of both relations accurately capture the idea units that are related to each other. Schematically, the segmentation in (18c) can be represented as in (19).

(19)



Because there is no relation on the discourse level between the main clause and the complement, there is no coherence relation indicated to hold between S_2 and $S_{1a\text{-}1b}$ in (19). In (18c), there is no closing square bracket after *see*, indicating that even though a new segment featuring in another coherence relation begins, *I do not want to see* is by itself not an independent discourse segment.

In the approach to segmenting fragments containing complement clauses we proposed in this section, we have adopted Schilperoord and Verhagen's (1998) notion of conceptual dependency, but changed the way in which we apply this notion to discourse segmentation. While they proposed the possibility of including only the first clause of a complement in its host clause to conceptually complete the main clause and allowing additional complement clauses to be independent discourse segments, we argued in favor of also having the option of segmenting the clauses within a complement. The entire coherence relation can then be used to conceptually complete the main clause. This approach, unlike Schilperoord and Verhagen's (1998), allows us to distinguish fragments in which a clause is related to a preceding complement from fragments in which a clause is related to a preceding main clause containing a complement construction, i. e., (14) versus (17). In the next sections, we will demonstrate that the same segmentation approach can be applied to fragments with clausal subjects, restrictive adverbial clauses, or restrictive relative clauses.

3.2 Clausal subjects, restrictive adverbial clauses, and restrictive relative clauses

Clausal subjects and restrictive adverbial clauses are similar to clausal complements, since in all these constructions, the main clause is conceptually dependent on the embedded clause. The segmentation approach proposed for text fragments with clausal complements can also be applied to fragments containing clausal subjects or restrictive adverbial clauses. (20) is an example of a sentence in which the subject is made up of two coordinated non-finite clauses.

Both *releasing terrorist prisoners* and *seeking to buy them off with places in rigged government* meet the structural criteria of discourse segments, as discussed in Section 2: they contain predicates, and they are not coordinated within a phrase. These clauses can therefore be segmented as in (20). The additive coherence relation as a whole can function as the subject of the sentence.

(20) [[*Releasing terrorist prisoners*]$_{S1a}$ *and* [*seeking to buy them off with places in rigged government*]$_{S1b}$ *is exalting terrorism*]$_{S1}$ *and* [*not eliminating it*.]$_{S2}$ {ep-01-09-05}

There will probably not be many fragments in which it is unclear whether only one clause or multiple clauses should be included in the clausal subject. The only scenario in which this could happen is if the sentence has a preceding coordinating sentence that could end with a non-finite clause.

For restrictive adverbial clauses there seems more room for ambiguity. Taking into account the propositional content of the segments in addition to the structural properties of the fragment, is therefore important. (21) and (22) have identical surface structures when considering the parts containing the restrictive adverbial clause. However, on the basis of the propositional content of the clauses, it can be determined that in (21) the final clause is part of the restrictive adverbial clause, while in (22) the final clause is connected to the preceding main clause, including its restrictive adverbial clause. The different segmentations of (21) and (22) reflect the differences in discourse structure between the two fragments. Note that Schilperoord and Verhagen (1998) would allow only the segmentation option in (22) (the fragment in (21) would be segmented in much the same way as (22), with *until* and the first clause following it included in the main clause, and the next clause as a separate discourse segment).

(21) *We are tired of the linkage of various directives in this package, with the excuse that we cannot look for a review of the European Works Council Directive until* [*it has been further bedded-in*]$_{S1}$ *and* [*the European company statute is in place*.]$_{S2}$ {ep-01-02-13}

(22) [*This cannot begin until there is a cessation of terrorism*]$_{S1}$ *and* [*only yesterday there was another suicide bomb attack in Jerusalem*.]$_{S2}$ {ep-01-09-04}

Restrictive relative clauses are slightly different from clausal complements, clausal subjects, and restrictive adverbial clauses, since they do not seem to

conceptually complete another clause, but rather a noun phrase, or referent. In contrast with restrictive relative clauses, non-restrictive relative clauses do not seem integral to the conceptualization of referents, and have traditionally been regarded as discourse segments. The segmentation of the relation signaled by *if* in (23) will therefore be allowed in most annotation approaches.

(23) *In addition, the Commission is now considering possible measures in the fields of technical assistance and trade,* [*which could be gradually extended*]$_{S1}$ *if* [*North Korea makes progress in the areas I have mentioned.*]$_{S2}$ {ep-01-01-17}

It seems, however, also possible for restrictive relative clauses to contain multiple clauses between which a coherence relation holds. In (24), in which dashes indicate clause boundaries, the noun phrase *the worried elderly people* is followed by a restrictive relative clause.

(24) *But on the BBC we saw a film recently //showing the deformed children and animals and the worried elderly people //who have decided to go back //because that* [*Chernobyl*] *was their home*, *//even though there is a risk*. {ep 96-04-17}

The segmentation strategy Schilperoord and Verhagen (1998) propose for clausal complements does not appear to be equally applicable to restrictive relative clauses. While it is technically possible to consider only the first clause after *people* as the relative clause, this results in a conceptually incomplete referent, since the group of people denoted in this fragment is more detailed than *the worried elderly people who have decided to go back*. In addition, including only the first clause of the restricted relative clause in the first segment, as in (24a), results in a segmentation that does not accurately represent the units of thought related to each other. The fact that Chernobyl was home to many people is *not* the reason why the speaker saw a film on the BBC.

(24a) # *But* [*on the BBC we saw a film recently showing the deformed children and animals and the worried elderly people who have decided to go back*]$_{S1}$ *because* [*that was their home*]$_{S2}$ ...

Again, a good way of arriving at a segmentation that represents the discourse structure without losing information is to allow segmentation within the embedded clause. This way, the segmentation does not only capture the causal within the relative clause, as indicated in (24b), but also the coherence relation signaled by *even though*, as indicated in (24c).

(24b) *... the worried elderly people* [*who have decided to go back*]$_{S1}$ *because* [*that was their home*,]$_{S2}$ *even though there is a risk.*

(24c) *... the worried elderly people* [[*who have decided to go back*]$_{S1a}$ *because* [*that was their home*,]$_{S1b}$]$_{S1}$ *even though* [*there is a risk.*]$_{S2}$

Text fragments that contain embedded clauses are prone to have multiple possible interpretations, since clauses adjacent to an embedded clause can be related to either the embedded clause or another clause in the discourse, usually a main clause. As argued above, interpretations can be differentiated by means of segmentation if we allow embedded clauses to potentially receive the status of discourse segments. In order to arrive at a segmentation that accurately reflects the inferred discourse structure, it seems important and perhaps even unavoidable to take into account the propositional content of the clauses when segmenting texts.

4 Stance markers and discourse segmentation

All of the fragments presented in Section 3 contained embedded structures. We demonstrated that by segmenting embedded clauses and allowing them to conceptually complete their superordinate structures, it is possible to arrive at segmentation options that accurately represent the discourse structure and leave no elements unaccounted for. For complement constructions such as the ones following *because* in (25) and (26), however, this option is not available.

(25) *I would like to put it to the Commissioner that* [*she lost the battle with her colleague Sir Leon Brittan on this*] *because* [*we understand that he is not very enthusiastic about dealing with the Norwegians and does not want to introduce restrictions.*] *He is frightened it might cause problems under the EEA agreement while we in the Committee on Fisheries and many people in Parliament take a different view.* {ep-97-01-16}

(26) *I have now been informed that* [*the Council will not deal with my question or ten other Members' questions*] *because* [*it claims it has not had time to prepare its replies.*] *I do not think that is acceptable.* {ep-02-04-10}

In (25) and (26), *we understand that* and *it claims* appear to not be part of the idea units related by *because*. In (25) it is not the speaker's understanding of Sir Leon Brittan's dislike of Norwegians that caused the Commissioner to lose her

battle, but rather Brittan's dislike of Norwegians itself. Similarly, in (26) it is not the Council's claim it did not have time to prepare replies that leads to the speaker's questions not being dealt with, but rather the Council's (supposed) lack of time. In these fragments it appears that the first segments relate to only the complements of the clauses following the connective; the only function of the superordinate clauses *we understand that* and *it claims* seems to be to modify the content of the complement clauses (a similar fragment can be found in PDTB Research Group 2007: 42, ex. 152). As was illustrated in Section 3, it is possible to leave initial matrix clauses outside the coherence relations, to have the entire coherence relations fall under their scope, and to connect the main clause, including its complement, to other parts of the discourse. Applying this approach to the second segments in (25) and (26), seems more problematic. First of all, the coherence relations would be indicated to hold between two units embedded under two different clauses. In addition, *we understand that* and *it claims* would be truly stranded. They cannot function as independent discourse units and are not related to other parts of the discourse. This would go against Mann and Thompson's (1988) criterion that all elements of a text should be included in the segmentation of that text. Yet, not excluding the superordinate clauses from the segments would go against our principle that discourse segments should represent the idea units that are related to each other, since *we understand that* and *it claims* do not seem to have a function within the coherence relations.

In this section, we will draw a parallel between fragments such as the ones in (25) and (26) and relations that contain stance adverbials and argue that discourse elements expressing stance can either have a function in the coherence relation as a whole, or merely modify one of the segments. After proposing that only the elements in a text that are part of the propositional content should obligatorily be included in the segmentation, we will present a solution to the segmentation problem fragments containing stance markers and complement-taking predicates represent.

4.1 Complement-taking predicates as stance markers

In Section 3 we focused mostly on the part of Schilperoord and Verhagen's (1998) conceptual dependency notion that stated that embedded clauses cannot enter into a relation with their host clause, but another aspect of the conceptual dependency criterion is that embedded clauses cannot enter into a relation with any other part of the discourse (that is not also embedded under the same

structure). However, this does appear to be the case in the relations in (25) and (26), since only the complements of the predicates following *because* seem to make up the idea units related to the first segments. Potential explanations are that either the definition of conceptual dependency is faulty, or that the complement constructions following *because* in (25) and (26) are not typical instances of clause embedding. Here we will argue that indeed the latter may be the case.

Schilperoord and Verhagen's (1998) definition of conceptual dependency implies that subordinate clauses may be more important than their matrix clauses. When it comes to predicates with object complements in particular, there has been a lot of discussion about the exact nature of the relation between the complement and its host clause. Although analyses of complement-taking verbs differ slightly in their specifics, what they seem to have in common is that they consider the complement to be central to the proposition being expressed. Both Givón (1993) and Verhagen (2001, 2005), for instance, propose that object complement constructions assign some proposition, expressed in the complement, to (the mental space of) a subject, expressed in the host clause. Fetzer (2014: 73) suggests that this aspect of complement-taking verbs makes them especially suitable to express epistemic stance about the proposition to which they are adjoined, since epistemic stance is "concerned with the speaker's evaluation of the certainty, possibility and probability of a state of affairs." Thompson (2002) even claims that complement-taking predicates (CTPs) are used to express epistemic stance, evidentiality, or evaluation in the majority of cases. Some complement-taking verb constructions, most of them with self-referencing subjects have grammaticalized and tend to be viewed as "parentheticals," the most notable example being *I think* (cf. Aijmer 1997; Brinton 2008; Traugott 1995). These parentheticals are generally analyzed as epistemic stance markers modifying the content of the following clause (Fetzer 2014; Hunter 2016).

Given the observed parallel between epistemic stance markers and CTPs, it is worthwhile exploring whether in discourse segmentation CTPs can be treated the same as stance markers. This comparison seems especially justified given CTPs' ability to express not just epistemic stance, but other types of stance as well. Conrad and Biber (2000: 57) identify three types of stance: *epistemic stance*, which comments on "the certainty (or doubt), reliability, or limitations of a proposition, including comments on the source of information," *attitudinal stance*, which conveys "the speaker's attitudes, feelings, or value judgements," and *style stance*, which describes "the manner in which

the information is being presented."[3] It appears that CTPs can also express attitudinal and style stance. In (27), for instance, the CTP expresses attitudinal stance, since the speaker conveys his positive attitude toward the proposition in the embedded clause. In (28) the CTP comments on the form in which the embedded clause is presented, and is thus an example of style stance.[4]

(27) *It is great that we are going to coordinate with the Americans*. {ep-00-06-14}

(28) *Let me just briefly reiterate that Parliament is provided in writing with a full list of the Commission's positions on each of the amendments*. {ep-02-10-22}

In the remainder of this section, we will demonstrate that stance adverbials can be part of the segments of a coherence relation, but can also occur outside of the relation, in which case they modify either the entire relation or one of the segments (Section 4.2). Subsequently, we will propose treating CTPs expressing stance in a way similar to adverbials of stance in discourse segmentation (Section 4.3).

4.2 Stance adverbials and segmentation

It seems possible to draw a parallel between (25) and (26), in which the second segments of the causal relations appear to be modified by their superordinate clauses, and relations in which S_2 is modified by a prototypical stance marker, for instance an adverbial, as in (29).

(29) [*I am glad that Commissioner Prodi is going to look at the EIB*] *because, frankly,* [*that institution is inefficient and ineffective in aiding those firms which could be innovative and competitive if they just had that helping hand*.] {ep-03-03-26}

3 In this paper we use Conrad and Biber's (2000) definition of epistemic stance, which includes evidentiality. Although we are aware of the ongoing debate on the exact relationship between evidentiality and epistemic stance (see cf. Cornillie 2009 for an overview), we do not feel that this issue is crucial to the current discussion.

4 Note that the segmentation problem posed by text fragments containing CTPs cannot be solved by annotating *attribution*, be it as a coherence relation, as in SDRT or RST, or as another type of construction, as in PDTB. Neither the CTP in example (30) nor the one in example (31) fits the definition of an attribution relation, which is to indicate who is responsible for the information in a fragment. Still, these examples do exhibit the same scopal properties as CTPs that do encode attribution.

In (29) it is not the case that the speaker's being frank about the EIB's inefficiency is the reason for the speaker to be glad it is being investigated. *Frankly* does not play a role in the coherence relation, but seems to merely modify S_2. This is in contrast to relations such as the one in (30).

(30) *They* [*transitory measures*] *are there for the time in which the market is still being directly regulated, but this whole package envisages a time when the entire market will operate under normal competition aspects*. [*Those transitory measures should be clearly identified*]$_{S1}$ *because* [*hopefully we will not need them in a few years' time*.]$_{S2}$ {ep-01-06-12}

S_2 in (30) also has a clause-initial stance adverbial, but *hopefully*, unlike *frankly* in (29), does seem to be part of the coherence relation: the speaker's hope that transitory measures will not be necessary in the future is the reason for his stating that they should be identified.

Stance adverbials can also have scope over an entire coherence relation, in which case they resemble complement constructions such as the ones in (14) and (18). Adverbials unequivocally have scope over a whole relation when they immediately precede the connective, as in (31a). Adverbials in other positions can also have scope over an entire relation: both (31b) and (31c) can, but need not, receive an interpretation similar to the relation in (31a).

(31a) [*The proportion of the complaints outside the mandate even increased slightly*,] *probably because* [*we received a growing number of complaints by e-mail*.] {ep-00-07-06}

(31b) *Probably*, [*the proportion of the complaints outside the mandate increased slightly*] *because* [*we received a growing number of complaints by e-mail*.]

(31c) [*The proportion of the complaints outside the mandate probably increased slightly*] *because* [*we received a growing number of complaints by e-mail*.]

Determining whether adverbials are part of the idea units related to each other, as in (30), or whether their function is to modify one of the segments, as in (29), or the relation as a whole, as in (31), can be crucial for the annotation of the fragments. One of the features of coherence relations important in many annotation approaches is whether a relation holds in the real world (or a fictional world), or whether it is constructed in the speaker's mind. This distinction has received many labels over the years: content vs. epistemic and speech act (Sweetser 1990), semantic vs. pragmatic (Sanders et al. 1992), internal vs. external (Halliday and

Hasan 1976), ideational vs. rhetorical (Mann and Thompson 1988; Redeker 1990), objective vs. subjective (Pander Maat and Sanders 2000), and others. Here, we will refer to this property of coherence relations as *source of coherence*, following Sanders et al. (1992). Certain adverbials can change a fact to a judgment, claim, or conclusion, e. g., *He is a judge*, vs. *He is* probably *a judge*, which can affect a relation's source of coherence and, consequently, the relation label ultimately attributed to a relation in annotation. Note that not all adverbials have potential consequences for annotation. Adverbials of time, for example, have the same scopal properties as other adverbials, but determining their scope will probably be less important in the process of discourse annotation than determining the scope of adverbials expressing stance.

4.3 Complement-taking verbs and discourse segmentation

Leaving an adverbial stranded, as in (29) and (31), seems less problematic than leaving an entire clause unaccounted for in the discourse structure, as in (25) and (26). However, CTPs and their complements do not always seem to correspond to typical host clause-embedded clause constructions, in which case the complement-taking predicate functions as a stance marker. Not incorporating a stance marker in the discourse structure seems acceptable, since stance markers are not part of the propositional content of a text, but rather "the lexical and grammatical expression of attitudes, feelings, judgments, or commitment concerning the propositional content of a message" (Biber and Finegan 1989: 93).

If we adopt the view that CTPs can potentially function as stance markers, fragments like (25) and (26) immediately become less problematic. The relations are no longer supposed to hold between two clauses embedded under different structures, and the only elements not being part of the idea units are stance markers rather than content elements of propositions. Both *we understand that* and *it claims* are instances of epistemic stance: they mention the source of information, and, especially in (26), comment on the speaker's idea of the actuality of the proposition.

The function of CTPs does not seem to be absolute. The same surface code, for instance *I know*, can have a different function depending on the context (Fetzer 2014). If CTPs can either express the mental space to which a proposition is assigned, or the speaker's stance toward a proposition, it is crucial for the process of discourse segmentation to determine which one is the case. If the main function of a CTP is judged to be assigning a proposition to a mental space, the predicate should be accounted for in the discourse structure, since it is part of the propositional content of a text. If, however, a CTP is judged to function as a stance marker, it should be treated in a way similar to other stance

markers, for instance adverbials. In that case, the CTP may be part of a segment, since a relation can be between a proposition including its stance and an another segment, as we have shown in Section 4.2, but can also modify only one of the segments and be left out of discourse segmentation.

It should be noted that stance markers also function as mental space builders in that they open the speaker space (cf. Dancygier and Sweetser 2012; Sanders and Redeker 1996). There is, however, a crucial difference in space building between CTP that function as stance markers and those that do not. If a CTP functions as a stance marker, the whole proposition, including the stance, is assigned to the mental space of the speaker. In (26), for instance, the status of *it has not had time to prepare its replies* is being questioned by the speaker. This process is different from the space building function of the CTP itself, which is to explicitly assign the contents of the complement to the mental space of the CTP's subject, which may, but certainly need not be the speaker.

Determining the function of a CTP within a specific text fragment relies heavily on its context: the exact same surface structure can function as a stance marker in one instance, and only connect a proposition to a mental space in another. There are, however, a few characteristics that seem to increase or decrease the chances of a CTP being a stance marker. Cognitive verbs with a first person singular pronoun, such as *I think, I mean, I hope*, or *I believe* seem to function as stance markers more often than other cognitive verbs (cf. Biber and Finegan 1989; Thompson 2002; Thompson and Mulac 1991). (32), for instance, is a colloquial example in which *I believe* functions as a stance marker: the speaker was not put in a small room because she believed there were no other rooms left. Instead, a more accurate paraphrase seems to be that she received the small room because it was the only available room, or so she thinks.

(32) *We got a small room because I believe it was the only one available.* (Tripadvisor 2009)

Despite cognitive verbs with a first person singular subject being more likely to function as stance markers, cognitive verbs with a different subject can also mark stance, as (25) and (26) illustrate.

CTPs can occur with or without a complementizer. Some have proposed that having a zero complementizer is the grammaticalized form of CTPs, and that CTPs without a complementizer can function as stance markers, while the function of CTPs with a complementizer is to assign a proposition to a mental space (cf. Aijmer 1997; Fetzer 2014). Others, however, propose that CPTs with complementizers can also function as stance markers (Kärkkäinen 2003; Thompson 2002). In addition, Kaltenböck (2009) argues that on the basis of

prosody there is no reason to assume that a complementizer affects a CTP's status, i. e., main clause versus stance marker. The presence of a complementizer therefore does not seem to be a reliable basis for excluding the possibility of a CTP functioning as a stance marker, although it may increase the likelihood of the CTP assigning a proposition to a mental space (Thompson and Mulac 1991).

This section explained examples such as (25), (26), and (32), in which S_2 starts with a CTP that does not seem to function in the relation, by arguing that CTPs and their complements are not always host clause-subordinate clause constructions. Instead, the CTP can function as a stance marker, in which case it is not part of the propositional content of the segment, but rather modifies the propositional content of S_2. Excluding a CTP from the representation of the discourse structure therefore seems justified when it functions as a stance marker, but when a CTP's main function is to assign a proposition to a mental space, it should be accounted for in discourse segmentation.

5 Discussion and conclusion

This paper has presented a theoretical approach to text segmentation and argued that segmentation without interpretation does not always result in an accurate representation of the discourse structure. The issues addressed in this paper were mainly illustrated by fragments taken from the Europarl corpus. This corpus consists of the written out proceedings of the European Parliament, which consist of a combination of prepared and spontaneous speech and contains both monologue and dialogue. As such, Europarl is a highly hybrid corpus. Some of the problems addressed in this paper may occur more often in written language, such as the complexity of some of the examples in Section 3, while other issues may be more essential to speech. Stance marking, for instance, seems to be generally more frequent in spoken than in written discourse (e. g., Biber 2006; Conrad and Biber 2000), and the use of CTPs as stance markers in particular has also been claimed to be especially frequent in speech (Thompson 2002). So even though coherence relations with an S_2 modified by a CTP seem to be very rare in written discourse (to our knowledge these have not been discussed anywhere else, with the exception of one example mentioned in PDTB Research Group 2007), we expect them to be more often encountered in spoken discourse. Our proposal for dealing with CTPs in discourse segmentation, whether they are located in S_1 or S_2, seems therefore particularly relevant now that discourse annotation is increasingly moving toward spoken and conversational data.

It should be noted that the account of complement-taking predicates in discourse presented in Section 4 focuses on English. While we believe that CTPs can function as stance markers in other languages as well, we question whether this fact alone always leads to constructions such as the ones in (25) and (26). When a CTP functions as a stance marker, the main clause has essentially become a function word, or discourse marker, while the subordinate clause functions as the main clause. This process appears to be mostly semantics-driven, since the basis seems to be the overlap in meaning between CTPs and other stance markers. English does not differentiate between main clauses and subordinate clauses in its word other or by any other means, which seems to enable such a change taking place. In languages that do syntactically distinguish main clauses from subordinate clauses, we do not expect to see discourse patterns similar to (25) and (26), since main clause/subordinate clause status is much more fixed. This, however, seems an issue worth exploring in future research.

Allowing embedded clauses to be segmented would lead to a more accurate representation of the structure of a discourse, but it would also increase transparency in discourse annotation, because the discourse segments will more accurately correspond to the units of thought that are inferred to be related to each other. If a fragment is, for example, segmented as in (17), partially repeated below, it can be assumed that the annotator interpreted the relation to hold between the main clause, including its embedded complement, and the clause following *because*. If, on the other hand, a fragment is segmented as in (14c) or (18a), both repeated below, it can be assumed that the annotator interpreted the relation to hold between the two clauses of the complement.

(17) ... [*it is an achievement that I am here tonight*]$_{S1}$ *because* [*Air France cancelled my flight at 2.10 p.m.*]$_{S2}$ *but I am here!*

(14c) [*He may remember that* [*I complimented him*]$_{S1a}$ *because* [*he had written an article in a journal complimenting Parliament on rescuing the internal market.*]$_{S1b}$]$_{S1}$

(18a) ... *I do not want to see* [*people priced out of jobs*]$_{S1}$ *because* [*social protection costs become unrealistically high.*]$_{S2}$

The segmentation would unambiguously indicate between which units of thought annotators considered the relation to hold. In case annotators have not attributed the same relation label to a fragment, differences in segmentation would immediately pinpoint the source of disagreement between annotators.

Although incorporating interpretation in the segmentation process leads to more accurate text segmentation, it does pose a problem for automatic text segmentation, which is an important and promising technique being developed both within the discourse community and in the field of NLP research. By identifying specific contexts in which multiple segmentation options should be considered, we can limit the amount of text for which we have to take into account meaning during segmentation. While automatic text segmentation systems will not be able to disambiguate fragments, it would be possible for them to flag, for instance, complement constructions. Only the crucial parts of a text would then have to be manually checked by a post editor. As constructions with multiple segmentation options, this paper pointed out complement constructions, restrictive relative clauses, restrictive adverbial clauses, or stance markers, but other linguistic contexts may also be identified as often being structurally ambiguous. Having an inventory of constructions that are especially prone to segmentation ambiguities can also help limit the amount of text for which meaning has to be taken into account in manual text segmentation. This would preserve the original concept of treating segmentation and annotation as two separate steps as much as possible.

This paper has argued that while the grammatical clause is a functional basis for identifying discourse segments, it is sometimes necessary to take into account the propositional content of the text to arrive at a segmentation of a text that accurately represents the discourse structure and in which the discourse segments correspond to the units of thought related to each other. One of the segmentation issues where meaning can play a role is ellipsis, in which case the situation model can be taken into account to determine whether a structure is a clause with an ellipted subject and main verb, or rather coordinated nouns within a single phrase functioning as a direct object. We also argued in favor of amending Mann and Thompson's (1988) completeness constraint, i. e., the criterion that all elements should be included in the segmentation of a text, to pertain only to the propositional content of a discourse. Stance markers, which are not part of the propositional content of the text, may for instance be left out. Determining whether a stance marker should be included in a text segment, can be done by considering the interpretation of the text. Finally, we demonstrated that for fragments with embedded clauses, for instance clausal complements or relative clauses, multiple segmentation options should be considered. Using the interpretation of a text fragment can help to distinguish between distinct syntactic structures that have identical surface structures, e. g., (17) versus (14c), and to arrive at an accurate representation of the discourse structure.

Funding: This work was funded through the SNSF Sinergia project MODERN (CRSII2_147653).

References

Aijmer, Karin. 1997. I think – an English modal particle. In Toril Swan & Olaf Jansen Westvik (eds.), *Modality in Germanic languages: Historical and comparative perspectives*, 1–48. Berlin & New York: Mouton de Gruyter.

Asher, Nicholas & Alex Lascarides. 2005. *Logics of conversation*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5. 97–116.

Biber, Douglas & Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9(1). 93–124.

Brinton, Laurel J. 2008. *The comment clause in English*. Cambridge: Cambridge University Press.

Carlson, Lynn & Daniel Marcu. 2001. *Discourse tagging reference manual*. ISI technical report ISI-TR-545. doi:ftp://128.9.176.20/isi-pubs/tr-545.pdf (accessed 23 July 2014).

Cartoni, Bruno, Sandrine Zufferey & Thomas Meyer. 2013. Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27(1). 23–42.

Conrad, Susan & Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In Geoff Thompson & Susan Hunston (eds.), *Evaluation in text: Authorial stance and the construction of discourse*, 56–74. Oxford: Oxford University Press.

Cornillie, Bert. 2009. Evidentiality and epistemic modality: On the close relationship between wo different categories. *Functions of Language* 16(1). 44–62.

Dancygier, Barbara & Eve E. Sweetser. 2012. *Viewpoint in language: A multimodal perspective*. Cambridge: Cambridge University Press.

Degand, Liesbeth. 1996. Causation in Dutch and French: Interpersonal aspects. In Ruqaiya Hasan, Carmel Cloran & David G. Butt (eds.), *Functional descriptions: Theory in practice*, 207–237. Amsterdam: John Benjamins.

Evers-Vermeul, Jacqueline. 2005. *The development of Dutch connectives: Change and acquisition as windows on form-function relations*. Utrecht University PhD thesis. Utrecht: LOT. http://www.lotpublications.nl/Documents/110_fulltext.pdf

Fetzer, Anita. 2014. *I think, I mean* and *I believe* in political discourse: Collocates, functions and distribution. *Functions of Language* 21(1). 67–94.

Givón, Talmy. 1993. *English grammar: A function-based introduction*. Amsterdam: Benjamins.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London & New York: Routledge.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive Science* 3. 67–90.

Hobbs, Jerry R. 1990. *Literature and cognition*. Stanford: CSLI.

Hunter, Julie. 2016. Reports in discourse. *Dialogue and Discourse* 7(4). 1–35.

Kaltenböck, Gunther. 2009. Initial *I think*: Main or comment clause? *Discourse and Interaction* 2 (1). 49–70.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation. A description of its interactional functions, with a focus on 'I think'*. Amsterdam: John Benjamins.

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*. Stanford: CSLI.

Kehler, Andrew, Laura Kertz, Hannah Rohde, & Jeffrey L. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics 25*(1). 1–44.

Koehn, Phillip. 2005. Europarl: A parallel corpus for statistical machine translation. Tenth Machine Translation Summit (MT Summit X). http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf (accessed 8 April 2014).

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281.

Matthiessen, Christian & Sandra A. Thompson. 1988. The structure of discourse and 'subordination.' In John Haiman & Sandra A. Thompson (eds.), *Clause combining and grammar in discourse*, 275–329. Amsterdam & Philadelphia: John Benjamins.

Pander Maat, Henk L. W. 2002. *Tekstanalyse* [text analysis]. Bussum: Coutinho.

Pander Maat, Henk L. W. & Ted J. M. Sanders. 2000. Domains of use or subjectivity? The distribution of three Dutch causal connectives explained. *Topics in English Linguistics* 33. 57–82.

Pareti, Sylvia. 2012. A database of attribution relations. *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. http://lrec-conf.org/proceedings/lrec2012/index.html (accessed 18 December 2015).

PDTB Research Group. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. IRCS technical report. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1203&context=ircs_reports (accessed 22 April 2014).

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12(5–6). 601–638.

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14(3). 367–381.

Reese, Brian, Julie Hunter, Nicholas Asher, Pascal Denis, & Jason Baldridge. 2007. *Reference manual for the analysis of rhetorical structure*. Unpublished manuscript. Austin, TX: University of Texas at Austin. http://timeml.org/jamesp/annotation_manual.pdf (accessed online 25 July 2014).

Renkema, Jan. 2009. *The texture of discourse: Towards an outline of connectivity theory*. Amsterdam & Philadelphia: John Benjamins.

Sanders, Ted J. M. 1997. Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes* 24(1). 119–147.

Sanders, José & Gisela Redeker. 1996. Speech and thought in narrative discourse. In Gilles Fauconnier & Eve E. Sweetser (eds.), *Spaces, worlds and grammar*, 290–317. Chicago: University of Chicago Press.

Sanders, Ted J. M., Wilbert P. M. S. Spooren, & Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15(1). 1–35.

Sanders, Ted J. M. & Carel H. van Wijk. 1996. PISA – A procedure for analyzing the structure of explanatory texts. *Text* 16(1). 91–132.

Schilperoord, Joost & Arie Verhagen 1998. Conceptual dependency and the clausal structure of discourse. In Jean-Pierre Koenig (ed.), *Discourse and cognition: Bridging the gap*, 141–163. Stanford: CSLI.

Stukker, Ninke, Ted J. M. Sanders, & Arie Verhagen. 2008. Causality in verbs and in discourse connectives. Converging evidence of cross-level parallels in Dutch linguistic categorization. *Journal of Pragmatics* 40(7). 1296–1322.

Sweetser, Eve E. 1990. *From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change*. Cambridge: Cambridge University Press.

Taboada, Maite & William C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies* 8(3). 423–459.

Thompson, Sandra A. 2002. "Object complements" and conversation: Towards a realistic account. *Studies in Language* 26(1). 125–163.

Thompson, Sandra A. & Anthony Mulac. 1991. The discourse conditions for the use of the complementizer that in conversational English. *Journal of Pragmatics* 15(3). 237–251.

Traugott, Elizabeth C. 1995. Subjectification in grammaticalization. In Dieter Stein & Susan Wright (eds.), *Subjectivity and subjectivisation*, 31–54. Amsterdam: Benjamins.

Tripadvisor review. 2009, August 31. Retrieved January 7, 2016, from http://www.tripadvisor.co.uk/ShowUserReviews-g211878-d654059-r39310708-Cloisters_Bed_BreakfastKinsale_County_Cork.html.

Verhagen, Arie. 2001. Subordination and discourse segmentation revisited, or: Why matrix clauses may be more dependent than complements. In Ted J. M. Sanders, Joost Schilperoord & Wilbert P. M. S. Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects*, 337–357. Amsterdam & Philadelphia: John Benjamins.

Verhagen, Arie. 2005. *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford: Oxford University Press.

Versley, Yannick & Anna Gastel. 2013. Linguistic tests for discourse relations in the TüBa-D/Z corpus of written German. *Dialogue and Discourse* 4(2). 142–173.

Wolf, Florian & Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics* 31(2). 249–287.