Discourse Segmentation and Ambiguity in Discourse Structure

Jet Hoek, Jacqueline Evers-Vermeul, Ted J.M. Sanders j.hoek@uu.nl

Universiteit Utrecht

Introduction

- Coherence relations connect two or more segments.
- Ideally, implementing segmentation rules results in text segments that correspond to the units of thought related to each other.
- The **clause** as the basis for assigning discourse segment status (e.g. Evers-Vermeul, 2005; Mann & Thompson, 1988; Sanders & van Wijk, 1996; Wolf & Gibson, 2005)
  - Theory-neutral
  - Can be objectively determined (although ellipsis can be an issue)
    - *Includes*: finite and non-finite clauses
    - *Excludes*: PPs, non-clausal adverbials / modifiers
- Sometimes multiple segmentation options should be considered, for instance for fragments containing **complement constructions**.
  → This also holds for fragments containing relative clauses and adverbial clauses.
- These fragments are especially prone to ambiguity, which affects segmentation.
  - **This can in turn affect the annotation of relation labels**
- For fragments with complement constructions, it seems necessary to take into account the **propositional content** of the fragment:

(1) 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}

(2) It is an achievement that I am here tonight **because** Air France cancelled my flight at 2.10 p.m. {ep-00-02-14}

S_1 in (1): Only complement
S_1 in (2): Complement + complement-taking predicate (CTP)

(1) [He may remember that [I complimented him]$_{S1a}$ **because** [he had written an article in a journal complimenting Parliament on rescuing the internal market.]$_{S1b}$]$_{S1}$ {ep-02-09-25}

(2) [It is an achievement that I am here tonight]$_{S1}$ **because** [Air France cancelled my flight at 2.10 p.m.]$_{S2}$ {ep-00-02-14}

→ **Here, we will use fragments containing complement constructions to illustrate 4 text features that can help solve structural ambiguity and help arrive at a segmentation that accurately represents the units of thought connected in the discourse.**

Disambiguating features

1. Subjectivity

- *Because* signals both objective and subjective causal relations.
- Dutch *want* predominantly signals subjective causal relations.

English original:

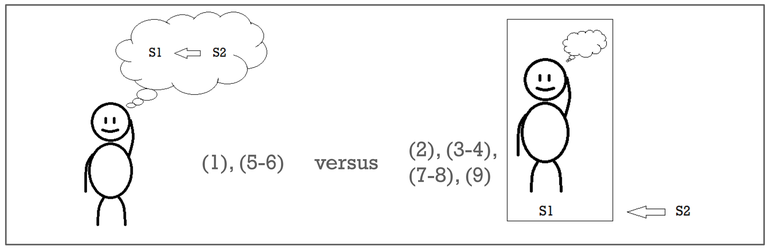
(3) It is a great pity indeed that Commissioner Barnier has been unable to be present here this morning, **because** this is a matter within his brief which is causing great concern not only in Scotland and Wales but in other parts of the Union. {ep-00-03-17}

Dutch translation:

(4) [Het is erg jammer dat commissaris Barnier hier vanmorgen niet kon zijn,] **want** [dit is een kwestie uit zijn bevoegdhedenpakket die niet alleen Schotland en Wales, maar ook andere regio's uit de Unie grote zorgen baart.]

- S_1 is complement: fact → relation is objective = Mismatch
- S_1 is complement + CTP: judgment → relation is subjective

→ **Include the complement-taking predicate in S_1**



2. Volitionality

- *Because* signals both volitional and non-volitional causal relations.
- Dutch *doordat* predominantly signals non-volitional causal relations.

English original:

(5) 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. {ep-00-02-15}

Dutch translation:

(6) *Ik ben voor sociale bescherming, en ik ben het ook eens met het oorspronkelijke document van de Commissie, maar* ik wil niet dat [mensen hun baan kwijtraken] **doordat** [de kosten van sociale bescherming onrealistisch hoog worden.]

- Complement: non-volitional event
- Complement + CPT: judgment (+explicit SoC) = Mismatch

→ **Only the complement as S_1**

3. Structural properties of the connective:

- Not all connectives can be embedded.
- *Because* can be embedded.
- Dutch *omdat* can be embedded, *want* cannot.

English original:

(7) I hope that Commissioner Vitorino feels the same way, **because** this would be a disastrous outcome. {ep-00-09-20}

Dutch translation:

(8) [Ik hoop dat commissaris Vitorino er ook zo over denkt,] **want** [anders zou het resultaat rampzalig zijn.]

S_1 is complement: connective is embedded = Mismatch
S_1 is complement + CTP: connective not embedded

→ **Include the complement-taking predicate in S_1**

4. Structural properties of the fragment:

- **Displacement** – an element of the embedded clause is positioned in the host clause (or vice versa), as the negation in (9):

(9) [We do not think these additions should be made] **unless** [it can be demonstrated that the consumer benefits.] {ep-00-04-10}

Complement and CTP are an integrated whole.

→ **Include the complement taking predicate in S_1**

Conclusion & Discussion

- Treating segmentation and annotation as a two-step process may not always result in segments that correspond to the units of thought related to each other.
- It is sometimes necessary to take into account the propositional content of the fragments during segmentation.
- Certain properties of connectives can help disambiguate between different interpretations of fragments and thus facilitate segmentation.
- Identifying for which types of constructions multiple segmentation options should be considered can help limit the number of fragments for which the propositional content has to be taken into account.
- Sometimes disambiguation may not be possible:

(10) The BBC recently produced evidence that 'wombs', as they described it, were for sale in Romania - that women were being paid to have children for export to Member States of the European Union. Furthermore, the BBC alleged that this was being done with the tacit approval of the Romanian authorities **because** it was bringing hard currency into Romania. {ep 00-03-15}

References

Carlson, L., & Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI technical report ISI-TR-545. Retrieved from http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf>.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit* (pp. 79-86). Phuket, Thailand.

Mann, W.C., & Thompson, S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text,* 8, 243-281.

Sanders, T.J.M., van Wijk, C. (1996). PISA - A procedure for analyzing the structure of explanatory texts. *Text,* 16, 91-132.

Schilperoord, J., & Verhagen, A. (1998). Conceptual dependency and the clausal structure of discourse. In J. Koenig (Ed.), *Discourse and cognition: Bridging the gap* (pp. 141-163). Stanford, CA: CSLI Publications.

Verhagen, A. (2001). Subordination and discourse segmentation revisited, or: Why matrix clauses may be more dependent than complements. In: T.J.M. Sanders, J. Schilperoord, & W.P.M.S. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 337-357). Amsterdam/Philadelphia: John Benjamins.

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

Acknowledgments

This work is part of the MODERN project, supported by SNSF grant CRSII2_147653