

Automatic Situation Entity Segmentation for Historical and Contemporary English and German

Hanna Schmück, Veronika Urban, Xaver Maria Krückl, Sonja Zeman,
Claudia Claridge, and Annemarie Friedrich
University of Augsburg, Germany

Historical corpora allow for studying how authors organised and presented knowledge over time, for example via the discourse structures they employ. Large-scale discourse analysis of cross-lingual and diachronic corpora requires discourse segmentation which cannot feasibly be done by hand at scale. The task has been approached from both syntactic (Bies et al., 1995) and discourse perspectives (Polanyi, 1995; Polanyi et al., 2004; Carlson et al., 2001), and operationalised in recent shared tasks on elementary discourse unit segmentation such as DISRPT (Braud et al., 2025). It is a non-trivial task due to its context-dependent nature.

In this work, we explore Situation Entity (SE) segmentation (Smith, 2003), a semantically grounded type of discourse segmentation that identifies clause-like units based on the situation type denoted by their main verb. SE segments can serve as the foundation for discourse mode classification (Smith, 2003), which captures shifts between, for example, Narrative and Information mode and can be used to study authorial style and genre conventions. SE segments further provide proposition-level information which can, for example, be used for Event extraction in historical texts. While SE segmentation guidelines have been developed for current English (Friedrich et al., 2016) and explored for contemporary German (Mavridou et al., 2015), guidelines for historical varieties have remained underspecified. The SE framework does not directly correspond to any of the formalisms currently represented in DISRPT.

For a systematic implementation of SE segmentation, we provide principled segmentation guidelines for four language varieties: Contemporary English, Late Modern English (LModE, 1700–1900), Contemporary German, and New High German (NHG, 1650–1900). The guidelines explicitly handle constructions that are frequent in historical text and commonly lead to annotator disagreements, including participial absolutes and reduced relative

clauses in LModE, and auxiliary ellipsis and less restricted verb positioning in NHG.

To validate the guidelines, we conduct an inter-annotator agreement (IAA) study with five trained annotators across all four varieties (two annotations per variety). The datasets used are the Corpus of Late Modern English Historiography (CLMEH; Claridge, 2025), the GiesKaNe corpus of New High German (Justus-Liebig-Universität Gießen, 2022), as well as MASC (Ide et al., 2008) and Wikipedia data for current English and Wikipedia and Blogs for current German. Human annotators achieve substantial agreement across all varieties (75–84% exact span match), confirming that the guidelines are learnable and consistent even for historically and typologically distant data.

Building on approximately 40,000 labelled SE segments from the SitEnt corpus of contemporary English (Friedrich et al., 2016), we train a new automatic segmenter by fine-tuning XLM-RoBERTa (Conneau et al., 2020), following a similar architecture as explored for German discourse unit segmentation (Frenzel et al., 2026), and replacing the prior approach that relied on post-processing via a legacy rule-based system (Soricut and Marcu, 2003; Friedrich et al., 2016). The model achieves an exact span match of 74% on contemporary English, approaching human performance. Applied in a zero-shot setting to the remaining varieties, however, performance degrades substantially to approximately 67% exact match for Contemporary German, and 46% for NHG which combines cross-lingual and temporal domain shift. Importantly, human annotators remain consistent under the same conditions, indicating that the difficulty lies not in the task itself but in the model’s dependence on surface features of contemporary English.

Future work will explore historical text normalisation as a preprocessing step and the incorporation of dependency-parsed clause structure to better inform segmentation decisions.

References

- Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. Technical report, University of Pennsylvania.
- Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. [The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–20, Suzhou, China. Association for Computational Linguistics.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. [Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory](#). In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.
- Claudia Claridge. 2025. [History writing](#). In M. Kytö and E. Smitterberg, editors, *The New Cambridge History of the English Language: Documentation, Sources of Data and Modelling*, pages 433–458. Cambridge University Press, Cambridge.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Steffen Frenzel, Maximilian Krupop, and Manfred Stede. 2026. [Discourse segmentation of german text with pretrained language models](#). *Journal for Language Technology and Computational Linguistics*, 39(1):1–31.
- Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. [Situation entity types: automatic classification of clause-level aspect](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.
- Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. [MASC: the manually annotated sub-corpus of American English](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Justus-Liebig-Universität Gießen. 2022. Syntaktische grundstrukturen des neuhochdeutschen—korpus. <https://www.uni-giessen.de/de/fbz/fb05/germanistik/forschung/sprache/gieskane/korpus>. Accessed: 2026-03-04.
- Kleio-Isidora Mavridou, Annemarie Friedrich, Melissa Peate Sørensen, Alexis Palmer, and Manfred Pinkal. 2015. [Linking discourse modes and situation entity types in a cross-linguistic corpus study](#). In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.
- Livia Polanyi. 1995. The linguistic structure of discourse. Csl technical report, CSLI, Stanford, CA.
- Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. [A rule based approach to discourse parsing](#). In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 108–117, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103 of *Cambridge Studies in Linguistics*. Cambridge University Press.
- Radu Soricut and Daniel Marcu. 2003. [Sentence level discourse parsing using syntactic and lexical information](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.