Cross-Linguistic Situation Entity Segmentation for Discourse Analysis in Diachronic English and German Text

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

Abstract

Situation Entity (SE) segmentation identifies clause-like discourse units focusing on verb constellations. While SE segmentation has been applied to contemporary English as a subtask of SE annotation, systematic guidelines for syntactically ambiguous constructions remain underspecified. We present principled SE segmentation guidelines for contemporary and historical varieties of English and German. Our inter-annotator agreement studies on Late Modern English (1700–1900) and New High German (1650–1900) corpora demonstrate substantial agreement. Using the existing SitEnt corpus in contemporary English, we implement a new automatic segmenter based on `XLM-RoBERTa`. Our evaluation examines cross-variety and cross-lingual generalization, demonstrating challenges both for human annotation efforts and in transferring segmenters trained on contemporary English to historical varieties. Our code and data are publicly available.[1]

1 Introduction

Discourse segmentation, the task of identifying meaningful units of text, is fundamental to computational discourse analysis. One framework for discourse analysis and segmentation are Smith's (2003) modes of discourse. These modes result from specific situation entity (SE) patterns, and SE segments are the focus of our annotation task. This framework is of particular interest since it offers a linguistically motivated approach of segmenting texts which is grounded in aspectual properties of clause-like elements. An example is visualized in Figure 1. SE segmentation serves as a prerequisite for two core downstream tasks: Firstly, it can be used as a basis for discourse mode classification, a valuable tool for the analysis of literary and historical corpora where shifts between, e.g., Narrative and Information mode reflect authorial style, genre conventions, and rhetorical structure. Secondly, SE segments correspond roughly to propositions and therefore constitute a linguistically motivated unit for information and event extraction in historical texts.

Figure 1: Situation annotation of New High German (top, Koralek (1889/1890)) and Late Modern English (bottom, Salmon (1724)) text snippets showing one segment per line and the respective Main Verb and Main Referent relations. Main Verbs can be connected to their respective Main Referent across segment borders.

Manual segmentation of text into clauses or clause-like units is not a trivial task, and requires the careful design of annotation guidelines. It has been approached both from a syntactic perspective (Bies et al., 1995) and a discourse perspective (Polanyi, 1995; Carlson et al., 2001; Polanyi et al., 2004). Automatic clause segmentation is also non-trivial (Tjong et al., 2001; Soricut and Marcu, 2003; Tofiloski et al., 2009). As Carlson et al. (2001) put it, "the boundary between discourse and syntax can be very blurry."

While SE annotation has been explored for contemporary German (Mavridou et al., 2015) and applied to English (Friedrich et al., 2016), existing guidelines for segmentation (Friedrich et al., 2015a, chapter 3) mainly specify how annotators are supposed to handle differences vs. the EDU segmentation produced by a legacy system (Soricut and Marcu, 2003).

[1] https://github.com/coling-unia/sitent-segmenter-law2026

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 95–112

In this work, we address these challenges by providing more systematic SE segmentation guidelines for both contemporary and historical varieties of English and German. We thereby close the gap left by previously undocumented approaches and generalize the idea to new language variants. We further conduct a small IAA study on SE segmentation of these varieties based on the Corpus of Late Modern English Historiography (CLMEH (Claridge, 2025)) and the GiesKaNe corpus (Justus-Liebig-Universität Gießen, 2022). Lastly, we fine-tune `XLM-RoBERTa` models for automatic SE segmentation and evaluate their cross-variety generalization abilities. Our automatic segmenter achieves a boundary F1 of 0.90 on contemporary English, and human annotators maintain substantial agreement across all four varieties (75–84% exact matches), demonstrating that the guidelines can be adapted to suit cross-lingual and historical settings.

2 Background

We provide an overview of SE types and discuss the broader research area of discourse segmentation.

2.1 Linguistic Background

Discourse modes (Smith, 2003) refer to different text types (Werlich, 1989; Biber, 1989; Adam, 2011), which are characterized by clusters of linguistic features. In Smith's *Narrative* and *Report* modes, for example, the reader perceived the discourse to move on predominantly via temporal relations. By contrast, progression is perceived as spatial in *Description* mode, like traversing through a scene. In *Information* and *Argument* mode, the discourse advances by focusing on different referents that are part of the domain of the discourse.

Discourse modes are not only distinguished by their type of progression: they also favor different distributions of *SE types*. SE types capture the aspectual form of (roughly) clauses. We illustrate and describe the inventory of SE types as proposed in prior work (Friedrich and Palmer, 2014b; Friedrich et al., 2016) in Table 1.

According to Smith (2003), SE types are assigned to *verb constellations*, i.e., a verb and its "primary referent" in the context of its arguments and modifiers. A crucial step in SE segmentation hence lies in determining what counts as a *main verb*. Friedrich et al. (2016) flesh out three features that help to distinguish SE types for new annotators or to obtain partial information if not all features can be determined. First, annotators determine whether the *main referent*, in English typically the subject of the main verb, is generic (Krifka et al., 1995) or whether it refers to a particular individual (see also Nedoluzhko, 2013; Friedrich and Pinkal, 2015b; Friedrich et al., 2015b). Generic main referents identify GENERIC SENTENCES. Second, the lexical aspectual class (Moens and Steedman, 1988; Klavans and Chodorow, 1992; Friedrich and Palmer, 2014a) plays a role, e.g., for distinguishing EVENTS from STATES. Third, if the main verb is *habitual*, i.e., if it indicates a situation that happens regularly, the SE must be either a GENERIC or a GENERALIZING SENTENCE (Vendler, 1957; Mathew and Katz, 2009; Friedrich and Pinkal, 2015a). The full correspondence between the lower-level aspectual features and SE types are provided in Appendix B.

SE type	Example
Eventualities	
STATE	The colonel owns the farm.
EVENT	John won the race.
REPORT	"…", said Obama.
General Statives	
GENERIC SENT.	The lion has a bushy tail.
GENERALIZING SENT.	Mary often fed the cat last year.
Abstract Entities	
FACT	I know that she refused the offer.
PROPOSITION	I believe that she refused the offer.
QUESTION	Who wants to come?
IMPERATIVE	Hand me the pen!

Table 1: SE type inventory (Smith, 2003; Friedrich et al., 2016; Friedrich and Palmer, 2014b).

Challenges regarding SE boundary identification intensify when working with historical language varieties. Late Modern English (LModE; approximately 1700–1900) and New High German (NHG; approximately 1650–1900) exhibit syntactic features that complicate SE boundary identification. LModE employs participial absolutes, reduced relative clauses, and infinitival purpose constructions with less explicit subordination marking than contemporary varieties. NHG shows a less restricted word order, particularly in verb positioning, when compared to present day German and makes use of participial clauses and final clause constructions with *damit* which introduce new boundary cases.

2.2 Existing Discourse Segmentation Guidelines

The first step when parsing text into discourse structures consists in determining which linguistic units correspond to semantic units in the discourse. Several frameworks have proposed different solutions to this problem. The Linguistic Discourse Model (Polanyi, 1995; Polanyi et al., 2004) defines Basic Discourse Units (BDUs) as segments with the potential to establish anchor points for future attachment, identifying syntactic constructions able to carry the necessary semantic information. In practice, BDUs are often rather small units, e.g. in the following case:

(1) [Germany elected Merz] [chancellor].

This is more fine-grained than SE annotation, since *chancellor* does not constitute a separate situation.

The Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004), distinguishes inter-sentential and intra-sentential discourse relations between segments called *argument spans*. The latter relate subordinating clauses, complement clauses, free to-infinitives, and nominalizations as well (Webber et al., 2019), as illustrated in examples (2)-(4).

(2) [Treasurys opened lower] $_{\text{ARG1}}$, [reacting negatively to news]$_{\text{ARG2}}$

(3) ... [a number of project veterans were on hand to watch the launch]$_{\text{ARG2}}$ [to watch the launch]$_{\text{ARG1}}$

(4) ... many are hoping [for major new liberalizations]$_{\text{ARG2}}$ if [he is returned firmly to power.]$_{\text{ARG1}}$

Rhetorical Structure Theory (RST, Mann et al., 1992; Carlson et al., 2001), on the other hand, defines Elementary Discourse Units (EDUs) essentially as clauses. Their size is in principle arbitrary, but the units should have independent functional integrity (Mann and Thompson, 1988). Clausal subjects, complements and restrictive relative clauses are considered as parts of the clause headed by their governing verb. Applying this intuition consistently at scale requires extensive rule sets motivated by the inventory of discourse relations. For more extended examples of EDUs as well as a Figure comparing the alternative segmentation approaches discussed in this section with SE segmentation see Appendix A.

2.3 Automatic Discourse Segmentation

The first steps in neural approaches to automatic discourse segmentation were taken by Wang et al. (2018) in `NeuralEDUSeg`, applying a `BiLSTM-CRF` architecture and addressing data sparsity through pre-trained word embeddings plus a restricted self-attention mechanism. In the context of EDU segmentation, the `DISRPT` shared tasks have driven further progress on neural approaches. The 2019 winning system, `ToNy` (Muller et al., 2019), combined a `BiLSTM-CRF` with multilingual `BERT`-based sequence prediction across 15 corpora spanning RST, SDRT and PDTB, outperforming prior models on nearly all languages. Next, Gessler et al. (2021) extended ToNy by incorporating token-level handcrafted features such as POS tags and dependency relations. More recent winning systems at `DISRPT` 2023 by (Braud et al., 2023) and 2025 by (Lalitha Devi et al., 2025) build on using `XLM-RoBERTa` (Conneau et al., 2020). Braud et al. (2023) additionally experiment with freezing specific layers to separate morpho-syntactic from semantic encoding. As a contrast, using a generative language model, Nayak (2024) showed that zero-shot prompting of GPT-3.5 turbo is still not competitive with smaller pretrained language models explicitly trained for segmentation. Frenzel et al. (2026) provide the most recent overview on the development from rule-based systems to neural approaches, and similarly adopt fine-tuning `XLM-RoBERTa` as the segmentation model in their approach on German data.

Previous methods for automatic SE segmentation (Friedrich, 2017) add a postprocessing step to `SPADE` to convert RST EDUs to SE segments. The segmenter of the discourse parser `SPADE` (Soricut and Marcu, 2003) is based on a probabilistic model learned from the RST Discourse Treebank. For all words in the vocabulary, the probability for inserting a boundary after a word w is estimated from the treebank and depends on a lexicalized version of the corresponding sentence's syntactic tree (Magerman, 1995).

3 Segmentation Guidelines for Situation Entities

Smith (2003) suggests that SEs are introduced by the clauses of a text, while noun phrases introduce individuals (e.g., people, places, objects or ideas) and tense and time adverbs introduce times. She does not further specify which linguistic construction she considers to be a clause. Following the idea that discourse segmentation depends on linguistic units to which the target categories can meaning-

fully be assigned, in this work, we define and operationalize a segmentation scheme for SE types for English and German.

3.1 Situation Segments

We now describe which syntactic units can be assigned SE types in English, i.e., which constructions we assume to function as verb constellations.

Finite clauses such as Example (5) are the most clear-cut case. The finite verb is the SE's main verb, its grammatical subject is the main referent. Earlier work on SEs (Friedrich et al., 2016) is based on Stanford dependencies (de Marneffe and Manning, 2008), which mark participles as dependents of the finite auxilaries or modal verbs. In the more recent Universal Dependencies framework (de Marneffe et al., 2021), auxiliaries are dependents of the participles because those carry the semantic meaning of the verbs. In both the earlier work and our recent extensions, the main verbs are marked on the same spans as the participle.

(5) John$_{\text{MREF}}$ built$_{\text{MV}}$ a house.
Whales$_{\text{MREF}}$ are$_{\text{MV}}$ huge.

When the finite verb is an auxiliary, the main verb for the purpose of SE annotation is still the meaning-carrying verb even if this is non-finite (see Example (6)).

(6) He$_{\text{MREF}}$ may join$_{\text{MV}}$ us later.
This$_{\text{MREF}}$ won't help$_{\text{MV}}$.

The segmentation rule also applies to expletive sentences where "it" functions as an empty subject.

(7) It is raining.
It turned dark.

While there is no clear main referent in these cases, it is typically still possible to decide whether a statement is made about a class or kind (generic) or not, as in the segments listed in (7).

Relative clauses clearly introduce SEs.

(8) (a) [My brother$_1$], [who lives in Chicago$_2$], [is visiting this weekend$_1$].
(b) [The book$_1$] [that fell down$_2$] [was mine$_1$].

As illustrated in example (8), SE segments are not necessarily contiguous spans of text, which makes them stand out compared to the discourse segmentation schemes discussed in Section 2.2. For practical reasons, we split the complete text into segments, but not each of the segments will ultimately be assigned SE features. The noun phrases "My brother" and "The book" will simply be labeled with NO-SE, however, they syntactically function as the subject and as the main referent of the matrix clauses. Annotators are instructed to consider the underlying syntactic dependencies when deciding on SE labels. We opted for this approach due to its simplicity during annotation and also to emphasize that SE types capture the form of linguistic realizations rather than more abstract semantic notions as, e.g., in Abstract or Unified Meaning Representations (Banarescu et al., 2013; Gysel et al., 2021).

Participles are not considered to evoke SEs if their use is purely adjectival as in "The dancing girl." The reasoning behind this is that we aim for a granularity that roughly corresponds to clauses. By contrast, participle clauses are also considered to be segments, e.g., if they function as reduced relative clauses (9) or participle clauses indicating temporal (10) or causal (11) relations. As a general rule, preposed participles without modifiers or compliments tend to be read as adjectival and therefore do not constitute their own segment whereas postponed participles are rarely read in purely adjectival manner and therefore invoke a separate situation.

(9) (a) [The man] [talking to John] [is my brother].
(b) [The book] [written by Orwell] [is famous].

(10) (a) [Walking home], [I met Sarah].
(b) [Having finished], [I got up].

(11) (a) [The bomb exploded], [destroying the bridge].
(b) [Loving her], [he proposed].

Our decision to treat participle clauses but not participles in adjectival use as SEs is rooted in the idea that they are more similar to the other types of clauses we define. We acknowledge, however, that this is not a clear-cut case.

Gerunds that are used as nouns, e.g., "running" or "walking" in the full segment (12) are, analogously to the BDU approach, not considered to invoke their own SEs. These gerunds clearly refer to concepts instead of situations.

(12) Running burns more calories than walking.

Asher (1993) investigates how eventualities (states and events) and abstract entities (propositions, properties, states of affairs and facts) are referred to in natural language. He provides an inventory of sentential nominals, i.e., syntactic structures whose meanings are correlated with sentences: derived nominals (13), gerund phrases (14), that-clauses (15), for-infinitival phrases (16), naked infinitive phrases (3a) and noun phrases involving common nouns that may combine with that-clauses

or gerund phrases (17).

(13) (a) The army's destruction of the city
(b) Franklin's favorite invention
(14) (a) The mayor's throwing of the pizza
(b) John's hitting Bill
(c) The gathering of the pecans
(15) that Sam greeted Susan
(16) (a) For Kim to win was unexpected.
(b) John wanted for Mary to be chair.
(17) (a) Mary's doubt that John was unhappy
(b) The fact that John was unhappy
(c) The letter explaining the situation

Some of these constructions such as (13a) or (15) clearly refer to events while others may also refer to objects (13b). Of the above sentential nominals, we mark only that-clauses as SE segments. We do not treat the rest of the above sentential nominals as invoking SEs, but simply as part of the larger situation segment in which they are embedded. We decided not to mark them because the boundary between event-denoting constructions (13a) on the one hand and phrases denoting concrete (13b) or abstract objects (17) on the other hand is not clear; additional annotation guidelines would be necessary.

Conjunctions and conditionals. When a clause starts with a conjunction or a subordination, the conjunction or subordination is segmented into the same span as the clause that they introduce.

(18) (a) [I hate] [and love him].
(b) [I believe] [that she called him].
(c) [I left] [after I had called him].
(d) [She left] [because I had called him].
(e) [I like to sit] [where the sun shines].
(f) [I like to run] [if the sun shines].

Please note that our general segmentation depends on whether aspectual features can be assigned to the verb construction expressed by a span of text. It does not matter which role the sub-clause plays in the argument/modifier structure of the embedding verb, e.g., in (18b), the embedded segment is a complement clause, while (18c) and (18d) are adjunct clauses.

To-infinitives are not considered to introduce SE segments (Friedrich et al., 2015b). In English, to-infinitives can fulfill different functions. In (19a) and (19b), the to-infinitives are predicative nominalizations and refer to abstract concepts. To-infinitives can also indicate purpose (19c) or function as adjectival or other complements as in (19d) and (19e). By definition, we assume these cases to be predicative uses rather segments that introduce their own situation entities.

(19) (a) [To travel alone can be exciting.]
(b) [My dream is to study abroad.]
(c) [She went to the library to study.]
(d) [They are happy to be here.]
(e) [He grew up to become a teacher.]

The core idea here is that SEs annotate the aspectual forms as chosen by the writer. While this does require semantic interpretations, we aim not to reformulate too much, as this introduces too many degrees of freedom.[2]

3.2 Extension to German

In historical language data, sentence boundaries cannot always be identified based on punctuation, as this differs significantly from that of modern English and German. In the case of the GiesKaNe corpus, this problem was already resolved during corpus creation, and sentence-level annotation, in accordance with the annotation guidelines, does not take place at the orthographic level. Sentences are thus determined grammatically, not by punctuation (Ágel and Henning, 2023). German syntax from the 17th to the 19th century also poses particular challenges for SE segmentation. The following subsections describe the resulting specific adjustments to the annotation scheme with regard to two selected phenomena. As work continues with additional data, further phenomena may be identified.

Participle clauses. In contemporary German, as in NHG, participles can be used in attributive function, in which case they are not classified as segments. In contrast to contemporary German, it is much more common in NHG to have participial clauses where the finite auxiliary is missing, also known as "auxiliary ellipsis" (Breitbarth, 2005; Thomas, 2018). This is actually a phenomenon of Early NHG, but its residual effects are still relevant for NHG data. Participial clauses without finite auxiliaries can be classified as segments, as has already been argued for English. In many cases, the heuristics from English do not work due to the less restricted word order (Ágel, 2015, 2000) and/or the missing finite verb in NHG. Even if parts of sentences are recognizable as reduced relative clauses (20a) or as clauses with auxiliary ellipsis (20b), only in some cases does the corresponding word order allow for segmentation in the NHG data.

[2] A special case where we allow minor hypothetic reformulations are participle clauses, e.g. *Sitting on the bench, I looked at the beach. – While I was sitting on the bench, I looked at the beach.*

(20) (a) [So sieht man denn selbst so genannte Gebildete, — subjektiv stumpf, objektiv peinigend — ... die lieblichsten, [durch die Natur vorgebildeten,] [durch die Kochkunst veredelten Produkte] naturalistisch und roh sich aneignen].[3] (Anthus, 1838)

(b) [Er erschiene unsäumig;] [und als Sie sich mit ihm an ein Fenster gesteuret] / [fragte Sie][4] (Birken, 1652)

Final clause construction. To mark final clause constructions in German, we follow Mavridou et al. (2015). German provides the construction of final clauses with "damit" or "so ... dass" ("such/so that"), indicating a purpose or goal (21a) or an actual event (21b).

(21) (a) [Erinnere mich nochmal,] [damit ich pünktlich komme.][5]

(b) [Da stopft Einer gedankenlos ... so viel Brod in den Mund,] [daß er unmöglich den spezifischen Geschmack irgend einer Speise perzipiren kann.][6] (Anthus, 1838)

As the subordinated clauses are final, we treat them as separate segments containing the conjunctions introducing the clause.

3.3 Extension to LModE

In addition to covering peculiarities of German with regards to segmentation we further provide a brief description of a special case arising when annotating LModE.

Absolute constructions. One significant difference between LModE and Present Day English is the higher frequency of use of the absolute construction (Van De Pol and Petré, 2015), a common boundary case. Example (22) begins with such a construction.

(22) [The Parliament being met on the 23d October,] [his Majesty refer'd them to what he had said to both Houses four Days before.] (Boyer, 1702)

These constructions present their own SEs without finite verb morphology or subordinating conjunctions. Their syntactic independence suggests separate segmentation, yet they can function as adverbial modifiers of the main clause. The lack of explicit conjunctions makes their relationship to the main clause SE less transparent than in equivalent finite subordinate clauses (*When the parliament had met...*). These purpose constructions occupy an intermediate position between arguments and adjuncts. They describe potential or foundational situations rather than actualized events, yet lack the full clausal structure of *that*-complements or *so that* result clauses. Despite this and due to their syntactic independence they are marked as separate segments in this project.

4 Annotation Study

In order to to extract relevant underspecified edge cases, and to evaluate IAA as well as model performance, we conducted a human annotation study spanning four language varieties, contemporary and historical English and German, using excerpts from written corpora.

4.1 Data Sources

Four datasets are used for this annotation study, corpus statistics including the number of annotated tokens per variety can be found in Table 2.

The Corpus of English Historiography (CLMEH (Claridge, 2025)) is used as the basis for segmentation of historical English text. For the manual annotation, two snippets from two different LModE authors, Abel Boyer and Thomas Salmon, were annotated. These texts were chosen since they are among the earliest texts in the corpus (published in 1702 and 1736 respectively), and therefore serve as a stress-test of the annotation guidelines.

The NHG data is a subset of the GiesKaNe Korpus (Justus-Liebig-Universität Gießen, 2022), similarly consisting of early texts by Johann Joachim Becher (1668) and Antonius Anthus (1838).

For modern German we use snippets from German Wikipedia articles as well as four blog articles downloaded from publicly available and CC-BY-SA-licensed blogs as listed in DWDS (Barbaresi and Würzner, 2014).

Lastly, to estimate human agreement for contemporary English, we also annotate present-day English based on a subset of four files from the the MASC+wiki dataset used in Friedrich (2017), one travel blog, a news story, an email, and an English Wikipedia page. Appendix C contains information on the text sources including author names where

[3] English: [Thus one sees even so-called educated people , — subjectively dull, objectively tormenting — ... appropriate the loveliest products, [pre-formed by nature,] [refined by the culinary arts] naturalistically and crudely.] Note that in the original, the descriptions of the products are preposed.

[4] English: [He appeared without delay;] [and when steered herself with him to a window] / [she asked]

[5] English: [Remind me again] [so that I arrive on time.]

[6] English: [There someone thoughtlessly stuffs so much bread into their mouth with every bite] [that they cannot possibly perceive the specific taste of any dish.]

available, as well as the year of publication for each manually annotated document, Table 2 contains the token counts per variety.

Language Variety	# Tokens	# Segs	Avg. Tok/Seg	Max. Tok/Seg
Current E	3,937	482	10.3	42
LModE	6,797	614	13.6	56
Current G	4,751	444	12.6	64
NHG	5,666	617	11.4	67

Table 2: Volume of annotated data per language variety.

4.2 Inter-Annotator Agreement

We collect two independent human segmentation annotations per language variety, which we call A and B[7]. The human segmentation was performed by five native German speakers as annotators, three authors of this paper and two paid student annotators. All annotators underwent several weeks of specific and language-/variety-dependent segmentation training. We evaluate annotation reliability using exact span match agreement, computed as the proportion of segments in a reference annotation that are recovered with identical span boundaries in the comparison annotation. Annotators were trained jointly but annotated independently. Results can be found in Table 3.

Agreement is computed over span boundaries derived from pre-tokenized BI representations using the same tokenizer as for model training (see Section 5). As the task is unitizing (segmentation without labels) over continuous text, standard chance-corrected measures such as Krippendorff's α_U are not applicable (Artstein and Poesio, 2008). The left column of Table 3 provides annotation statistics and agreement scores across all four varieties. Overall average agreement reaches 79.1%, with all varieties falling within a narrow 5-point range, supporting the general robustness and cross-linguistic transferability of the guidelines. Contemporary English and Contemporary German show the largest exact span match percentages – on average 81.4% and 80.2% respectively. As expected, agreement for historical variants is harder to achieve and sits at 76.86% on average for NHG and 77.98% for LModE.

To gain a better understanding of the root causes of all disagreements, they were explored by one of the authors and manually assigned a disagreement group and label.

A closer look at the inter-annotator disagreement (see Figure 3 in Appendix C for an exact breakdown disagreements) reveals that inconsistencies concentrate in syntactic and discourse-related phenomena. Across all four conditions, *clause missed* — scenarios where a clause containing a valid meaning-carrying verb was disregarded — constitutes the single most frequent source of disagreement. As explored above, the NHG condition narrowly shows the largest overall disagreement counts, with particularly high occurrences of *clause missed* disagreements and its sub-category of missed *relative clause*s. We attribute this to phenomena such as frequent (auxiliary) ellipses, long sentences containing multiple clause chains that are unusual to the modern annotator, and multiple valid boundary placement options due to a less restricted word order.[8] Current German disagreements cluster around *inlay* and *arbitrary boundary* placement since word order is less restricted leading to multiple valid splitting points within long sentences. Current English exhibits elevated disagreement for *verbless* and *participle* constructions alongside *arbitrary boundary* decisions, pointing to uncertainty in identifying implicit or non-canonical clause structures. LModE shows a similar profile, with persistent disagreement in *clause missed* and *verbless* constructions, as well as the discourse-adjacent category of *absolute constructions*[9]. Orthographic disagreements remain rare across all datasets. These findings suggest that disagreement primarily arises from complex syntactic boundary decisions and discourse-level interpretation, and is amplified in historical data where structurally ambiguous constructions are more frequent.

[7]Please note that the A and B labels do not always refer to the same human annotators

[8]See this excerpt from Anthus as an Example: *Wie aber die Gelehrten noch dar über schwitzen , zu bestimmen , wo die Pflanze zu m Thier wird , wo die Grenzen des Pflanzen- und Thierreiches fest zu stecken seien , eben so schwierig ist es , zu bestimmen , welcher unteren Thierreihe man zuerst die Fähigkeit eines eigentlichen Essens zu zu gestehen hat.* English: *But as the scholars still sweat over how to determine, where the plant becomes the animal, where the borders of the plant and animal kingdoms are to be placed, as hard is it, to determine, to whom amongst the order of animals one should first grant the property of a true food.*

[9]See this excerpt from Boyer (LModE) as an example: *On the 29th of October 1689, came on the usual Solemnity of the Lord Mayor of London, and Sir Thomas Pilkinton being continued for the Year 1690, and the King and Queen, the Prince and Princess of Denmark, and both Houses of Parliament, having been pleas'd to accept his Invitation to his Dinner, their Majesties attended by their Royal Highnesses, [...].*

Variety	Inter-annotator agreement (exact match) Ann. B vs. A	Ann. A vs. B	Model performance (exact match) Model vs. Ann. A	Model vs. Ann. B
Contemporary G	76.48±14.25	83.92±11.83	69.32±12.22	64.20± 6.73
NHG	78.36± 0.75	75.36±11.56	42.68±30.51	48.87±38.06
Contemporary E*	80.62±16.69	82.15±12.25	77.06± 8.58	86.91± 9.36
LModE	79.34± 1.04	76.62± 4.88	64.28± 3.21	67.50± 2.90

Table 3: Inter-annotator exact span match (%) per variety. Agreement is reported once with annotator A treated as the reference (gold) annotation and once with annotator B treated as reference,i.e., the first column shows how many of A's annotations were exactly matched by B and the second column vice versa. Model performance as exact span match (%) versus human annotator A and B. *As the model's training set contains the test documents for Contemporary English, we trained a separate model for this setting without these test documents in the training data.

5 Automatic Segmentation Model

We implement a new SE segmentation model that does not rely on legacy systems, using about 40,000 labelled English SE segments provided by Friedrich et al. (2016).[10] The training data includes texts from a variety of genres taken from the Manually Annotated Subcorpus of the OANC (MASC) (Ide et al., 2008) and Wikipedia. The original rule-based segmentation model was based on a legacy system (Soricut and Marcu, 2003) which itself had been trained on only just over 7,000 sentences. While we do not yet have sufficient training data for German and historic English variants, we apply our model to these variants in a zero-shot way, achieving promising results.

5.1 Modeling

`XLM-RoBERTa` (Conneau et al., 2020) has proven itself a robust baseline for EDU segmentation not only for English (Braud et al., 2023; Lalitha Devi et al., 2025), but also for German data (Frenzel et al., 2026). Thus, we similarly treat segmentation as a sequence tagging task and fine-tune `XLM-RoBERTa` (Conneau et al., 2020), a multilingual transformer-based encoder model producing contextualized token representations. We preserve the predefined train-test splits on document level and sample a development set from the training split across documents, resulting in a final train/dev/test ratio of 0.68/0.10/0.22. First, we segment each document into sentences using `spaCy`'s (Explosion AI, 2025) `en_core_web_sm` pipeline and treat each sentence as an independent input. [11] To mark segment boundaries, we use a `BI` sequence tagging scheme. Following common practice, we backpropagate the loss for each first `SentencePiece` (Kudo and Richardson, 2018) token corresponding to a `spaCy` token. We fine-tune `XLM-RoBERTa-large` with a linear classification head and a Conditional Random Field (CRF) (Lafferty et al., 2001) decoding layer.

Metric	Avg.	SD
B F1	90.5	0.1
B Precision	90.4	0.7
B Recall	90.6	0.8
Exact Match	74.3	0.3
WindowDiff	7.7	0.1

Table 4: Situation segmentation performance (detecting start boundary tokens) on the SitEnt test set using `XLM-RoBERTa-large`, averaged over 5 random seeds. This test dataset had been single-annotated by Friedrich et al. (2016); the type of data is most similar to Contemporary E, where human agreement reached up to 82.15 % exact match.

5.2 Evaluation Metrics

Next to precision, recall and F1 score, we also report exact matches and WindowDiff (Pevzner and Hearst, 2002) using the NLTK implementation (Bird and Loper, 2004). WindowDiff evaluates segmentation quality by sliding a fixed-size window over both the reference and predicted boundary sequences and comparing the number of boundaries within each window. Instead of requiring exact boundary matches, it penalizes differences in local boundary counts. See Appendix D for details on hyperparameters and training procedure.

5.3 Experimental Results

Results for `XLM-RoBERTa-large` on the SitEnt test set are reported in Table 4. The averaged results over five random seed runs show a well-

[10] https://github.com/annefried/sitent

[11] This modular design ensures that, in future work, we can train sentence segmenters using any sentence-segmented labeled data, which is important particularly for the historical language variants, where training data is limited and conventional tokenization may fail, allowing to fine-tune segmenters on smaller, sentence-aligned datasets.

balanced trade-off between precision and recall, with a slight tendency of the model to overpredict segments. While Exact Match is considerably lower than the boundary detection scores, the low WindowDiff suggests that most errors stem from minor boundary misalignments (e.g., slightly shifted EDU boundaries or punctuation handling), rather than complete segmentation failures.

In the next step, we apply our best performing model on boundary F1 across the hyperparameter grid to annotate the datasets described in Section 4. Table 3 reports exact matches between the model output and each of the two independent human annotations per language variety. Exact match is again calculated as in the IAA evaluation above. For each document, the model annotation is compared independently against each of the two human annotations, yielding two exact match scores per document.

5.4 Discussion of Computational Experiments

The results in Table 3 reflect the expected degradation in model performance the more the language variety diverges from the model's training domain of contemporary English. On Contemporary E, the model reaches mean exact match scores of 77.06% (vs. Ann. A) and 86.91% (vs. Ann. B), approaching human agreement levels (IAA = 81.39%), proving that the model captures segmentation patterns in in-domain data reasonably well. It is much closer to Ann. B who has been part of the training data annotation. Applying the model cross-lingually to Contemporary G reveals a substantial drop to 69.32/64.20%, despite human IAA remaining stable at 83.92%.

While human annotators remain highly consistent in German, the model fails to fully transfer its segmentation knowledge across languages. For the historical varieties a similar pattern of degradation becomes apparent. For LModE, the model achieves 64.28%/67.50%, compared to a human IAA of 79.34%/76.62%, reflecting a certain difficulty with historical English. A low cross-document variability (SD = 3.21/2.90) suggests that this difficulty is consistent across texts. Contemporary G and LModE show highly similar model performance ranges, suggesting that cross-lingual variation in contemporary German and diachronic variation in Late Modern English impose a comparable level of difficulty for the model. NHG represents the most challenging setting for the model, combining cross-lingual and temporal domain shift, leading to the lowest exact match at 42.68/48.87%, against a human IAA of 78.36%/75.36. A high cross-document variability (SD = 30.51/38.06) indicates a substantial variation across the annotation of the two evaluated texts. While human annotators remain relatively consistent, the model fails to recover more than half of the same boundaries under this combined condition.

6 Conclusion

We have presented principled SE segmentation guidelines for contemporary and historical varieties of English and German, filling a gap left by prior work which relied largely on implicit or underspecified segmentation criteria. Our inter-annotator agreement study demonstrates substantial and consistent human agreement across all language varieties, with average exact match scores ranging between 76.86% and 81.4%. This confirms that the guidelines are, in principle, learnable and robust even for syntactically challenging historical texts. Our new `XLM-RoBERTa`-based segmenter achieves a boundary F1 of 90.5 on contemporary English, matching human agreement on its training domain. However, model performance drops substantially under cross-lingual and temporal domain shift, most severely for historical German, where exact match falls to 45.78%. Although the segmentation task itself is consistent across varieties, the bottleneck lies in the availability of in-domain training data. We release our guidelines, annotated data, and model code to support future work on multilingual and diachronic discourse analysis.

Acknowledgement

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the BayernKI project v110ee. BayernKI funding is provided by Bavarian state authorities. We thank Manfred Pinkal and Alexis Palmer for their helpful discussion related to this annotation scheme. We also extend our thanks to the annotators as well as to the anonymous reviewers for their constructive and useful feedback.

References

Jean-Michel Adam. 2011. *Les textes: types et prototypes: récit, description, argumentation, explication et dialogue*. Armand Colin.

Vilmos Ágel. 2000. Syntax des neuhochdeutschen bis zur mitte des 20. jahrhunderts. In Werner Besch, Anne Betten, Oskar Reichmann, and Stefan Sonderegger, editors, *Sprachgeschichte, Part 2*. Walter de Gruyter, Berlin, New York.

Vilmos Ágel. 2015. Die umparametrisierung der grammatik durch literalisierung. online- und offlinesyntax in gegenwart und geschichte. In Ludwig Eichinger, editor, *Sprachwissenschaft im Fokus*, pages 121–156. DE GRUYTER.

Vilmos Ágel and Mathilde Henning. 2023. Annotationshandbuch des dfg-projekts syntaktische grundstrukturen des neuhochdeutschen. zur grammatischen fundierung eines referenzkorpus neuhochdeutsch.

Antonius Anthus. 1838. *Vorlesungen über Esskunst.* Wigand, Leipzig.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*, volume 50 of *SLAP*. Kluwer, http://www.wkap.nl/.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Adrien Barbaresi and Kay-Michael Würzner. 2014. For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *Proceedings of the NLP4CMC Workshop (KONVENS 2014)*, pages 2–10. Hildesheim University Press.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27.1:3–43.

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.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Sigmund von Birken. 1652. *Die Fried-erfreuete Teutonje.* Dümmler, Nürnberg.

Abel Boyer. 1702. *The History of King William the Third: In III Parts*. Printed for A. Roper and F. Coggan, London.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Anne Breitbarth. 2005. *Live fast, die young: The short life of Early Modern German auxiliary ellipsis: Zugl.: Tilburg, Univ., Diss., 2005*, volume 115 of *LOT*. LOT, Utrecht.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, Information Sciences Institute (ISI), Marina del Rey, CA.

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. *Preprint*, arXiv:1911.02116.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

David Denison. 1999. SYNTAX. In Suzanne Romaine, editor, *The Cambridge History of the English Language*, 1 edition, pages 92–329. Cambridge University Press.

Explosion AI. 2025. spacy: Industrial-strength natural language processing in python.

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. 2017. *States, events, and generics: computational modeling of situation entity types*. Ph.D. thesis, Universität des Saarlandes.

Annemarie Friedrich, Kleio-Isidora Mavridou, and Alexis Palmer. 2015a. *Situation Entity Types Annotation Manual*.

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015b. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 21–30, Denver, Colorado, USA. Association for Computational Linguistics.

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.

Annemarie Friedrich and Manfred Pinkal. 2015a. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal. Association for Computational Linguistics.

Annemarie Friedrich and Manfred Pinkal. 2015b. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1272–1281, Beijing, China. Association for Computational Linguistics.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jens Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, W. Bruce Croft, Chu-Ren Huang, Jan Hajic, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35:343 – 360.

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.

Judith L. Klavans and Martin Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Ottilie Koralek. 1889/1890. Lamentatio intermissa I. Tagebucharchiv Emmendingen. Unpublished Transcription (Hollmann).

Manfred Krifka, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An introduction. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 1–124. University of Chicago Press.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning.*, page 282–289.

Sobha Lalitha Devi, Pattabhi Rk Rao, and Vijay Sundar Ram. 2025. SeCoRel: Multilingual discourse analysis in DISRPT 2025. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 79–86, Suzhou, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

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

William C. Mann, Christian M.I.M. Matthiessen, and Sandra A. Thompson. 1992. Rhetorical Structure Theory and Text Analysis. In William C. Mann and Sandra A. Thompson, editors, *Pragmatics & Beyond New Series*, volume 16, page 39. John Benjamins Publishing Company.

Thomas A. Mathew and Graham E. Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Proceedings of the Sixth Midwest Computational Linguistics Colloquium*, Bloomington, Indiana. Indiana University.

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.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Kota Nayak. 2024. Does chatgpt measure up to discourse unit segmentation?a comparative analysis utilizing zero-shot custom prompts. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111, Sofia, Bulgaria. Association for Computational Linguistics.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Livia Polanyi. 1995. The linguistic structure of discourse. Csli 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.

Thomas Salmon. 1724. *A Review of the History of England. In Two Volumes*, 2 edition, volume 1. Printed for Charles Rivington, London.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. 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.

William Stubbs. 1891. *The Constitutional History of England, Vol. I*. Oxford at the Clarendon Press, Oxford. Digitized by the Digital Library of India (Osmania University).

Victoria Thomas. 2018. *Auxiliary Ellipsis in Early Modern German 1350-1800*. Phd, The University of Manchester, Manchester.

Erik F. Tjong, Kim Sang, and Hervé Déjean. 2001. Introduction to the CoNLL-2001 shared task: Clause identification. In *Proceedings of the 2001 Workshop on Computational Natural Language Learning (CoNLL)*, Manchester, UK.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 77–80, Singapore.

Nikki Van De Pol and Peter Petré. 2015. Why is there a Present-Day English absolute? *Studies in Language*, 39(1):199–229.

Zeno Vendler. 1957. *Linguistics in Philosophy*, chapter Verbs and Times. Cornell University Press, Ithaca, New York.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Egon Werlich. 1989. *Typologie der Texte*. UTB für Wissenschaft.

A Discourse Segmentation Guidelines in Related Work

This section takes a closer look at related work on discourse segmentation including more extensive examples. In the Linguistic Discourse Model (LDM; Polanyi (1995)), the atomic units of discourse are called Basic Discourse Units (BDUs); they are segments that have the potential to establish an anchor point for future attachment of other segments. Polanyi et al. (2004) start by defining the semantic basis for functioning as a segment and then identify syntactic constructions that are able to carry the semantic information needed for discourse segment status. They observe that in written text, often a subsequent but not necessarily adjacent segment continues the development of material introduced in a sub-sentential, often subordinate, clause. Example (23) shows an LDM annotation which treats the post-modifier "braying" next door as a separate segment, as it can be interpreted as "who was braying next door."

(23) [The donkey [braying next door] was loud.]

Nominal gerunds "Singing" and nominalizations "rationalization" are not always considered BDU segments. However, Polanyi et al. consider the nominal "the destruction of the old town hall" to be a BDU. To the best of our knowledge, no guidelines exist for distinguishing these cases.

Rhetorical Structure Theory (RST; Mann et al. (1992), presents another approach to discourse segmentation. Here, the segments are called Elementary Discourse Units (EDUs). EDUs are defined essentially as clauses, but clausal subjects, complements and restrictive relative clauses are considered as parts of the clause headed by their governing verb. When building the RST Discourse Treebank, Carlson et al. (2001) note that applying this intuitive notion is difficult when aiming for a large and consistently annotated corpus. They develop an extensive set of rules for identifying EDUs based on syntactic constituents with the aim of obtaining a balance between tagging granularity and the ability to identify units consistently (Carlson and Marcu, 2001). The rules are motivated by RST's inventory of discourse relations (schemas). For example, while infinitival constructions are generally not considered to constitute their own EDU, they are if they introduce a purpose clause as in (24) because the infinitival clause corresponds to the satellite of a Purpose relation here. Prepositional phrases with clausal objects (25) are EDUs, while other non-finite clausal objects are not.

(24) [... officials at Southern Co. conspired to cover up their accounting for spare parts] [to evade federal income taxes.]

(25) [Canadian Utilities isn't alone] [in exploring power generation opportunities in Britain.]

For ease of comparison, we provide an overview of the differences between SE segmentation and other discourse segmentation schemes in Figure 2. A core difference is that SE segmentation is motivated by the ability of assigning aspectual features, i.e., situation entity types (Smith, 2003), to a segment, while RST, LDM, and PDTB are more concerned with whether and what discourse relations can be identified between segments. Segmentation in the PDTB can be either finer-grained or coarser-grained than SE segmentation; LDM and RST apply either the same or a more fine-grained segmentation as illustrated in Figure 2. Connectives typically constitute a separate data type and are not included in the arguments, while SE segmentation performs an exhaustive text segmentation, simply grouping them with the clause they introduce.

B Background on Linguistic Framework and Annotation Guidelines

Table 5 shows the full inventory of SE types including their description. SE annotation into these classes is a downstream task building on SE segmenation. In a similar vein, Table 6 defines the relationship between the main verb's and the main referent's aspectual features and the SE type of the entire segment.

C Manually Annotated Data

Table 7 shows the data source for each manually annotated file as well as its publication year. Token counts are available in Table 2. Figure 3 shows the inter-annotator disagreement analysis across annotation conditions and linguistic categories. Darker cells indicate higher disagreement frequencies.

SE type	Description	Example
Eventualities		
STATE	introduce properties	The colonel owns the farm.
EVENT	happenings	John won the race.
REPORT	for attribution	"…", said Obama.
General Statives		
GENERIC SENTENCE	generalizations over kinds	The lion has a bushy tail.
GENERALIZING SENTENCE	habituals; generalizations over situations	Mary often fed the cat last year.
Abstract Entities		
FACT	clausal complements of verbs of knowledge	I know that she refused the offer.
Proposition	clausal complements of verbs of belief	I believe that she refused the offer.
QUESTION		Who wants to come?
IMPERATIVE		Hand me the pen!

Table 5: Inventory of SE types, as adapted from Smith (2003) in previous work (Friedrich et al., 2016; Friedrich and Palmer, 2014b).

SE type	main referent	aspectual class	habituality
EVENT	non-generic	dynamic	episodic
	generic		
STATE	non-generic	stative	static
GENERIC SENTENCE	generic	dynamic	habitual
		stative	static, habitual
GENERALIZING SENTENCE	non-generic	dynamic	habitual
		stative	

Table 6: SE types and their corresponding features, adapted from Friedrich (2017).

Variety	Text	Source	Year
Contemp. E*	Email	MASC	1993
	News WSJ0135	MASC	>1990
	Travelblog Dublin	MASC	<2008
	Trees	Wikipedia	<2016
LModE	Historiography - Boyer	CLMEH	1702
	Historiography - Salmon	CLMEH	1736
Contemp. G	bitblokes	Blog	2023
	phantanews	Blog	2026
	iphoneblog	Blog	2025
	literaturblog	Blog	2017
	Literatur	Wikipedia	2026
	Wald	Wikipedia	2026
NHG	Lecture - Anthus	GiesKaNe	1838
	Academic - Becher	GiesKaNe	1668

Table 7: Text sources used in the annotation study. *Contemporary E texts are excluded from model training.

D Computational Modeling

Training is performed using the AdamW optimizer (Loshchilov and Hutter, 2019) with default hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, a batch size of 64, a maximum of 20 epochs, and early stopping after three non-improving epochs based on the boundary (B) tag's F1 score. A small grid search is conducted over five random seeds, learning rates (4e−5, 5e−5), and weight decay values (0.001, 0.005) to select the best model. We use linear learning rate decay with a 10% warmup phase, gradient clipping with a maximum norm of 1.0, and BF16 mixed precision training on four NVIDIA H100 GPUs. The best individual run reached an F1 of 90.65, converging after 8 epochs with a learning rate of 4×10^{-5} and weight decay 0.005, random seed 100.

WindowDiff We calculate the WindowDiff (Pevzner and Hearst, 2002) using the NLTK implementation (Bird and Loper, 2004) as follows: Given a sentence of length N with reference boundaries r_i and predicted boundaries h_i (where $r_i = 1$ if token i starts a new segment, 0 otherwise), we:

1. Set the window size to
$$k = \max\left(1, \left\lfloor \frac{N}{2B} \right\rfloor\right),$$
where B is the total number of reference boundaries in the sentence.

2. Slide a window of size k across the sentence, comparing the boundary count in each win-

PDTB vs. SE:	... [and many are hoping [for major new liberalizations]$_{ARG2}$]$_{SE}$ [if [he is returned firmly to power.]$_{ARG1}$]$_{SE}$
PDTB vs. SE:	[[Here in this new center for Japanese assembly plants just across the border from San Diego,]$_{SE}$ [turnover is dizzying, infrastructure shoddy, bureaucracy intense.]$_{SE}$ [Even after-hours drag;]$_{SE}$ ["karaoke" bars,]$_{SE\dagger}$ [where Japanese revelers sing over recorded music,]$_{SE}$ [are prohibited by Mexico's powerful musicians union.]$_{SE}$]$_{ARG2}$ [Still, [20 Japanese companies,]$_{SE\ddagger}$ [including giants such as Sanyo Industries Corp., Matsushita Electronics Components Corp. and Sony Corp.]$_{SE}$ [have set up shop in the state of Northern Baja California.]$_{SE}$]$_{ARG1}$
LDM vs. SE:	[[California elected Schwarzenegger]$_{BDU}$ [governor]$_{BDU}$.] $_{SE}$
LDM vs. SE:	[[The donkey]$_{SE\dagger}$ [[braying next door]$_{SE}$] [was annoying.]$_{SE}$]$_{BDU}$
RST vs. SE:	[[The company announced]$_{SE}$]$_{EDU}$ [[that it will shut down its plant]$_{SE}$]$_{EDU}$ [[and dismiss several hundred employees.]$_{SE}$]$_{EDU}$
RST vs. SE:	[[The company plans to shut down its plant]$_{SE}$ [and dismiss several hunded employees.]$_{SE}$]$_{EDU}$

Figure 2: Comparison of SE segmentation to RST, LDM, and PDTB. † are examples of segments that are not complete SEs on their own, but they also do not belong to the following situation (unless by coreference relations). As illustrated in Figure 1, they can still contain main referents of other segments and hence be part of another segment as defined by syntactic dependency structure.

dow:

$$\mathrm{WD} = \frac{1}{N-k}\sum_{i=1}^{N-k} \mathbf{1}\left(\sum_{j=i}^{i+k-1} r_j \neq \sum_{j=i}^{i+k-1} h_j\right)$$

The score ranges from 0 (perfect match) to 1 (complete mismatch). Unlike exact match, WindowDiff tolerates small boundary shifts: if a boundary is off by one or two positions, only the windows containing that region are penalized, not the entire sentence.

We compute WindowDiff per sentence and average across all sentences with $N \geq 2k+1$, where M is the number of such sentences:

$$\mathrm{WD}_{\text{final}} = \frac{1}{M}\sum_{s=1}^{M} \mathrm{WD}^{(s)}.$$

E Additional Annotation Peculiarities and Annotation Examples

E.1 Late Modern English

Relative clauses Segmentation decisions are affected by variation in relative clause formation that can be observed in Late Modern English. The use of *which* for human antecedents is declining in this period and zero relatives are used more commonly in their place (Denison, 1999), which creates ambiguity in identifying clause boundaries. Examples (26a) and (26b), taken from Stubbs (1891), illustrate this variation. Example (26b) shows the special case of reduced relative clauses in the past tense as discussed in the general guidelines for English.

(26) (a) [The Saxons, Angles, and Jutes,][although speaking the same language,][worshipping the same gods][and using the same laws,][had no political unity like the Franks of Clovis;]

(b) [the Saxons in Germany were still a pure nationality,][unconquered by the Franks,][untainted by Roman manners,][and still heathen.]

Since the Late Modern English period covers an ongoing usage change with regards to relative clauses and their positions, the guideline for using preposition or post-position to approximate adjectival use cannot be relied upon. Segmentation, instead, is based on the annotators perception given the larger semantic embedding.

E.2 Late Modern English

Figure 4 reveals how he model treats an entire sentence as a single segment, failing to introduce a boundary at the transition to the main clause. This error highlights a limitation of the model in handling subordination structures, particularly in historical German, where clause boundaries may be less clearly signaled by punctuation or conjunctions. Whilst humans reliably segment along clause

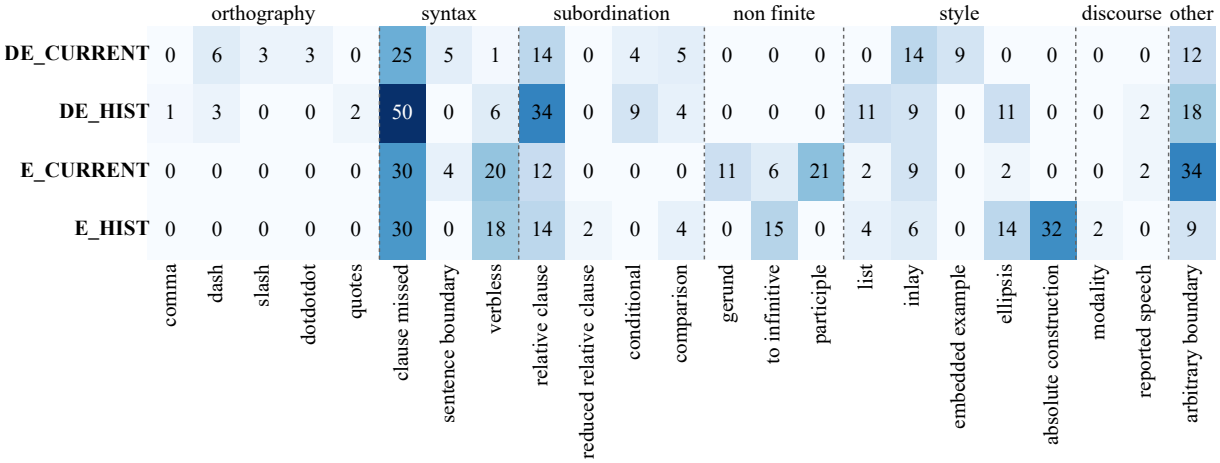

Figure 3: Inter-annotator disagreement analysis across annotation conditions and linguistic categories. Darker cells indicate higher disagreement frequencies.

boundaries, the model under-segments here. Figure 5 shows one further NHG example that illustrates a genuinely non-trivial case due to ellipses and multiple coordinated and subordinated structures where inter annotator differences can be seen as valid alternative annotations.

E.3 Contemporary German

Figure 6 illustrates a case of annotator segmentation disagreement involving the German coordinating conjunction “und.” Annotator 1 and the model introduce a boundary at the conjunction, treating the coordinated clauses as separate SEs, whereas Annotator 2 merges the coordinated clauses into a single segment. The example highlights a linguistically motivated source of variability in SE annotation: coordinated clauses can plausibly be interpreted either as independent segments or as a unified discourse unit.

Token *English*	Daß *That*	bei *in*	den *the*	Pflanzen *plants*	von *of*	keinem *no*	eigentlichen *actual*	Essen *food*
Model	B	I	I	I	I	I	I	I
Annotator 1	B	I	I	I	I	I	I	I
Annotator 2	B	I	I	I	I	I	I	I

Token *English*	die *the*	Rede *talk*	sein *be*	kann *can*	, *,*	**wird** ***will***	sich *itself*	aus *from*
Model	I	I	I	I	I	**I**	I	I
Annotator 1	I	I	I	I	I	**B**	I	I
Annotator 2	I	I	I	I	I	**B**	I	I

Token *English*	dem *the*	Begriffe *concept*	des *of*	Essens *eating*	später *later*	ergeben *result*	. .
Model	I	I	I	I	I	I	I
Annotator 1	I	I	I	I	I	I	I
Annotator 2	I	I	I	I	I	I	I

Figure 4: Example of segmentation disagreement on a subordinated clause. The model treats the entire sentence as a single segment, while human annotators introduce a segment boundary at the main clause (“wird”).

Token *English*	[...] *,*	sehen *we*	wir *see*	auf *on*	jener *that*	ersten *first*	Stufe *level*	Wesen *beings*	, *,*	welche *which*	, *,*	mit *with*
Model	I	I	I	I	I	I	I	I	I	**B**	I	I
Annotator 1	I	I	I	I	I	I	I	I	I	**B**	I	I
Annotator 2	I	I	I	I	I	I	I	I	I	**B**	I	I

Token *English*	Ausnahme *exception*	des *of*	Salzes *salt*	und *and*	Wassers *water*	, *,*	weder *neither*	eßbar *edible*	sind *are*	noch *nor*	essen *eat*	; *;*
Model	I	I	I	I	I	I	I	I	I	I	I	I
Annotator 1	I	I	I	I	I	I	I	I	I	I	I	I
Annotator 2	I	I	I	I	I	I	I	I	I	I	I	I

Token *English*	auf *on*	der *the*	zweiten *second*	eßbare *edible*	, *,*	aber *but*	nicht *not*	essende *eating*	Wesen *beings*	; *;*	auf *on*	der *the*
Model	I	I	I	I	I	I	I	I	I	I	I	I
Annotator 1	**B**	I	I	I	I	I	I	I	I	I	I	I
Annotator 2	I	I	I	I	I	I	I	I	I	I	I	I

Token *English*	dritten *third*	Stufe *level*	endlich *finally*	Wesen *beings*	, *,*	welche *which*	essen *eat*	und *and*	gegessen *are eaten*	werden *become*	. .
Model	I	I	I	I	I	I	I	I	I	I	I
Annotator 1	I	I	I	I	I	**B**	I	**B**	I	I	I
Annotator 2	I	I	I	I	I	I	I	I	I	I	I

Figure 5: Example of segmentation disagreement in a complex sentence with multiple coordinated and subordinated structures.

Token *English*	Die *The*	App *App*	unterstützt *supports*	natürliche *natural*	Spracheingabe *speech recognition*
Model	B	I	I	I	I
Annotator 1	B	I	I	I	I
Annotator 2	B	I	I	I	I

Token *English*	**und** ***and***	bietet *offers*	intelligente *intelligent*	Vorschläge *suggestions*	. .
Model	**B**	I	I	I	I
Annotator 1	**B**	I	I	I	I
Annotator 2	I	I	I	I	I

Figure 6: Example of segmentation disagreement on a coordinated clause. Annotator 1 and the model introduce a new segment boundary at “und”, while Annotator 2 treats the entire sentence as a single segment.