
Making Sense of human-annotated text structure using automatic segmentation contours

Sophia Conrad¹ & Florian Frenken²

¹University of Zurich, ²RWTH Aachen University

sophia.conrad@uzh.ch, florian.frenken@rwth-aachen.de

To date, computational linguistics has primarily approached the issue of text structure from a semantic perspective, relying on lexical cohesion to identify boundaries of topically consistent segments in discourse. Hearst's (1997) seminal TextTiling algorithm was pivotal in this development. Functional segmentation, which has traditionally been the focus of research in genre theory (Swales 1990), on the other hand, has received comparatively little attention, not least due to the inherent subjectivity of the concept, which complicates automation. Besides overarching cultural conventions, different production contexts, i.e. register, also play an important role in shaping how texts unfold over time. An automatic approach to functional segmentation should therefore take both perspectives into account.

To this end, human annotations of textual functions are necessary for training and evaluation. However, they are both time-consuming and typically show low inter-annotator agreement. We theorize that the observed differences may be a result of annotators attending on various levels in the hierarchy of textual organization (cf. Kintsch & Van Dijk 1978). The aggregate annotations could therefore be interpreted as a statistical distribution of boundary strength, which may correlate with attention to structural elements at the macro, meso, and micro levels, where linguistic changes differ in perceptual salience.

The basis of our analyses are text segmentations by human annotators, the TextTiling algorithm, and LLMs of 90 texts from the International Corpus of English (ICE) (Greenbaum 1991). We investigate their agreement at different granularities in relation to *segmentation contours* defined here as frequencies of linguistic features over text time obtained via sliding windows (see Ströbel et al. 2016). This way, we aim to evaluate to what extent automatic segmentation can be used as a suitable proxy for costly human annotation.

References: • Greenbaum, S. (1991). ICE: The International Corpus of English. *English Today* 7(4), 3–7. • Hearst, M. A. (1997). Text Tiling: Segmenting Text Into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23(1), 33–64. • Kintsch, W. & T. A. Van Dijk (1978). Toward a Model of Text Comprehension and Production. *Psychological Review* 85(5), 363–394. • Ströbel, M., Kerz, E., Wiechmann, D. & S. Neumann. (2016). CoCoGen – Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-window Technique. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. • Swales, J. M. (1990). *Genre Analysis*. Cambridge University Press.