Complete and independent? Reconsidering discourse segmentation basics

Jet Hoek
Utrecht University
j.hoek@uu.nl

Jacqueline Evers-Vermeul
Utrecht University
j.evers@uu.nl

Ted J.M. Sanders
Utrecht University
t.j.m.sanders@uu.nl

AG2

This paper theoretically approaches discourse segmentation and focuses on two issues concerning segmentation that were proposed by Mann and Thompson (1988) in their introduction of RST, but that have been implemented in many other discourse annotation approaches as well: 1) 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 & Mann 2006), and 2) the completeness constraint, which poses that the segmentation of a text has to include all elements of that text. Taking the clause as the syntactic basis for the identification of discourse segments, we discuss English fragments, mainly from the Europarl corpus (Koehn 2005), that present segmentation difficulties. 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 can be at the expense of the quality of the segmentation. For fragments with embedded clauses, e.g., clausal complements, multiple segmentation options should be considered. Using the interpretation of a text fragment can help to distinguish between distinct syntactic constructions that have identical surface structures.

We argue in favor of amending Mann and Thompson's (1988) completeness constraint to pertain only to the propositional content of a discourse. Stance markers, which are not part of the propositional content of the text (Biber & Finegan 1989), 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.

References: • Biber, D. & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect.*Text* 9(1),93-124. • Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit X.* • Mann, W.C. & Thompson, S.A. (1988). RST: Toward a functional theory of text organization. *Text* 8(3), 243-281. • Taboada, M. & Mann, W.C., (2006). RST: Looking back and moving ahead. *Discourse Studies* 8(3), 423-459.