Steffen Frenzel, Maximilian Krupop, Manfred Stede

Discourse Segmentation of German Text with Pretrained Language Models

Abstract

Segmenting text into so-called "elementary discourse units" (EDUs) is a task that is relevant for several NLP applications, including discourse parsing or argument mining. In recent years, EDU segmentation has been addressed as part of a shared task on multilingual discourse parsing ("DISRPT"), where BERT-based encoder models proved particularly successful. The German language has been represented in DISRPT with the Potsdam Commentary Corpus (Stede, 2004), but recently, more German data with EDU segmentation has been published. In this paper, we conduct detailed tests on the German-language datasets that are currently available. We test a multilingual off-the-shelf model, several BERT-based encoders, and the current generation of LLMs. The results are analyzed both qualitatively and quantitatively and are compared to the multilingual state-of-the-art. We are making the best-performing model available as a tool that can be used by the community.

1. Introduction

Discourse segmentation as a computational task arose in the context of RST parsing (Rhetorical Structure Theory, Mann and Thompson (1988)) in the early 2000s. Analyzing a text according to RST requires the preparatory step of segmenting the text into Elementary Discourse Units (EDUs), which are then recursively connected to each other by so-called coherence relations. Hence, EDUs are the smallest units that make an independent contribution to the discourse structure; semantically they correspond to propositions or speech acts, and syntactically – broadly – to clauses. Beyond this general characterization, the definition can get a bit complicated, as we will discuss in Section 2.1.

Over the years, EDU segmentation has become relevant also outside the realm of discourse parsing. For example, in the dialogue community, speaker turns have been broken into units that can be labeled as a separate 'dialogue act', which bears resemblance to EDUs, though spoken language has a number of additional intricacies for a systematic segmentation. For text summarization, it was shown that extractive approaches can benefit from the presence of EDU boundaries (Li, Thadani, & Stent, 2016). In the field of argument mining, the notion of 'argumentative discourse unit' has been defined as an extension of the EDU (e.g., Seyfried, Reed, and Kamide (2024)). And recently, EDUs have been proposed as the base unit for aligning 'semi-parallel text', i.e., versions of a text that have been edited to some degree (Frenzel & Stede, 2025).

In recent years, several papers on EDU segmentation have been published, particularly in the context of the "DISRPT" shared tasks (Braud, Zeldes, Li, Liu, & Muller, 2025). These models are

generally multilingual, but there has been no work on EDU segmentation specifically on German texts for quite some time. The resource situation has improved significantly in recent years with the publication of the APA and PARADISE datasets (Hewett, 2023; Seemann, Shahmohammadi, Stede, & Scheffler, 2024; Shahmohammadi & Stede, 2024). Since the DISRPT shared tasks do not yet include these new resources, this paper aims to train and evaluate various models explicitly for German EDU segmentation.

(1) **[**Die spezifische Qualität seines Ruhmes änderte sich in keinem Lande und in keinem Jahrzehnt**:]** **[**immer wieder stand die Auflagenhöhe seiner Werke in keinem Verhältnis zu der immer noch anwachsenden Literatur über ihn**]** **[**oder zu dem immer noch sich vertiefenden und verbreiternden Einfluß**,]** **[**den dieses Werk auf die Schriftsteller der Zeit ausübt**.** (Arendt, 2018, p. 97)

An evaluation will also be carried out on Hannah Arendt's syntactically quite complex essay "Franz Kafka" (Arendt, 2018). Example 1 already shows that Arendt uses long sentences in her essays, which can usually be divided into several EDUs. She uses various subordinate clause constructions, putting the annotation guidelines and our classification models to the test. These texts are therefore particularly well-suited for evaluating our work as harder-than-average cases.

In this paper we make the following contributions:

- We test several current approaches to automatic discourse segmentation of German text, including GPT-5 and three fine-tuned XLM-RoBERTa models.
- We evaluate all models using the same method that was used in the DISRPT shared tasks to make our results directly comparable to the state-of-the-art.
- We make the best segmentation model publicly available.

In the following, we discuss the notion of EDU and some related work in Section 2, then we describe our annotation process in Section 3. We report on experiments with automatic segmentation in Section 4 and conclude in Section 5.

2. Background & Related Work

2.1. Elementary Discourse Units

EDUs may be characterized as forming the intersection between the grammar of a language and its use: Prevot and Muller (2025) describe them on the one hand as the largest possible unit of traditional grammar and on the other as the smallest possible unit of discourse analysis.

If that discourse analysis is done in terms of RST (Mann & Thompson, 1988), which defines some 20 coherence relations in terms of speaker intentions, then the aim of an annotation is to reconstruct the author's "text plan" from the perspective of the reader. Although EDUs represent the unit on which the RST trees are based, Mann and Thompson (1988) do not clearly define the rules according to which EDU segmentation should be done. Carlson, Marcu, and Okurowski (2003) list a number of possible definitions from the early research phase of text segmentation, which shows that the exact definition is not only a linguistic matter but also depends on the

technical perspective. Similarly, Polanyi, Culy, van den Berg, Thione, and Ahn (2004) state that a "discourse theory must specify how 'segments' should be identified in light of the questions the theory is set up to answer".

For example, EDUs were repeatedly defined as clauses (Givón, 1983; Longacre, 1996), however, Taboada and Mann (2006) state that this segmentation rule can be problematic if applied to multilingual data or spoken language. Depending on the research goal, EDUs were therefore also described as prosodic units (Hirschberg & Litman, 1993) or equated with sentences (Polanyi, 1988). Prevot and Muller (2025) assign these works to different related research fields and cluster them under the terms *Interactional Units*, *Conversational Discourse Units*, *Basic Discourse Units* and *Illocutionary Units*. However, all definitions share two fundamental properties: the segmentation must be exhaustive (i.e., no tokens remain) and EDUs must not overlap.

Our annotation guidelines (see Section 3.1) are closely related to the guideline document by Stede et al. (2015), which was explicitly created for RST analyzes of German text. There, EDUs are described as follows:

> "An EDU corresponds to a recognizable, independent speech act (illocution). However, this does not have to be structurally "complete" in the narrow sense: any elisions should be filled in during the assessment, and anaphoric (or cataphoric) references should be replaced by their antecedents." (Stede et al., 2015, p. 149)

"Filling in" and "replacing" are meant to be mental operations of the annotator who decides on segment boundaries. In the guidelines, main clauses and main clause fragments ("Passt mir!" / *Suits me!*) are always considered to be an independent EDU. As a shortcut for making segmentation simpler, parenthetical insertions that appear in the middle of an EDU are not considered to be an independent EDU. Subordinate clauses are categorized for their syntactic type and semantic function, and some of them are designated as EDUs while others (e.g., complement clauses) are not. Fragmentary material can be considered an EDU if it forms an independent illocution, but this decision remains subjective to a certain degree. For detailed rules that also address the role of certain prepositional phrases, see Stede et al. (2015).

2.2. Segmentation Models

Early automatic discourse segmenters usually exploited surface signs and used rule-based approaches for segment identification (e.g. Le, Abeysinghe, and Huyck (2004); Tofiloski, Brooke, and Taboada (2009)). Sidarenka, Peldszus, and Stede (2015), who implemented a syntax-oriented model for German texts, pointed out that these approaches would be "ideal" from a computational perspective, because the form-based unit identification could be seen as a step that can be clearly separated from the interpretation of these units. At the same time, rule-based approaches run into robustness problems and often have trouble with complex sentences and with fragmentary material that appears in many kinds of text. Nonetheless, rule-based approaches have been developed for various languages including Spanish (Da Cunha, SanJuan, Torres Moreno, Lloberes Salvatella, & Castellón Masalles, 2010) and Dutch (Van der Vliet, 2010). In these approaches, the task is either casted as binary classification on word level (e.g., Fisher and Roark (2007); Joty, Carenini, and Ng (2015); Subba and Eugenio (2007)) or as a sequence labeling problem (e.g., Braud, Lacroix,

and Søgaard (2017a, 2017b); Hernault, Bollegala, and Ishizuka (2010); Xuan Bach, Le Minh, and Shimazu (2012)).

Recent approaches shifted emphasis from form to meaning. In 2018, Wang, Li, and Yang (2018) presented `NeuralEDUSeg`, a model that uses a BiLSTM-CRF approach. This work was one of the first to exclusively use a neural network for segmentation, though it faced the problem that training data was sparse and neural networks like the BiLSTM required large amounts of data. Also, EDU-boundaries are sometimes not determined locally, when clauses are deeply embedded or interrupted by parentheticals. Long-distant dependencies could be extracted from parse trees (which were created by rule-based dependency parsers) to some extent, but are difficult to identify by neural models. The authors tried to solve these problems by employing a restricted self-attention mechanism (Vaswani et al., 2023) and using pre-trained word embeddings (Peters et al., 2018).

In the last years, the workshop on Discourse Relation Parsing and Treebanking (DISRPT) dedicated several shared tasks to EDU segmentation. In 2019, the winning approach was `ToNy` by Muller, Braud, and Morey (2019). They tested a BiLSTM-CRF against a BERT-based sequence prediction model and used a total of 15 corpora for their experiments. The data included different languages as well as all three main frameworks for discourse - RST, SDRT and PDTB. By using multilingual BERT embeddings for sequence prediction, the authors were able to outperform existing models for almost all languages.

The winning system in DISRPT 2021 by Gessler et al. (2021) extended `ToNy` by introducing handcrafted features like POS-tags, dependency relations, head distance etc. The features are computed token-wise and are then added to the word embeddings.

The 2023 winning approach by Braud et al. (2023) used a segmentation model based on XLM-RoBERTa. They assume that lower layers of a neural network encode mainly morpho-syntactic information, while higher layers mainly encode semantic information (see Bender and Koller (2020); Kovaleva, Romanov, Rogers, and Rumshisky (2019); Rogers, Kovaleva, and Rumshisky (2021)). Therefore, they experiment with freezing certain layers to improve performance and make the models more efficient.

In 2025, the best results for the German PCC data were achieved by `SeCoRel`, another approach based on XLM-RoBERTa (Lalitha Devi, Rk Rao, & Sundar Ram, 2025). It was submitted by the team from AU-KBC Research Center. They achieved an f1 score of 0.944 on the plain tokenized PCC data, which was slightly worse than the winners from 2023 and 2021.

We compare all DISRPT winning approaches with our best models for German in Section 4.

A new approach was taken by Nayak (2024), who tested ChatGPT as a zero-shot segmentation tool in combination with various prompting strategies. The main outcome, however, was that LLMs are still not well suited for this task, since the models often 'hallucinated' and the output could not compete with much smaller models that were explicitly trained for this task.

2.3. German EDU-Segmented Datasets

APA-RST: The APA-RST dataset consists of German newspaper articles that are available in several versions with varying degrees of simplification. The articles originate from the Austrian Press Agency and were written between 2018 and 2022. The original articles were manually

simplified to the levels B1 and A2 according to the Common European Framework of Reference for Languages (CEFR). APA-RST consists of 25 articles per complexity level, which add up to 75 documents in total. The dataset was annotated according to the RST guidelines of Stede, Taboada, and Das (2017) and published by Hewett (2023). The EDU segments can be derived from the RST trees.

PCC: The Potsdam Commentary Corpus (PCC) was initially published by Stede (2004) and has since been modified and extended. The latest version was published by Bourgonje and Stede (2020). It contains a total of 176 newspaper articles from the German *Märkische Allgemeine Zeitung*. The dataset was annotated in terms of sentence syntax, coreference, aboutness topics and discourse structure according to both RST and PDTB (Prasad et al., 2008). EDUs can again be derived from the RST trees.

	APA	PCC	PARA	Kafka
Documents (total)	75	176	78	1
Sentences (total)	902	1896	740	76
EDUs (total)	1358	3112	1387	219
Avg. tokens per sentence	14.91	15.81	20.18	33.58
Avg. tokens per EDU	9.91	9.27	10.81	11.71
Avg. EDUs per sentence	1.51	1.64	1.87	2.87
Avg. EDUs per doc	18.11	17.68	17.78	219

Table 1: Corpus statistics for all datasets.

PARADISE: The PARADISE (PARAllel DIScoursE) corpus was published by Seemann et al. (2024) and was created to analyze differences in spoken vs. written discourse in the context of German computer-mediated communication. The corpus consists of 69 podcasts and 69 corresponding blogposts from the 'business' and 'science & culture' domains. Manual EDU segmentation and the annotation of discourse relations according to RST were based on the guidelines of Stede et al. (2017). We use the part of the dataset that is annotated for discourse structure, including a total of 78 documents.

Kafka: In addition to these published datasets, we also use an essay by philosopher Hannah Arendt entitled "Franz Kafka." The essay describes the work and reception of the German author and is available in several versions. It was first published in English in the US, but we use a German version from 1948.[1] The EDU segmentation for this text was created in the course of our manual annotation study, which is described in more detail in Section 3. In comparison to

[1]The essay is published as part of the research project 'Hannah Arendt: Kritische Gesamtausgabe'. (`https://hannah-arendt-edition.net/home`)

the newspaper and social media texts mentioned above, the essay is characterized by long and complex sentences and therefore represents a particularly difficult case for EDU segmentation. We use this text both to check the quality of our annotation guidelines in the course of the manual segmentation and to test the models for automatic segmentation. The text is not used to train the models.

3. Manual Annotation

In order to create a gold segmentation of the Kafka essay we conducted a manual annotation study. The annotation guidelines, which we derived from the related work, are presented in Section 3.1, the human agreement is described in Section 3.2 and the disagreements are analyzed in more detail in Section 3.3.

3.1. Annotation Guidelines

The EDU segmentation guidelines for our experiments are a slightly modified version of the guidelines by Stede et al. (2015), which can – in a short form – also be found in Sidarenka et al. (2015). Their aim was a flexible segmentation that provides the basis for a variety of subsequent analyzes (e.g., rhetorical structure, argumentation structure, illocutions). The authors therefore targeted a segmentation that was as fine-grained as possible and could afterwards be flexibly adapted for the respective analysis task.

For our purposes, we used an abridged version of those guidelines. The overarching formal principle is that EDU segmentation must be exhaustive and non-overlapping; each token must therefore be assigned to exactly one EDU.

Annotators should start by searching for sentence-final punctuation symbols (SFPS). Full stops, exclamation marks and question marks as well as colons and semi-colons usually constitute EDU boundaries. Therefore, the segment between two SFPS has to be inspected: When ignoring any connectives and (mentally) replacing anaphoric material with their antecedents, the segment should contain a complete independent proposition in order to qualify as an EDU.

Sometimes a segment between two SFPS only contains one illocution – in that case, no further action is needed. However, some sentences are more complex and can be divided further. Three cases indicate the need to further segment a sentence:

Parataxis: Multiple main clauses can appear in one sentence, if they are linked by a coordinating conjunction. Each one forms an independent EDU:

(2) **[**Meier öffnete die Flasche,**]** **[**und er trank sie in einem Zug leer.**]**
(Stede et al., 2015, p. 33)

Hypotaxis: Some types of subordinate clauses can form independent EDUs as well:

- Non-restrictive relative clauses can provide an independent information about the head-nominal phrase:

  (3) **[**Wohin mit den vom Urlaub übrig gebnliebenen Münzen,**]** **[**die bald ohnehin nichts mehr wert sind?**]** (Stede et al., 2015, p. 37)

- Continuative subordinate clauses can provide independent information by referring to the complete illocution of the previous clause - instead of referring to a specific element of that clause:

  (4) **[**Er hat uns eingeladen,**]** **[**was uns sehr freut.**]** (Stede et al., 2015, p. 38)

- Adverbial clauses are related in terms of content to the superordinate main clause and therefore provide an independent information – the have to be identified as an individual EDU as well:

  (5) **[**Er ging schwimmen,**]** **[**obwohl er eine leichte Grippe hatte.**]** (Stede et al., 2015, p. 36)

Parentheticals: Parentheticals are inserted into a clause. These units "interrupt" the clause and are marked by commas, hyphens, or parentheses. Example 6 shows such a case. However, mark only those that correspond to a complete proposition or a clearly identifiable illocution. Do not mark them as an independent EDU if fragments are created to the left or right that cannot be assigned to another EDU.

A more detailed version of our annotation guidelines can be found in the LLM system prompt in Appendix A.5.

3.2. Inter-Annotator Agreement

To assess the annotation guidelines, we conducted an agreement study with two annotators. Documents were selected from two different datasets – on the one hand a subset of the relatively short APA documents, and on the other hand the Kafka essay, which we had already mentioned earlier as a difficult case. We use the same metrics that we will also use later on to evaluate our segmentation models: Precision, Recall and f1 like they were used in the DISRPT evaluation (Braud et al., 2024). In addition, the WindowDiff metric (Pevzner & Hearst, 2002) (WinDiff) is applied, which considers sequences of EDU boundaries and thus captures deviations in segmentation structure. The window size is adaptively determined based on the average EDU length, ensuring a realistic sensitivity of the metric. Unlike f1, WinDiff is an error metric on a scale between 0 and 1, small values are therefore desirable. All evaluation metrics are described in more detail in Section 4.3.

As neither of the two annotations is a gold standard, we modify the application of the evaluation metrics, computing it in both directions and then reporting the average of both scores.

	Prec	Recall	f1	WinDiff
APA	0.980	0.934	0.957	0.1395
Kafka	0.830	0.784	0.807	0.357

Table 2: Agreement of the manual annotations.

The scores shown in Table 2 imply that our difficulty estimate of the different datasets was correct. While the agreement on the APA data is satisfactory, there are significant differences between the annotations of the Kafka essay. We do not have comparable results on these datasets

from the related work, but we can report the results of an agreement study on the PCC data as a reference. The study was conducted by Sidarenka et al. (2015) and they report a mean WinDiff score of 0.177 on the PCC data. This indicates that the PCC is placed in between the APA and Kafka data in terms of annotation difficulty. Incidentally, this also corresponds to the results of our experiments with automatic segmentation, which are reported in Section 4.4.

The next section describes some errors in more detail. Since there was no existing gold standard for the Kafka essay, the errors were discussed based on the annotation guidelines and a gold standard was then distilled from the two annotations.

3.3. Analysis of Disagreements

In order to identify recurring sources of error in the manual annotation, a qualitative analysis of the two annotations of the Kafka essay was performed. We checked all segment boundaries that were not identified by both annotators.

First, parentheticals stand out:

(6) Annotation 1:

[Das einzige, was den Leser in Kafkas Werk lockt und verlockt, ist die Wahrheit selbst,**]** **[**und diese Verlockung ist Kafka in seiner stillosen Vollkommenheit**]** **[**– jeder »Stil« würde durch seinen eigenen Zauber von der Wahrheit ablenken –**]** **[**bis zu dem unglaublichen Grade geglückt, daß seine Geschichten auch dann in Bann schlagen,**]** **[**wenn der Leser ihren eigentlichen Wahrheitsgehalt erst einmal nicht begreift.**]** (Arendt, 2018, p. 97)

(7) Annotation 2:

[Das einzige, was den Leser in Kafkas Werk lockt und verlockt, ist die Wahrheit selbst,**]** **[**und diese Verlockung ist Kafka in seiner stillosen Vollkommenheit – jeder »Stil« würde durch seinen eigenen Zauber von der Wahrheit ablenken – bis zu dem unglaublichen Grade geglückt, daß seine Geschichten auch dann in Bann schlagen,**]** **[**wenn der Leser ihren eigentlichen Wahrheitsgehalt erst einmal nicht begreift.**]** (Arendt, 2018, p. 97)

These cases are problematic, because, according to our definition, the parentheticals themselves often constitute independent EDUs. However, since they can be inserted in the middle of a sentence, fragments sometimes arise to the left and right of them that cannot be considered independent EDUs. Since such fragments must not be left behind, no segment boundary is to be introduced in these cases. This is illustrated by Example 6 – segments 2 and 4 in this annotation do not constitute complete EDUs, so no segment boundaries should be introduced at the dashes. Instead, Example 7 shows the correct annotation in this case.

A similar interesting case are sentences like the one in Example 9. The long subordinate clauses, which were evaluated by annotator 2 as an independent EDU, are inserted into the sentence in such a way that the main clause is interrupted and fragments are created. Such problems could in principle be solved by a hierarchical segmentation (cf. Sidarenka et al. (2015)) that allows for EDU embedding. While being more appropriate from the linguistic viewpoint, embeddings are

known to be problematic for automatic approaches, and hence we here restrict the process to a "flat" annotation that as a result misses some complete propositions that are embedded in others.

(8) Annotation 1:

[Der Prozeß, über den eine kleine Bibliothek von Auslegungen in den zwei Jahrzehnten, die seit seinem Erscheinen verstrichen sind, veröffentlicht worden ist, ist die Geschichte des Mannes K.,...**]** (Arendt, 2018, p. 97)

(9) Annotation 2:

[Der Prozeß,**] [**über den eine kleine Bibliothek von Auslegungen in den zwei Jahrzehnten, die seit seinem Erscheinen verstrichen sind, veröffentlicht worden ist,**] [**ist die Geschichte des Mannes K.,...**]** (Arendt, 2018, p. 97)

Apart from that, there were various individual errors; for example, segment boundaries were overlooked when stringing together several main clauses. However, these cases are rare and play only a minor role.

4. Automatic Segmentation

4.1. Training Data, Pre- and Postprocessing

We use the three corpora described in Section 2.3 for model training: PCC (176 documents), APA-RST (75), and PARADISE (78). A corpus overview and global statistics are given in Table 1. We adopt an 80/10/10% document-level split (train/dev/test) that is stratified at the article level. The resulting counts can be found in Table 3.

Gold EDUs are extracted from the RST tree files by reading the `<segment>` elements. We reconstruct each document's segments into a continuous text as input to our models and remove headlines when present as they are not connected to the RST tree. For APA-RST, the original texts were occasionally split across multiple RST tree files due to their length. For these cases, we concatenate the different parts to recreate the original document. For PCC and PARADISE, we automatically normalize extraneous whitespace around punctuation. All normalization and headline removal occur before segmentation and do not alter EDU boundaries.

Corpus	Train	Dev	Test	Total
PCC	141	17	18	176
APA	61	7	7	75
PARA	62	8	8	78
All	264	32	33	329

Table 3: Document counts per split for each corpus

Evaluation-time post-processing (DPLP only): To undo formatting artifacts introduced by the DPLP parser (to be introduced in the next section), we apply boundary-safe whitespace normalization to DPLP outputs. Specifically, we use a rule-based pass to (i) remove spaces before punctuation, (ii) normalize spacing around brackets and quotation marks, (iii) preserve hyphenation at line ends, and (iv) standardize frequent abbreviations (e.g., *e.V.*) and rare forms (*Trainer*in*). We do a further manual correction for cases when quotation marks span multiple EDUs or when hyphens appear as prefixes (*In- und Ausland*), in compounds (*Asien-Pazifik-Region*), or as dashes. All edits are restricted to line level and change only spacing characters: we never insert or delete newline characters and do not rejoin hyphenated items split across lines. Thus, EDU boundaries produced by DPLP remain unchanged.

4.2. Model Designs

Detailed descriptions of all model settings and the prompts used for the LLM approaches can be found in Appendix A.1 to Appendix A.5.

4.2.1. DPLP Discourse Parser

The original DPLP is a feature-based, shift-reduce RST parser (Ji & Eisenstein, 2014). It builds RST trees over EDUs using surface and syntactic cues (indices, token/POS, dependency label/head) and Brown-cluster features for lexical generalization. We include DPLP as a non-neural reference model.

For German, Shahmohammadi and Stede (2024) modernized the codebase and containerized runtime dependencies. In this work we use only the segmentation component (`discoseg`) from their Github fork. The segmenter takes raw text and returns one EDU per line. We run it from the fork's Docker image and retrain it on our split. Preprocessing and evaluation-time normalization follow Section 4.1, where the latter edits whitespace only and does not change EDU line breaks. See Appendix A.1 for DPLP settings.

4.2.2. DMRST Discourse Parser

We retrain the DMRST parser (Liu, Shi, & Chen, 2021) using the publicly available adaptation by Chistova (2024). We include this model as it is built on an XLM-RoBERTa (Conneau et al., 2019) backbone which we also use for our own models. DMRST is an end-to-end RST framework that integrates token-level segmentation with a hierarchical encoder and a top-down span-splitting decoder (plus nuclearity/relation labeling) for RST trees. In our experiments, we use only DMRST's segmentation component and do not perform full tree or relation parsing. Preprocessing (including headline removal and pre-segmentation normalization) matches Section 4.1. No evaluation-time post-processing is applied to DMRST outputs.

While Chistova (2024) modifies the original framework by replacing the token classifier with a BiLSTM-CRF segmenter (`ToNy`) and adding a BiLSTM local encoder, we keep the original DMRST token classifier and select the configuration that performed best on our token-boundary metric: `XLM-RoBERTa-base` with a linear token-level segmenter and sentence-boundary hints. At inference time, the model takes raw text and outputs one EDU per line. Training follows the

same document-level split as in Section 4.1. Hyperparameters are listed in Appendix A.2 and were tuned for segmentation quality on our dev set rather than end-to-end RST parsing.

4.2.3. RoBERTa

We also fine-tune two XLM-RoBERTa models ourselves, following recent approaches to discourse segmentation (e.g., Metheniti, Braud, Muller, and Rivière (2023)). In the first approach, we use the multilingual XLM-RoBERTa (Conneau et al., 2019), directly for token-level classification (`XLM-RoBERTa-base`). In the second approach, we additionally incorporate a Conditional Random Field (CRF) layer above the classification head (`XLM-RoBERTa-crf`). The CRF explicitly models label dependencies across tokens, encouraging consistent labeling sequences (e.g., disallowing an 'I' without a preceding 'B').

To improve fine-tuning stability, we adopt gradual unfreezing (Howard & Ruder, 2018). Initially, all transformer layers of the XLM-RoBERTa encoder are frozen. During training, we progressively unfreeze the higher layers in scheduled steps, allowing the model to gradually adapt pretrained representations to the EDU segmentation task. For the base model, training minimizes token-level cross-entropy loss. In the CRF variant, loss is computed as the negative log-likelihood under the CRF. Optimization is performed with AdamW (Loshchilov & Hutter, 2017) and standard learning rate schedules.

4.2.4. LLM-based Segmentation

In another approach, we segment the test set of our data split (see Section 4.1) into EDUs by prompting two different OpenAI models via the Responses API. A single, fixed prompt is used for all documents: it combines a short task description, a minimal rule set distilled from our annotation guidelines (Section 3.1), and a one-shot example from the PCC corpus paired with its gold segmentation as the expected model output. We do not enable tools or retrieval and request plain text, line-separated EDUs as output. The exact prompt and a minimal inference call are provided in Appendix A.5 and Appendix A.6 (Listings 1 and 2).

We evaluate two model families: GPT-4.1[2] as a non-reasoning model and GPT-5[3] as a reasoning model. For GPT-4.1 we set `temperature=0` and `top_p=1`. For GPT-5 we leave decoding and any reasoning controls at defaults. Each document is processed once (single pass, no multi-seed). We reuse the same one-shot PCC example for all domains, which simplifies setup but may introduce domain-mismatch effects; exploring domain-matched exemplars or prompt variants is left to future work.

Outputs are post-processed only for formatting: we strip leading/trailing whitespaces per line and drop blank lines. We do not merge or split lines or edit punctuation. Thus, EDU boundaries remain exactly as produced by the model.

The corpora we use are publicly available. We cannot rule out that some source texts (or derivatives) appear in web-scale pretraining data for proprietary models. To ensure comparability with our trained models, we evaluate the LLMs on the same held-out test split.

[2] `gpt-4.1-2025-04-14`
[3] `gpt-5-2025-08-07`

4.3. Evaluation Metrics

Evaluating the performance of models for discourse segmentation is more difficult than for most other tasks. In the literature, different metrics are used, and in many cases it is not clear exactly how the values were arrived at.

Most approaches treat segmentation as a binary classification problem at the token level: For each token, it is being decided whether it initiates a new segment or not. This is done, for example, in the DISRPT shared tasks. However, when we use LLMs for segmentation, hallucinations or shifts in the model's output occur occasionally. In addition, the models use different (often integrated) tokenizers that cannot be easily manipulated. If segment boundaries are then determined using the index positions in the tokenized model output, these discrepancies can have a major impact on the evaluation metrics.

We eventually decided to use the official DISRPT evaluation script to make our results comparable to the state-of-the-art models (Braud et al., 2024). However, since we faced the previously mentioned tokenization issues with the GPT models, we decided to rebuild plain text from all model outputs and then use the same tokenizer (in our case `spacy`) to create a unified tokenization. We also included a preprocessing step to find out if words were added, deleted or changed by the models. Then we created `.tok` files that are compatible with the DISRPT evaluation and used their script to evaluate all models. The DISRPT evaluation uses micro-averaged Precision, Recall and f1.

Additionally, we report the previously-mentioned WinDiff as a segmentation-specific error metric. While Precision, Recall, and f1 treat segmentation as a boundary classification task – thereby answering the question of whether the set boundaries are in exactly the right positions – WinDiff measures the structural consistency of the segmentations by sliding a fixed-size window across both segmentations and comparing the number of segments within each window. Over- and undersegmentation are thus penalized without evaluating the exact positions. Instead, this metric attempts to answer the question of whether a model recognizes the approximate correct size and distribution of EDUs in a text (Pevzner & Hearst, 2002).

For each document, the evaluation metrics are calculated separately and then averaged across the entire test dataset.

4.4. Results

A total of six different models were evaluated on four different datasets.[4] In all cases, the same split was used for training and evaluation. In all cases, the evaluation was performed by using the official DISRPT evaluation script (Braud et al., 2024). Additionally, we calculated the WinDiff as described in Section 4.3.

Looking first at the overall performance of the various models (shown in Table 4), it is striking that no single model is clearly superior in all cases. Instead, different segmenters show strengths and weaknesses in connection with the different data sets. Nevertheless, some basic conclusions can be drawn:

[4] Our best-performing model is available on Github: `https://github.com/discourse-lab/eduseg_de`

	APA		PCC		PARA		Kafka	
	f1	WinDiff	f1	WinDiff	f1	WinDiff	f1	WinDiff
DPLP	0.919	0.066	0.904	0.176	0.794	0.318	0.692	0.451
DMRST	0.903	0.065	0.866	0.167	0.796	0.259	0.763	0.400
GPT-4.1	0.917	0.075	0.911	0.164	0.863	0.268	0.740	0.358
GPT-5	0.932	**0.064**	0.914	0.167	0.904	0.155	0.758	0.399
XLM-RoBERTa-base	0.943	0.074	0.938	0.108	**0.947**	**0.094**	**0.776**	**0.324**
XLM-RoBERTa-crf	**0.991**	0.072	**0.949**	**0.095**	0.901	0.165	0.752	0.415

Table 4: Micro-averaged f1 scores (DISRPT) and WinDiff for all models and datasets

The first thing that stands out is that the DPLP and DMRST parsers deliver the worst results overall, DPLP delivering the worst performance on PARADISE and Kafka and DRMST performing worst on APA and PCC. Since the DPLP parser is built on a rule-based architecture and is by far the oldest model in this comparison, its performance may not be surprising, but the DMRST parser should have worked better overall, given that it is also based on an XLM-RoBERTa model.

Secondly, it can be seen that in a direct comparison of the LLMs, the newer GPT-5 model outperforms the older GPT-4.1 model in all cases, even though the differences between the models are fairly small. Both models were instructed with the same prompt, so the results can be compared directly.

Thirdly, it can be observed that the two fine-tuned XLM-RoBERTa models produce the best results overall. On the Kafka data, which is particularly difficult to segment in comparison, XLM-RoBERTa-base has an edge over all competitors. However, the ranking differs depending on the evaluation metric: WinDiff ranks GPT-5 as the best model on the APA dataset, while the DISRPT-based f1 score favors the XLM-RoBERTa models. As mentioned earlier, the two evaluation metrics measure different things, therefore differences in the rankings are possible. Nevertheless, a systematic analysis of these evaluation techniques for discourse segmentation is noted as a task for future work.

Apart from that, we can also compare the performance of our models with the DISRPT models of the past years. The PCC was included in their shared tasks since the first iteration in 2019, however, we do not have information about their exact train-test split.

We chose the best performing approaches from the last three DISRPT iterations for our comparison. As can be seen in Table 5, the results for all models are fairly similar. Even the 2021 approaches are still competitive, especially the 2021 `disCut` version (Kamaladdini Ezzabady, Muller, & Braud, 2021). Our XLM-RoBERTa models can almost match this performance on

	This work		DISRPT 2025	DISRPT 2023	DISRPT 2021
	`rob-base`	`rob-crf`	`SeCoRel`	`DisCut`	`disCut`
Precision	0.959	0.957	0.949	**0.968**	0.947
Recall	0.918	0.942	0.939	0.918	**0.966**
f1	0.938	0.949	0.944	0.942	**0.956**

Table 5: Comparison of our XLM-RoBERTa models with `SeCoRel` (Braud et al., 2025), `DisCut` (Braud et al., 2023) and `disCut` (Zeldes et al., 2021). All models were tested on plain tokenized PCC data.

the PCC data, but the other models that we tested are not fully competitive in this comparison. However, it should be mentioned that the DISRPT models have the important advantage of being trained on a much larger amount of annotated data – the 2025 edition included a total of 39 annotated datasets across 16 languages (Braud et al., 2025).

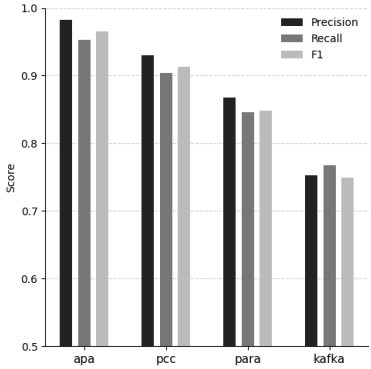

Figure 1: Averaged performance of all models per dataset

We also calculated the average f1 score of all models per dataset. In the results shown in Figure 1, the varying difficulty of the test data becomes clearly apparent. While the average f1 score of all models on the APA data is above 0.9, it falls below 0.8 on the Kafka data. Frenzel and Stede (2025) calculated that the APA texts contain an average of only 1.12 EDUs per sentence, while Kafka contains an average of 2.57 EDUs per sentence, showing again that these sentences are far more complex. This is also reflected in the average sentence length, which is approximately 10 words per sentence for APA, but approximately 30 words per sentence for Kafka. However, from

f1	RoBERTa-CRF	RoBERTa-base	GPT5	GPT4	DMRST	DPLP
DPLP	0.914	0.900	0.839	0.819	0.820	1.000
DMRST	0.880	0.886	0.831	0.803	1.000	
GPT4	0.912	0.908	0.855	1.000		
GPT5	0.907	0.899	1.000			
RoBERTa-base	0.953	1.000				
RoBERTa-crf	1.000					

Table 6: Agreement (f1) amongst models on test data

these numbers we can also calculate that regardless of the average sentence length, the average EDU length stays roughly the same for all datasets.

Next, we look at the amount of agreement between the predictions of the various models; these are shown in Table 6. Again, the micro f1 score from DISRPT was used.

The two RoBERTa models show the highest agreement of all models. Since these models are – apart from the CRF layer – very similarly designed, these results are not surprising. Interestingly, the XLM-RoBERTa-CRF model shows quite high agreement with most models. The DMRST parser has, on average, the lowest overlap with the other models. It should be noted, however, that all agreement scores are distributed within a range of 0.15 f1 and are thus relatively close to each other.

Finally, we can also measure how many EDUs are predicted by each model in comparison to the gold standard. We compute these scores as a total count across all datasets. The results are shown in Figure 2.

The graph shows the deviations in the absolute count of EDU segments in the model predictions compared to the gold standard, which contains a total of 756 EDUs. Interpreting these figure is somewhat tricky: Although the smallest possible deviation from the gold standard is desirable here, it does not allow any conclusions to be drawn as to whether the predicted segment boundaries are also in the correct positions. This is evident from the fact that the ranking of the models here differs from that in the evaluation using the f1 score.

The RST parsers DPLP and DMRST show rather small deviations from the gold standard in the number of segment boundaries. However, since these parsers in particular produced poor overall

Figure 2: Total count of EDU segments on the test data: deviation from gold standard

f1 scores, it can be concluded that many of the segment boundaries set are in the wrong positions. Both LLM approaches are way off in this comparison. GPT-4.1 sets far too few boundaries overall and has the largest deviation from the gold standard. GPT-5 sets too many boundaries and is a close second to last in this comparison. The XLM-RoBERTa-base model also predicts too few boundaries. The XLM-RoBERTa-crf model shows the best overall performance in this comparison, predicting only two boundaries more than can be found in the gold standard.

4.5. Error Analysis

In this section, we take a closer look at the outputs of the models that achieve the best results on average: the RoBERTa models as well as GPT-5. We manually analyze the model outputs on the easy APA dataset and the difficult Kafka dataset and try to identify general error trends in the models.

XLM-RoBERTa: There are very few deviations from the gold standard in the APA dataset. This is mainly because very few sentences can be divided into multiple EDUs. Nevertheless, it speaks in favor of the models that no incorrect EDU boundaries are drawn in this case. A striking example in which an EDU boundary was overlooked by both RoBERTa models is the following sentence:

(10) „Der Kandidatenkreis war am Anfang sehr groß, wurde aber von Tag zu Tag kleiner."

These are two main clauses connected by the conjunction "aber". However, this is a stylized construction, as the subject in the second main clause has been omitted and the predicate has been

moved to the front. The sentence would be clearer (for segmentation purposes) if the subject was mentioned again and the conjunction was placed in front:

(11) „Der Kandidatenkreis war am Anfang sehr groß, aber er wurde von Tag zu Tag kleiner."

Verbs like 'wurde' are not usually expected to mark a segmentation boundary, but conjunctions like 'aber' appear frequently as boundary markers. Therefore, it stands to reason that the models made an incorrect decision here. Apart from these special constructions, however, hardly any errors have appeared in the APA dataset.

In comparison, the Kafka essay shows a different picture. Errors are more frequent in this case and occur in various constructions. In addition, the differences between the two RoBERTa models are greater here. While the RoBERTa-base model predicts the same overall number of EDU boundaries as contained in the gold standard, the evaluation metrics already show that some boundaries must nevertheless be set incorrectly. Just as in the manual annotation, the parentheticals that are typical of Arendt's writing stand out in this case again. The models often annotate inconsistently here, as can be seen in example 12:

(12) **[**Daß solche Irrtümer möglich waren**] [**– und dies Mißverständnis ist nicht weniger fundamental,**] [**wenn auch weniger vulgär, als das Mißverständnis der psychoanalytischen Auslegungen Kafkas –, liegt natürlich im Werke Kafkas selber.**]** (Arendt, 2018, p. 97)

The beginning of the parenthetical is recognized here as an EDU boundary, but the end is not. In this case, it would be correct to define the parenthetical as a separate EDU, since there are no fragments remaining on the left or right. However, since this is not always the case (as seen in example 6), insertions cannot be defined as separate EDUs across the board. Defining only the beginning or end of the parenthetical as a boundary is always incorrect, since an insertion itself is not connected to the surrounding parts of the sentence and, in case of doubt, is only annotated together with its context because it has "destroyed" its environment.

GPT-5: Prompting GPT-5 for segmenting the APA test split gives results very close to gold. While the simplified texts (A2 and B1 difficulty) are almost segmented perfectly, the model tends to insert superfluous splits which do not occur in the gold data, as in the examples 13 and 14:

(13) **[**Das heißt,**] [**man muss nicht mehr genesen oder geimpft sein.**]**

(14) **[**Heizöl kostete 61 Prozent mehr**] [**und Diesel fast 35 Prozent.**]**

We suspect these extra boundaries reflect the model adhering to the literal instructions in the system prompt (*"subordinate or coordinated clauses usually constitute their own EDU"*), whereas the gold annotation follows a more conservative convention that keeps complements inside their matrix clause. For the original text (OR) though, the model violates the instructions in certain cases by under-segmenting and merging multiple clauses into a single EDU. This occurs especially over clause boundaries in quotations and parentheticals as well as leaving supporting frames unsplit, such as example 15, which should have been split after the comma:

(15) Dass man sich für eine billige Lösung entschieden habe, verneinte Barisic vehement.

Both types of segmentation errors can be seen on the model output for the Kafka essay. Here GPT-5 over-segments, where the additional boundaries arise from clause-by-clause splitting after almost every comma. Again, we suspect this could be the model following the instructions too closely and thereby consistently segmenting subordinate clauses, as in example 16, which should not have been split into two EDUs:

(16) **[**Es ist durchaus charakteristisch für die Wirkung der Kafkaschen Prosa,**]** **[**daß die verschiedensten »Schulen« ihn für sich in Anspruch zu nehmen suchen;**]** (Arendt, 2018, p. 97)

At the same time, the model seems to ignore the system prompt rules for dash insertions and coordinated clauses, for instance by collapsing two EDUs into a single one in example 17:

(17) **[**Der Prozeß,**]** **[**über den eine kleine Bibliothek von Auslegungen in den zwei Jahrzehnten, die seit seinem Erscheinen verstrichen sind, veröffentlicht worden ist,**]** (Arendt, 2018, p. 97)

Missed boundaries by the model can be mainly traced back to long quotations, relative clauses, and dash-separated clauses.

5. Conclusion and Outlook

The results demonstrated that transformer-based encoders are currently the best solution for (German) EDU segmentation. While the GPT-5 results are also not bad, they introduce the problem of occasionally altering the original text during segmentation. This is a severe issue because it disrupts evaluation and because preserving the original text structure is essential for many subsequent tasks. Furthermore, as we pointed out, even the latest GPT version, despite extensive prompts including a one-shot example, still does not quite achieve the segmentation quality of the XLM-RoBERTa models that were explicitly fine-tuned for this task. One route for future work can be to test an open LLM; its vanilla performance may not be likely to beat that of GPT-5, but potentially an extra fine-tuning step could close the gap.

While segmentation now works very well on simple datasets such as the existing RST datasets that we described, problems still arise with complex texts such as Hannah Arendt's essays. For now, due to the small amount of data currently available, we have only used the Kafka essay in our experiments for testing purposes, not for training. But we are currently planning to annotate a larger amount of such political essays, which can be fruitful also for further "sharpening" the annotation guidelines when they are confronted with difficult syntactic and semantic phenomena and their combinations; currently (as we reported) not only the automatic performance but also the human agreement on these texts has room for improvement.

However, we also acknowledge that EDU segmentation itself is not fully objective. Even with detailed guidelines, annotators may reasonably disagree about plausible boundary placements, especially in complex German syntax and argumentative prose. More generally, this perspective

aligns with prior work on human label variation, arguing that disagreement can often be considered signal rather than noise (e.g., Plank (2022)). This means that the available gold-standard segmentation should be treated as a single reference annotation instead of an absolute ground truth. Accordingly, part of what we count as model errors may correspond to alternative yet valid segmentations. Future work is therefore left to collecting multiple annotations per text, analyzing disagreement patterns, and exploring disagreement-aware training and evaluation setups that allow for multiple valid boundary placements.

Last but not least, another item for future work concerns the evaluation methods for the segmentation task. Although in preparatory work we ran extensive experiments with different evaluation regimes at the character and token levels, we ultimately decided to report results with the approach that has been used at the recent DISRPT workshops, in the interest of comparability. However, our complete set of experiments showed that evaluation results can differ considerably between metrics. Therefore it can be beneficial to conduct a detailed study of the calculations of different evaluation metric variants for EDU segmentation and to analyze them in detail for their relative merits, which can lead to a better-motivated choice of method for the task.

Acknowledgements

We thank our student assistant Dietmar Benndorf for annotating training data, and we are grateful to the anonymous reviewers for their helpful feedback. Our work is supported by the Deutsche Forschungsgemeinschaft (DFG), project (524057241) "Semi-automatische Kollationierung verschiedensprachiger Fassungen eines Textes“.

A. Reproducibility details

A.1. DPLP settings

Item	Value
Runtime	Docker image `mohamadisara20/dplp-env:ger`
Code base	`https://github.com/mohamadi-sara20/DPLP-German`
Task used	EDU segmentation only (full parser trained but not used)
Training policy	repository defaults (no custom hyperparameters)
Split	as in Section 4.1 (no randomized seeding)
Preprocessing	as in Section 4.1 (headline removal, normalization)
Post-processing	as in Section 4.1 (boundary-safe spacing only, no line edits)
Relation mapping	unified PCC, APA-RST, PARADISE label set

Notes. Training used the fork's defaults; both parser and segmenter are LinearSVC models. Relation mapping was supplied for parser compatibility; segmentation results do not depend on it.

A.2. DMRST settings & hyperparameters

Item	Value
Code base	`https://github.com/tchewik/bilingualrsp`
Backbone	`xlm-roberta-base`
Segmentation head	linear boundary classifier (break vs. no-break)
Tagging scheme	binary (boundary / not)
Sentence-boundary hints	on (`use_sent_boundaries=true`, de)
Context window / padding	400 / 55 tokens
Freeze first transformer layers	10
Batch size / epochs	1 / 6
Learning rate (head)	`1e-4`
Encoder (LM) LR	2×10^{-5} (= head LR $\times$ 0.2)
Seed	40
Decoding	greedy argmax (no CRF)
Preprocessing	as in Section 4.1 (headline removal, normalization)
Relation mapping	unified PCC, APA-RST, PARADISE label set
Checkpoint	epoch 5 (selected by best dev F1)
Dev segmentation F1	≈ 0.76

Note. Relation mapping was supplied for parser compatibility; segmentation results do not depend on it.

A.3. RoBERTa-base settings

Item	Value
Backbone	`xlm-roberta-base` (from HuggingFace)
Task head	Token classification (`AutoModelForTokenClassification`)
Labels	["I", "B"]
Tokenizer	`AutoTokenizer.from_pretrained("xlm-roberta-base")`
Input handling:	
max_length	512
stride	128 (for sliding windows on long docs)
padding	`"max_length"`
truncation	`True`
return_offsets_mapping	`True` (for reconstruction)
Training:	
Optimizer	`AdamW`
Learning rate	`5e-5`
Batch size	8
Epochs	3
Loss	CrossEntropyLoss (built into HF model)
Hardware	CUDA if available

A.4. RoBERTa-crf settings

Backbone	`"xlm-roberta-base"` (encoder only)
Custom head	Linear classifier + CRF layer (`torchcrf.CRF`)
Labels	["I", "B", "O"] – required by CRF
Tokenizer	`AutoTokenizer.from_pretrained("xlm-roberta-base")`
Input handling:	
max_length	512
stride	128 (for sliding windows on long docs)
padding	`"max_length"`
truncation	`True`
return_offsets_mapping	`True` (for reconstruction)
Training:	
Loss	Negative log-likelihood from CRF
Optimizer	`AdamW`
Learning rate	`3e-5`
Batch size	8
Epochs	3–5
Hardware	CUDA if available

A.5. LLM System Prompt

```
You are an expert in German linguistics and discourse structure (
    Rhetorical Structure Theory, RST).

TASK
Segment the input text into Elementary Discourse Units (EDUs). EDUs are
    minimal units of text typically corresponding to clauses or
    sentences that can function as building blocks for rhetorical
    analysis. For example, a subordinate clause or a coordinated clause
    usually constitutes its own EDU. However, not every conjunction or
    punctuation indicates a new EDU -- only meaningful rhetorical or
    syntactic boundaries do.

ABSOLUTE RULES
1. Treat sentence-final punctuation (., !, ?, ;, :) as a candidate
    boundary.
2. Break **after** a dash-delimited or parenthetical insertion
   (--- ... ---, - ... -, (...)) **if** the insertion forms a complete
       proposition.
3. Non-restrictive relative, adverbial, and continuative subordinate
    clauses
   (ANR, WEI) are separate EDUs. Restrictive relatives stay inside their
        host.
4. Coordinated main clauses joined by *und, oder, aber, doch* each form
    an EDU.
5. Do **not** split inside abbreviations, numbers with commas, or
    hyphenated
   compounds (z. B., 3,5 km, Bundes-Umwelt-Ministerium).
6. Preserve original punctuation and spacing; output **one EDU per line
    **,
   no numbering, headers, or commentary. Cover the entire text -- no
       overlaps.

EXAMPLE
Input:
"""Die Ketziner und Paretzer werden heute aus der Zeitung erfahren, dass
     sie ab Mittwoch nach Uetz, Marquardt und Potsdam nur noch ueber
    Falkenrehde fahren koennen. Weder das Wasserstrassenneubauamt Berlin
     als Bauherr, noch die bauausfuehrende Firma oder die
    Strassenverkehrsbehoerde Nauen hatten bis gestern Informationen dazu
     veroeffentlicht. Abgesehen von der in Paretz herrschenden
    generellen Ablehnung gegen den Havelausbau und den damit verbundenen
     Brueckenneubau - die Buerger quasi bis fuenf Minuten vor Zwoelf im
    Regen stehen zu lassen, ist nicht zu fassen bei solch einem
    Grossprojekt. Zwar ist die Umleitung laut Bauherr nur 4,5 Kilometer
    lang. Aber was machen die Wanderer und Radfahrer? Und noch schlimmer
     - wie kommen die Landwirte mit ihren Traktoren ab Mittwoch auf die
    Felder jenseits des Kanals? Hier hat das Wasserstrassenneubauamt
    versagt - ohne Wenn und Aber. Wofuer die Behoerde jedoch nichts kann
    , ist der ungluecklich Zeitpunkt des Baubeginns so kurz nach der
    Wiedereroeffnung des Schlosses. Seit einem Jahr ist klar, dass es
```

```
    nach der Buga los geht. Der enorme Zeitverzug beim Schloss war nicht
     vorhersehbar."""

Output:
Die Ketziner und Paretzer werden heute aus der Zeitung erfahren,
dass sie ab Mittwoch nach Uetz, Marquardt und Potsdam nur noch ueber
    Falkenrehde fahren koennen.
Weder das Wasserstrassenneubauamt Berlin als Bauherr, noch die
    bauausfuehrende Firma oder die Strassenverkehrsbehoerde Nauen hatten
     bis gestern Informationen dazu veroeffentlicht.
Abgesehen von der in Paretz herrschenden generellen Ablehnung gegen den
    Havelausbau und den damit verbundenen Brueckenneubau
- die Buerger quasi bis fuenf Minuten vor Zwoelf im Regen stehen zu
    lassen, ist nicht zu fassen bei solch einem Grossprojekt.
Zwar ist die Umleitung
laut Bauherr
nur 4,5 Kilometer lang.
Aber was machen die Wanderer und Radfahrer?
Und noch schlimmer
- wie kommen die Landwirte mit ihren Traktoren ab Mittwoch auf die
    Felder jenseits des Kanals?
Hier hat das Wasserstrassenneubauamt versagt
- ohne Wenn und Aber.
Wofuer die Behoerde jedoch nichts kann, ist der ungluecklicher Zeitpunkt
    des Baubeginns so kurz nach der Wiedereroeffnung des Schlosses.
Seit einem Jahr ist klar, dass es nach der Buga los geht.
Der enorme Zeitverzug beim Schloss war nicht vorhersehbar.

INSTRUCTIONS
Now segment the following German text into EDUs:

"""{text}"""
```

Listing 1: System prompt used for segmentation

A.6. Minimal Python call (OpenAI Responses API)

```
from openai import OpenAI
client = OpenAI()

MODEL = "gpt-4.1-2025-04-14"  # or: "gpt-5-2025-08-07"
resp = client.responses.create(
    model=MODEL,
    # GPT-4.1 (non-reasoning):
    temperature=0.0,
    top_p=1.0,
    # GPT-5 (reasoning): keep defaults
    input=[
        {"role": "system", "content": SYSTEM_PROMPT},
        {"role": "user",   "content": raw_text}
```

```
    ],
)
# Minimal post-processing
edus = [l.strip() for l in resp.output_text.splitlines() if l.strip()]
```

Listing 2: Minimal inference code (OpenAI Responses API)

Environment & versions. Python 3.11.5; `openai` (Python client) = 1.77.0; OS: WSL2.

A.7. Full Results on Test Split

	Prec	Rec	f1	WinDiff	Docs	EDUs Gold	EDUs Pred
APA							
DPLP	0.956	0.885	0.919	0.066	7	122	113
DMRST	0.981	0.836	0.903	0.065	7	122	104
GPT-4	0.932	0.902	0.917	0.075	7	122	118
GPT-5	0.954	0.912	0.932	**0.064**	7	122	117
RoBERTa-base	0.976	0.911	0.943	0.074	7	122	114
RoBERTa-crf	1.000	0.982	**0.991**	0.072	7	122	116
PCC							
DPLP	0.918	0.881	0.904	0.176	18	328	316
DMRST	0.920	0.819	0.866	0.167	18	328	291
GPT-4	0.950	0.875	0.911	0.164	18	328	302
GPT-5	0.875	0.957	0.914	0.167	18	328	359
RoBERTa-base	0.959	0.918	0.938	0.108	18	328	314
RoBERTa-crf	0.957	0.942	**0.949**	**0.095**	18	328	323
PARA							
DPLP	0.780	0.810	0.794	0.318	8	162	163
DMRST	0.837	0.759	0.796	0.259	8	162	147
GPT-4	0.889	0.840	0.863	0.268	8	162	153
GPT-5	0.890	0.919	0.904	0.155	8	162	166
RoBERTa-base	0.959	0.934	**0.947**	**0.094**	8	162	132
RoBERTa-crf	0.907	0.895	0.901	0.165	8	162	133
Kafka							
DPLP	0.651	0.740	0.692	0.451	1	219	248
DMRST	0.711	0.822	0.763	0.400	1	219	252
GPT-4	0.937	0.612	0.740	0.358	1	219	143
GPT-5	0.688	0.845	0.758	0.399	1	219	269
RoBERTa-base	0.776	0.776	**0.776**	**0.324**	1	219	219
RoBERTa-crf	0.707	0.804	0.752	0.415	1	219	249

References

Arendt, H. (2018). *Sechs Essays* (B. Hahn, Ed.). Göttingen: Wallstein Verlag.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463

Bourgonje, P., & Stede, M. (2020). The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1061–1066). Marseille, France: European Language Resources Association. Retrieved from `https://aclanthology.org/2020.lrec-1.133/`

Braud, C., Lacroix, O., & Søgaard, A. (2017a). Cross-lingual and cross-domain discourse segmentation of entire documents. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 237–243). Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/P17-2037

Braud, C., Lacroix, O., & Søgaard, A. (2017b). Does syntax help discourse segmentation? Not so much. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2432–2442). Copenhagen, Denmark: Association for Computational Linguistics. doi: 10.18653/v1/D17-1258

Braud, C., Liu, Y. J., Metheniti, E., Muller, P., Rivière, L., Rutherford, A., & Zeldes, A. (2023). The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In C. Braud et al. (Eds.), *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)* (pp. 1–21). Toronto, Canada: The Association for Computational Linguistics. doi: 10.18653/v1/2023.disrpt-1.1

Braud, C., Zeldes, A., Li, C., Liu, Y. J., & Muller, P. (2025). The DISRPT 2025 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In C. Braud, Y. J. Liu, P. Muller, A. Zeldes, & C. Li (Eds.), *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)* (pp. 1–20). Suzhou, China: Association for Computational Linguistics. doi: 10.18653/v1/2025.disrpt-1.1

Braud, C., Zeldes, A., Rivière, L., Liu, Y. J., Muller, P., Sileo, D., & Aoyama, T. (2024). DISRPT: A Multilingual, Multi-domain, Cross-framework Benchmark for Discourse Processing. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4990–5005). ELRA and ICCL.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. W. Smith (Eds.), *Current and New Directions in Discourse and Dialogue* (pp. 85–112). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-010-0019-2_5

Chistova, E. (2024). Bilingual Rhetorical Structure Parsing with Large Parallel Annotations. In

L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9689–9706). Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.577

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, *abs/1911.02116*. Retrieved from http://arxiv.org/abs/1911.02116

Da Cunha, I., SanJuan, E., Torres Moreno, J. M., Lloberes Salvatella, M., & Castellón Masalles, I. (2010). *Diseg: un segmentador discursivo automático para el español* (Procesamiento del Lenguaje Natural - Nº 45). University of Alicante. Retrieved from http://hdl.handle.net/10045/14690

Fisher, S., & Roark, B. (2007). The utility of parse-derived features for automatic discourse segmentation. In A. Zaenen & A. van den Bosch (Eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 488–495). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P07-1062/

Frenzel, S., & Stede, M. (2025). Sentence-Alignment in Semi-parallel Datasets. In A. Kazantseva, S. Szpakowicz, S. Degaetano-Ortlieb, Y. Bizzoni, & J. Pagel (Eds.), *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)* (pp. 87–96). Association for Computational Linguistics. doi: 10.18653/v1/2025.latechclfl-1.9

Gessler, L., Behzad, S., Liu, Y. J., Peng, S., Zhu, Y., & Zeldes, A. (2021). DisCoDisCo at the DISRPT 2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection. In A. Zeldes, Y. J. Liu, M. Iruskieta, P. Muller, C. Braud, & S. Badene (Eds.), *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)* (pp. 51–62). Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.6

Givón, T. (1983). Topic continuity in spoken english. In *Topic continuity in discourse: A quantitative cross-language study* (p. 343-363). John Benjamins Publishing Company. doi: 10.1075/tsl.3.08giv

Hernault, H., Bollegala, D., & Ishizuka, M. (2010). A Sequential Model for Discourse Segmentation. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 315–326). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hewett, F. (2023). APA-RST: A Text Simplification Corpus with RST Annotations. In M. Strube, C. Braud, C. Hardmeier, J. J. Li, S. Loaiciga, & A. Zeldes (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)* (pp. 173–179). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.codi-1.23

Hirschberg, J., & Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, *19*(3), 501–530. Retrieved from https://aclanthology.org/J93-3003/

Howard, J., & Ruder, S. (2018). Fine-tuned Language Models for Text Classification. *CoRR*, *abs/1801.06146*. Retrieved from http://arxiv.org/abs/1801.06146

Ji, Y., & Eisenstein, J. (2014). Representation Learning for Text-level Discourse Parsing. In *Annual Meeting of the Association for Computational Linguistics*. Retrieved from https://

api.semanticscholar.org/CorpusID:16391334

Joty, S., Carenini, G., & Ng, R. T. (2015). CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, *41*(3), 385–435. doi: 10.1162/COLI_a_00226

Kamaladdini Ezzabady, M., Muller, P., & Braud, C. (2021). Multi-lingual Discourse Segmentation and Connective Identification: MELODI at Disrpt 2021. In A. Zeldes, Y. J. Liu, M. Iruskieta, P. Muller, C. Braud, & S. Badene (Eds.), *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)* (pp. 22–32). Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.3

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the Dark Secrets of BERT. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4365–4374). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-1445

Lalitha Devi, S., Rk Rao, P., & Sundar Ram, V. (2025). SeCoRel: Multilingual Discourse Analysis in DISRPT 2025. In C. Braud, Y. J. Liu, P. Muller, A. Zeldes, & C. Li (Eds.), *Proceedings of the 4th shared task on discourse relation parsing and treebanking (disrpt 2025)* (pp. 79–86). Suzhou, China: Association for Computational Linguistics. doi: 10.18653/v1/2025.disrpt-1.6

Le, H., Abeysinghe, G., & Huyck, C. (2004). Automated discourse segmentation by syntactic information and cue phrases. In *Iasted international conference on artificial intelligence and applications (aia 2004)* (p. 293-298). Innsbruck, Austria: ACTA Press.

Li, J. J., Thadani, K., & Stent, A. (2016). The Role of Discourse Units in Near-Extractive Summarization. In R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, & A. Gainer (Eds.), *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 137–147). Association for Computational Linguistics. doi: 10.18653/v1/W16-3617

Liu, Z., Shi, K., & Chen, N. (2021). DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse* (pp. 154–164). Punta Cana, Dominican Republic and Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.codi-main.15

Longacre, R. E. (1996). *The Grammar of Discourse. Interdisciplinary Contributions to Archaeology Topics in Language and Linguistics*. Springer Science & Business Media.

Loshchilov, I., & Hutter, F. (2017). Fixing Weight Decay Regularization in Adam. *CoRR*, *abs/1711.05101*. Retrieved from http://arxiv.org/abs/1711.05101

Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, *8*, 243-281.

Metheniti, E., Braud, C., Muller, P., & Rivière, L. (2023). DisCut and DiscReT: MELODI at DISRPT 2023. In C. Braud et al. (Eds.), *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)* (pp. 29–42). The Association for Computational Linguistics. doi: 10.18653/v1/2023.disrpt-1.3

Muller, P., Braud, C., & Morey, M. (2019). ToNy: Contextual embeddings for accurate mul-

tilingual discourse segmentation of full documents. In A. Zeldes, D. Das, E. M. Galani, J. D. Antonio, & M. Iruskieta (Eds.), *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019* (pp. 115–124). Minneapolis, MN: Association for Computational Linguistics. doi: 10.18653/v1/W19-2715

Nayak, K. S. R. (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*. (https://caiac.pubpub.org/pub/1fomf4fq)

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. doi: 10.18653/v1/N18-1202

Pevzner, L., & Hearst, M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, *28*(1), 19-36. doi: 10.1162/089120102317341756

Plank, B. (2022). The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10671–10682). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, *12*(5), 601-638. doi: https://doi.org/10.1016/0378-2166(88)90050-1

Polanyi, L., Culy, C., van den Berg, M., Thione, G. L., & Ahn, D. (2004). A Rule Based Approach to Discourse Parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004* (pp. 108–117). Cambridge, Massachusetts, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W04-2322/`

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In N. Calzolari et al. (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA). Retrieved from `https://aclanthology.org/L08-1093/`

Prevot, L., & Muller, P. (2025). A Few Shades of Supervision for Discourse Segmentation: Experiments on a French conversational corpus. *Dialogue & Discourse*, *16*, 35-73. doi: 10.5210/dad.2025.202

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842-866. doi: 10.1162/tacl_a_00349

Seemann, H. J., Shahmohammadi, S., Stede, M., & Scheffler, T. (2024). Spoken vs. Written Computer-Mediated Communication. In C. Poudat & M. Guernut (Eds.), *Proceedings of the 11th International Conference on CMC and Social Media Corpora for the Humanities 2024 (CMC-2024).*

Seyfried, C., Reed, C., & Kamide, Y. (2024). Defining Argumentative Discourse Units as Clauses: Psycholinguistic Evidence. *Computational Models of Argument*.

Shahmohammadi, S., & Stede, M. (2024). Discourse Parsing for German with new RST Corpora. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, & M. Wiegand (Eds.), *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)* (pp. 65–74). Vienna, Austria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.konvens-main.7

Sidarenka, U., Peldszus, A., & Stede, M. (2015). Discourse Segmentation of German Texts. *JLCL*, *30*(1), 71–98.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the Workshop on Discourse Annotation* (pp. 96–102). Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-0213/

Stede, M., Mamprin, S., Peldszus, A., Herzog, A., Kaupat, D., Chiarcos, C., & Warzecha, S. (2015). *Handbuch Textannotation* (M. Stede, Ed.). Universität Potsdam.

Stede, M., Taboada, M., & Das, D. (2017). *Annotation Guidelines for Rhetorical Structure*. Potsdam. (Unpublished manuscript)

Subba, R., & Eugenio, B. M. D. (2007). Automatic Discourse Segmentation using Neural Networks. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*. Retrieved from https://api.semanticscholar.org/CorpusID:17566178

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, *8*(3), 423-459. doi: 10.1177/1461445606061881

Tofiloski, M., Brooke, J., & Taboada, M. (2009). A Syntactic and Lexical-Based Discourse Segmenter. In *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Short Papers)* (pp. 77–80). Suntec, Singapore: Association for Computational Linguistics.

Van der Vliet, N. (2010). Syntax-based Discourse Segmentation of Dutch Text. In M. Slavkovik (Ed.), *Proceedings of the 15th Student Session, ESSLLI* (pp. 203 – 210). (2010/n.h.van.der.vliet/pub001)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2023). *Attention Is All You Need*. Retrieved from https://arxiv.org/abs/1706.03762

Wang, Y., Li, S., & Yang, J. (2018). Toward Fast and Accurate Neural Discourse Segmentation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 962–967). Brussels, Belgium: Association for Computational Linguistics. doi: 10.18653/v1/D18-1116

Xuan Bach, N., Le Minh, N., & Shimazu, A. (2012). A Reranking Model for Discourse Segmentation using Subtree Features. In G. G. Lee, J. Ginzburg, C. Gardent, & A. Stent (Eds.), *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 160–168). Seoul, South Korea: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W12-1623/

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., & Badene, S. (2021). The DISRPT 2021 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In A. Zeldes, Y. J. Liu, M. Iruskieta, P. Muller, C. Braud, &

S. Badene (Eds.), *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)* (pp. 1–12). Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.1

Correspondence

Steffen Frenzel

Universität Potsdam
Applied CL / Discourse Research Lab
Potsdam, Germany
steffen.frenzel@uni-potsdam.de

Maximilian Krupop

Universität Potsdam
Applied CL / Discourse Research Lab
Potsdam, Germany
maximilian.krupop@uni-potsdam.de

Manfred Stede

Universität Potsdam
Applied CL / Discourse Research Lab
Potsdam, Germany
stede@uni-potsdam.de